2-2009

Dynamics of Bayesian updating with dependent data and misspecified models

Cosma R. Shalizi
Carnegie Mellon University, cshalizi@andrew.cmu.edu

Published In
Electron. J. Statist. , 3, 1039-1074.

Submitted to the Annals of Statistics
arXiv: 0901.1342

DYNAMICS OF BAYESIAN UPDATING WITH DEPENDENT DATA AND MISSPECIFIED MODELS

By Cosma Rohilla Shalizi

Statistics Department, Carnegie Mellon University

Santa Fe Institute

Recent work on the convergence of posterior distributions under Bayesian updating has established conditions under which the posterior will concentrate on the truth, if the latter has a perfect representation within the support of the prior, and under various dynamical assumptions, such as the data being independent and identically distributed or Markovian. Here I establish sufficient conditions for the convergence of the posterior distribution in non-parametric problems even when *all* of the hypotheses are wrong, and the data-generating process has a complicated dependence structure. The main dynamical assumption is the generalized asymptotic equipartition (or "Shannon-McMillan-Breiman") property of information theory. I derive a kind of large deviations principle for the posterior measure, and discuss the advantages of predicting using a combination of models known to be wrong. An appendix sketches connections between the present results and the "replicator dynamics" of evolutionary theory.

1. Introduction. The problem of the convergence and frequentist consistency of Bayesian learning goes as follows. We encounter observations $X_1, X_2, \ldots$, which we would like to predict by means of a family Θ of models or hypotheses (indexed by θ). We begin with a prior probability distribution Π_0 over Θ, and update this using Bayes's rule, so that our distribution after seeing $X_1, X_2, \ldots X_t \equiv X_1^t$ is Π_t. If the observations come from a stochastic process with infinite-dimensional distribution P, when does Π_t converge P-almost surely? What is the rate of convergence? Under what conditions will Bayesian learning be consistent, so that Π_t doesn't just converges but its limit is P?

Since the Bayesian estimate is the whole posterior probability distribution Π_t rather than a point or set in Θ, we need a special notion of consistency. The usual approach is to define some sufficiently strong set of neighborhoods

AMS 2000 subject classifications: Primary 62C10, 62G20, 62M09; secondary 60F10, 62M05, 92D15, 94A17

Keywords and phrases: Posterior convergence, Bayesian consistency, asymptotic equipartition, large deviations, Bayesian nonparametrics, replicator dynamics, sofic systems

of P in the space of probability distributions on X_1^∞, and say that Π_t is consistent when, for each such neighborhood N, $\lim_{t\rightarrow\infty} \Pi_t(N) = 1$. When this holds, the posterior increasingly approximates a delta distribution centered at the truth.

The greatest importance of these problems, perhaps, is their bearing on the objectivity and reliability of Bayesian inference; consistency proofs and convergence rates are, as it were, frequentist licenses for Bayesian practices. Moreover, if Bayesian learners starting from different priors converge rapidly on the same posterior distribution, there is less reason to worry about the subjective or arbitrary element in the choice of priors. (Such "merger of opinion" results are also important in economics and game theory [8].) Recent years have seen considerable work on these problems, especially in the non-parametric setting where the model space Θ is infinite-dimensional [24].

Pioneering work by Doob [14], using elegant martingale arguments, established that when any consistent estimator exists, and P lies in the support of Π_0, the set of sample paths on which the Bayesian learner fails to converge to the truth has prior probability zero. (See [9] and [33] for extensions of this result to non-IID settings, and also the discussion in [16].) This is not, however, particularly reassuring, since P generally also has prior probability zero, and it would be unfortunate if these two measure-zero sets should happen to coincide. Indeed, Diaconis and Freedman established that the consistency of Bayesian inference depends crucially on the choice of prior, and that even very natural priors can lead to inconsistency (see [13] and references therein).

More recent work has shown that, no matter what the true data-generating distribution P, Bayesian updating converges along P-almost-all sample paths, provided that (a) P is contained in Θ, (b) every Kullback-Leibler neighborhood in the Θ has some positive prior probability (the "Kullback-Leibler property"), and (c) certain restrictions hold on the prior, amounting to versions of capacity control, as in the method of sieves or structural risk minimization. These contributions also make (d) certain dynamical assumptions about the data-generating process, most often that it is IID [4, 20, 49], independent non-identically distributed [9, 23], or, in some cases, Markovian [22, 23]; [21] and [45] in particular discuss rates of convergence (in the IID setting).

The goal of the present paper is to provide sufficient conditions for the convergence of the posterior without assuming (a) or (b), and substantially weakening (c) and (d). Even if one uses non-parametric models, cases where one knows that the true data generating process is exactly represented by one of the hypotheses in the model class are scarce. Moreover, while IID data can be produced, with some trouble and expense, in the laboratory or in a well-conducted survey, in many applications the data are not just heterogeneous and dependent, but their heterogeneity and dependence is precisely what is of interest. This raises the question of what Bayesian updating does when the truth is not contained in the support of the prior, and observations are dependent to boot.

To answer this question, I first weaken the dynamical assumptions to the asymptotic equipartition property (Shannon-McMillan-Breiman theorem) of information theory, i.e., for each hypothesis θ, the log-likelihood per unit time converges almost surely. This log-likelihood per unit time is basically the growth rate of the Kullback-Leibler divergence between P and θ, $h(\theta)$. As observations accumulate, areas of Θ where $h(\theta)$ exceeds its essential infimum $h(\Theta)$ tend to lose posterior probability, which concentrates in divergence-minimizing regions. Some additional conditions on the prior distribution are needed to prevent it from putting too much weight initially on hypotheses with high divergence rates but slow convergence of the log-likelihood. As the latter assumptions are strengthened, more and more can be said about the convergence of the posterior.

Using the weakest set of conditions (Assumptions 1–3), the long-run exponential growth rate of the posterior density at θ cannot exceed $h(\Theta) - h(\theta)$ (Theorem 1). Adding assumptions 4–6 to provide better control over the integrated or marginal likelihood establishes (Theorem 2) that the long-run growth rate of the posterior density is in fact $h(\Theta) - h(\theta)$. A final extra assumption (7) then lets us conclude (Theorem 3) that the posterior distribution converges, in the sense that, for any set of hypotheses A, the posterior probability $\Pi_t(A) \rightarrow 0$ unless the essential infimum of $h(\theta)$ over A equals $h(\Theta)$. In fact, we then have a kind of large deviations principle for the posterior measure (Theorem 4), as well as a bound on the generalization ability of the posterior predictive distribution (Theorem 5).

For the convenience of reader, the development uses the usual statistical vocabulary and machinery. It may be of some interest, however, that the results were first found via an apparently-novel analogy between Bayesian updating and the "replicator equation" of evolutionary dynamics, which is a formalization of the Darwinian idea of natural selection. Individual hypotheses play the role of distinct replicators in a population, the posterior distribution being the population distribution over replicators and fitness being proportional to likelihood. Appendix A gives details.

2. Preliminaries and Notation. As usual, let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $X_1, X_2, \ldots$, for short X_1^∞, be a sequence of random variables, taking values in the measurable space $(\Xi, \mathcal{X})$, whose infinite-dimensional distribution is P. The natural filtration of this process is $\sigma\left(X_1^t\right)$. The only dynamical properties are those required for the Shannon-McMillan-Breiman theorem (Assumption 3); more specific assumptions such as P being a product measure, Markovian, exchangeable, etc., are not required. Unless otherwise noted, all probabilities are taken with respect to P, and $\mathbf{E}\left[\cdot\right]$ always means expectation under that distribution.

Statistical hypotheses, i.e., distributions of processes adapted to $\sigma\left(X_1^t\right)$, are denoted by F_θ, the index θ taking values in the hypothesis space, a measurable space $(\Theta, \mathcal{T})$, generally infinite-dimensional. For convenience, assume that P and all the F_θ are dominated by a common reference measure, with respective densities p and f_θ. I do not assume that $P \in \Theta$, and *a fortiori* not that $P \in \operatorname{supp} \Pi_0$ — i.e., quite possibly all of the available hypotheses are false.

We will study the evolution of a sequence of probability measures Π_t on $(\Theta, \mathcal{T})$, starting with a *non-random* prior measure Π_0. (A filtration on Θ is not needed; the measures Π_t change but not the σ-field $\mathcal{T}$.) Without loss of generality, assume all Π_t are absolutely continuous with respect to a common reference measure — Π_0 will do, if nothing else — with densities π_t.

Let $L_t(\theta)$ be the conditional likelihood of x_t under θ, i.e., $L_t(\theta) \equiv f_\theta(X_t = x_t | X_1^{t-1} = x_1^{t-1})$, and let $\langle L_t \rangle$ be conditional integrated likelihood, i.e., $\int_\Theta L_t(\theta) d\Pi_t(\theta)$. Bayesian updating of course means that, for any $A \in \mathcal{T}$,

$$\Pi_{t+1}(A) = \frac{\int_A L_t(\theta) d\Pi_t(\theta)}{\int_\Theta L_t(\theta) d\Pi_t(\theta)} = \frac{\int_A L_t(\theta) d\Pi_t(\theta)}{\langle L_t \rangle}$$

or, in terms of the density,

$$\pi_{t+1}(\theta) = \frac{L_t(\theta)\pi_t(\theta)}{\langle L_t \rangle}$$

It will also be convenient to express Bayesian updating in terms of the prior and the total likelihood:

$$\Pi_t(A) = \frac{\int_A d\Pi_0(\theta) f_\theta(x_1^t)}{\int_\Theta d\Pi_0(\theta) f_\theta(x_1^t)} = \frac{\int_A d\Pi_0(\theta) \frac{f_\theta(x_1^t)}{p(x_1^t)}}{\int_\Theta d\Pi_0(\theta) \frac{f_\theta(x_1^t)}{p(x_1^t)}} = \frac{\int_A d\Pi_0(\theta) R_t(\theta)}{\langle R_t \rangle}$$

where $R_t(\theta) \equiv \frac{f_\theta(x_1^t)}{p(x_1^t)}$ is the ratio of model likelihood to true likelihood. (Note that $0 < p(x_1^t) < \infty$ for all t P-a.s.) Similarly,

$$\pi_t(\theta) = \pi_0(\theta) \frac{R_t(\theta)}{\langle R_t \rangle}$$

Remark on the topology of Θ *and on* $\mathcal{T}$. The hope in studying posterior convergence is to show that, as t grows, with higher and higher (P) probability, Π_t concentrates more and more on sets which come closer and closer to P. The tricky part here is "closer and closer": points in Θ represent infinite-dimensional stochastic process distributions, and the topology of such spaces is somewhat odd, and irritatingly abrupt, at least under the more common measures of distance. Any two ergodic measures are either equal or have completely disjoint supports [25], so that the Kullback-Leibler divergence between distinct ergodic processes is always infinity (in both directions), and the total variation and Hellinger distances are likewise maximal. Most previous work on posterior consistency has restricted itself to models where the infinite-dimensional process distributions are formed by products of fixed-dimensional base distributions (IID, Markov, etc.), and in effect transferred the usual metrics' topologies from these finite-dimensional distributions to the processes. It *is* possible to define metrics for general stochastic processes [25], and if readers like they may imagine that $\mathcal{T}$ is a Borel σ-field under some such metric. This is not necessary for the results presented here, however.

2.1. *Example.* The following example will be used to illustrate the assumptions (§2.2.1 and Appendix B), and, later, the conclusions (§3.5).

The data-generating process P is a stationary and ergodic measure on the space of binary sequences, i.e., $\Xi = \{0, 1\}$, and the σ-field $\mathcal{X}$ is naturally 2^{Ξ}. The measure is most easily represented as a function of a two-state Markov chain S_1^{∞}, with $S_t \in \{1, 2\}$. The transition matrix is

$$\mathbf{T} = \begin{bmatrix} 0.0 & 1.0 \\ 0.5 & 0.5 \end{bmatrix}$$

so that the invariant distribution puts probability $1/3$ on state 1 and probability $2/3$ on state 2; take S_1 to be distributed accordingly. The observed process is a binary function of the latent state transitions, $X_t = 0$ if $S_t = S_{t+1} = 2$ and $X_t = 1$ otherwise. Figure 1 depicts the transition and observation structure. Qualitatively, X_1^{∞} consists of blocks of 1s of even length, separated by blocks of 0s of arbitrary length. Since the joint process $\{(S_t, X_t)\}_{1 \leq t \leq \infty}$ is a stationary and ergodic Markov chain, X_1^{∞} is also stationary, ergodic and mixing.

This stochastic process comes from symbolic dynamics [30, 34], where it is known as the "even process", and serves as a basic example of the class of *sofic* processes [50], which have finite Markovian representations, as in Figure 1, but are not Markov at any finite order. Despite their simplicity, these models arise naturally when studying the time series of chaotic dynamical systems [3, 10, 11, 43].

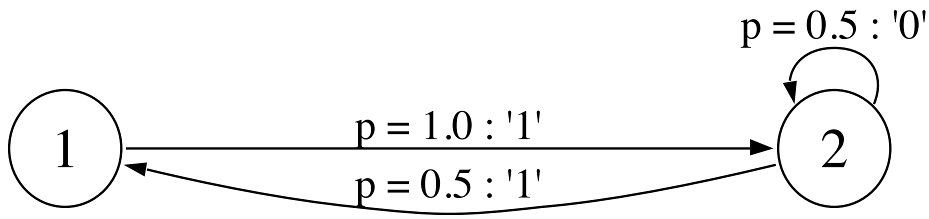

Fig 1. *State-transition diagram for the "even process". The legends on the transition arrows indicate the probability of making the transition, and the observation which occurs when the transition happens. The observation $X_t = 1$ when entering or leaving state 1, otherwise it is 0. This creates blocks of 1s of even length, separated by blocks of 0s of arbitrary length. The result is a finite-state process which is not a Markov chain of any order.*

Let Θ_k be the space of all binary Markov chains of order k with strictly positive transition probabilities and their respective stationary distributions; each Θ_k has dimension 2^k. (Allowing some transition probabilities to be zero creates uninteresting technical difficulties.) Since each hypothesis is equivalent to a function $\Xi^{k+1} \mapsto (0, 1]$, we can give Θ_k the topology of pointwise convergence of functions, and the corresponding Borel σ-field. We will take $\Theta = \bigcup_{k=1}^{\infty} \Theta_k$, identifying Θ_k with the appropriate subset of Θ_{k+1}. Thus Θ consists of all strictly-positive stationary binary Markov chains, of whatever order, and is infinite-dimensional.

As for the prior Π_0, it will be specified in more detail below (§2.2.1). At the very least, however, it needs to have the "Kullback-Leibler rate property", i.e., to give positive probability to every ϵ "neighborhood" $N_\epsilon(\theta)$ around every $\theta \in \Theta$, i.e., the set of hypotheses whose Kullback-Leibler divergence from θ grows no faster than ϵ:

$$N_\epsilon(\theta) = \left\{ \theta' : \epsilon \geq \lim_{t \to \infty} \frac{1}{t} \int dx_1^t f_\theta(x_1^t) \log \frac{f_\theta(x_1^t)}{f_{\theta'}(x_1^t)} \right\}$$

(The limit exists for all θ, θ' combinations [26].)

This example is simple, but it is also beyond the scope of existing work on Bayesian convergence in several ways. First, the data-generating process P is not even Markov. Second, $P \notin \Theta$, so all the hypotheses are wrong, and the truth is certainly not in the support of the prior. (P can however be approximated arbitrarily closely by distributions from Θ in various process metrics.) Third, because P is ergodic, and ergodic distributions are extreme points in the space of stationary distributions [15], it cannot be represented as a mixture of distributions in Θ. This means that the Doob-style theorem of Ref. [33] does not apply, and even the subjective certainty of convergence is not assured.

Ref. [38] describes a non-parametric procedure which will adaptively learn to predict a class of discrete stochastic processes which includes the even process. Ref. [44] introduces a frequentist algorithm which consistently reconstructs the hidden-state representation of sofic processes, including the even process. Ref. [47] considers Bayesian estimation of the even process, using Dirichlet priors for finite-order Markov chains, and employing Bayes factors to decide which order of chain to use for prediction.

2.2. *Assumptions.* The needed assumptions have to do with the dynamical properties of the data generating process P, and with how well the dynamics meshes both with the class of hypotheses Θ and with the prior distribution Π_0 over those hypotheses.

Assumption 1 *The likelihood $R_t(\theta)$ is $\sigma\left(X_1^t\right) \times \mathcal{T}$-measurable for all t.*

The next two assumptions actually need only hold for Π_0-almost-all θ. But this adds more measure-0 caveats to the results, and it is hard to find a natural example where it would help.

Assumption 2 *For every $\theta \in \Theta$, the Kullback-Leibler divergence rate from P,*

$$h(\theta) = \lim_{t\to\infty} \frac{1}{t}\mathbf{E}\left[\log \frac{p(X_1^t)}{f_\theta(X_1^t)}\right]$$

exists (possibly being infinite) and is $\mathcal{T}$-measurable.

As mentioned, any two distinct ergodic measures are mutually singular, so there is a consistent test which can separate them. ([41] constructs an explicit but not necessarily optimal test.) One interpretation of the divergence rate [26] is that it measures the maximum exponential rate at which the power of such tests approaches 1, with $d = 0$ and $d = \infty$ indicating sub- and supra-exponential convergence, respectively.

Assumption 3 *For each $\theta \in \Theta$, the generalized or relative asymptotic equipartition property holds, and so*

$$\lim \frac{1}{t} \log R_t(\theta) = -h(\theta) \tag{1}$$

with P-probability 1.

Refs. [1, 26] for give sufficient, but not necessary, conditions sufficient for Assumption 3 to hold for a given θ. The ordinary, non-relative asymptotic equipartition property, also known as the Shannon-McMillan-Breiman theorem, is that $\lim t^{-1} \log p(x_1^t) = -h_P$ a.s., where h_P is the entropy rate of the data-generating process. (See [26].) If this holds and h_P is finite, one could rephrase Assumption 3 as $\lim t^{-1} \log f_\theta(X_1^t) = -h_P - h(\theta)$ a.s., and state results in terms of the likelihood rather than the likelihood ratio. (Cf. [19, ch. 5].) However, there are otherwise-well-behaved processes for which $h_P = -\infty$, at least in the usual choice of reference measure, so I will restrict myself to likelihood ratios.

The meaning of Assumption 3 is that, relative to the true distribution, the likelihood of each θ goes to zero exponentially, the rate being the Kullback-Leibler divergence rate. Roughly speaking, an integral of exponentially-shrinking quantities will tend to be dominated by the integrand with the slowest rate of decay. This suggests that the posterior probability of a set $A \subseteq \Theta$ depends on the smallest divergence rate which can be attained at a point of prior support within A. Thus, adapting notation from large deviations theory, define

$$\begin{aligned} h(A) &\equiv \operatorname*{ess\,inf}_{\theta \in A} h(\theta) \\ J(\theta) &\equiv h(\theta) - h(\Theta) \\ J(A) &\equiv \operatorname*{ess\,inf}_{\theta \in A} J(\theta) \end{aligned}$$

where here and throughout ess inf is the essential infimum with respect to Π_0, i.e., the greatest lower bound which holds with Π_0-probability 1.

Our further assumptions are those needed for the "roughly speaking" and "should" statements of the previous paragraph to be true, so that, for reasonable sets $A \in \mathcal{T}$,

$$\lim \frac{1}{t} \log \int_A d\Pi_0(\theta) R_t(\theta) = -h(A)$$

Let $I \equiv \{\theta : \; h(\theta) = \infty\}$.

Assumption 4 $\Pi_0(I) < 1$

If this assumption fails, then every hypothesis in the support of the prior doesn't just diverge from the true data-generating distribution, it diverges so fast that the error rate of a test against the latter distribution goes to zero faster than any exponential. (One way this can happen is if every hypothesis has a finite-dimensional distribution assigning probability zero to some event of positive P-probability.) In these situations of extreme mis-specification, the methods of this paper seem to be of no use.

Assumption 5 *There exists a sequence of sets $G_n \to \Theta$ such that*

1. *$\Pi_0(G_n) \geq 1 - \alpha \exp\{-n\beta\}$, for some $\alpha > 0$, $\beta > 2h(\Theta)$;*
2. *The convergence of Eq. 1 is uniform in θ over $G_n \setminus I$;*
3. *$h(G_n) \to h(\Theta)$.*

Comment: Recall that by Egorov's theorem [29, Lemma 1.36, p. 18], if a sequence of finite, measurable functions $f_t(\theta)$ converges pointwise to a finite, measurable function $f(\theta)$ for Π_0-almost-all $\theta \in G$, then for each $\epsilon > 0$, there is a (possibly empty) $B \subset G$ such that $\Pi_0(B) \leq \epsilon$, and the convergence is uniform on $G \setminus B$. Thus the first two parts of the assumption really follow for free from the assumption that likelihoods and divergence rates are measurable in θ. (That β needs to be at least $2h(\Theta)$ becomes apparent in the proof of Lemma 5, but that could always be arranged.) The extra content comes in the third part of the assumption, which could fail if the lowest-divergence hypotheses were also the ones where the convergence was slowest, consistently falling into the bad sets B allowed by Egorov's theorem.

For each measurable $A \subseteq \Theta$, for every $\delta > 0$, there exists a natural number $T(A, \delta, \omega)$ such that

$$t^{-1} \log \int_A R_t(\theta, \omega) d\Pi_0(\theta) \leq \delta + \limsup_t t^{-1} \log \int_A R_t(\theta, \omega) d\Pi_0(\theta)$$

for all $t > T(A, \delta, \omega)$, provided the lim sup is finite. (Here I am explicit in the dependence of the likelihood on the sample path to emphasize that the rate of convergence may be path-dependent.) We need this random last-entry time $T(A, \delta)$ to state the next assumption.

Assumption 6 *The sets G_n of the previous assumption can be chosen so that, for every δ, the inequality $n \geq T(G_n, \delta)$ holds a.s. for all sufficiently large n.*

The meaning of this is that, fixing δ, we can arrange our sequence of good sets so that (at least eventually) we start using G_n only after it has δ-converged.

Finally, to show convergence of the posterior measure, we need to be able to control the convergence of the log-likelihood on sets smaller than the whole parameter space.

Assumption 7 *The sets G_n of the previous two assumptions can be chosen so that, for any set A with $\Pi_0(A) > 0$, $h(G_n \cap A) \to h(A)$.*

Assumption 7 could be replaced by the logically-weaker assumption that for each set A, there exist a sequence of sets $G_{n,A}$ satisfying the equivalents of Assumptions 5 and 6 for the prior measure restricted to A. Since the most straight-forward way to check such an assumption would be to verify Assumption 7 as stated, the extra generality does not seem worth it.

2.2.1. *Verification of Assumptions for the Example.* Since every $\theta \in \Theta$ is a finite-order Markov chain, and P is stationary and ergodic, Assumption 1 is unproblematic, while Assumptions 2 and 3 hold by virtue of [1].

It is easy to check that $\inf_{\theta \in \Theta_k} h(\theta) > 0$ for each k. (The infimum is not in general attained by any $\theta \in \Theta_k$, though it could be if the chains were allowed to have some transition probabilities equal to zero.) The infimum over Θ as a whole, however, is zero. Also, $h(\theta) < \infty$ everywhere (because none of the hypotheses' transition probabilities are zero), so the possible set I of θ with infinite divergence rates is empty, disposing of Assumption 4.

Verifying the remaining assumptions means building a sequence G_n of increasing subsets of Θ on which the convergence of $t^{-1} \log R_t$ is uniform and sufficiently rapid, and ensuring that the prior probability of these sets grows fast enough. This will be done by exploiting some finite-sample deviation bounds for the even process, which in turn rest on its mixing properties and the method of types. Details are referred to Appendix B. The upshot is that the sets G_n consist of chains whose order is less than or equal to $\frac{\log n}{2/3+\epsilon} - 1$, for some $\epsilon > 0$, and where the absolute logarithm of all the transition probabilities is bounded by Cn^{γ}, where the positive constant C is arbitrary but $0 < \gamma < \frac{2/3+\epsilon/2}{2/3+\epsilon}$. The exponential rate $\beta > 0$ for the prior probability of G_n^c can be chosen to be arbitrarily small.

3. Results. I first give the theorems here, without proof. The proofs, in §§3.1–3.4, are accompanied by re-statements of the theorems, for the reader's convenience.

I establish five theorems. The first gives an upper bound on the posterior density at a given point θ in Θ. The second matches the upper bound on the posterior density with a lower bound, together providing the growth-rate for the posterior density. The third is that $\Pi_t(A) \to 0$ for any set with $J(A) > 0$, showing that the posterior concentrates on the divergence-minimizing part of the hypothesis space. The fourth is a kind of large deviations principle for the posterior measure. Finally, the fifth bounds the asymptotic Hellinger and total variation distances between the posterior predictive distribution and the actual conditional distribution of the next observation.

The first result uses only Assumptions 1–3. (It is not very interesting, however, unless 4 is also true.) The latter three, however, all depend on finer control of the integrated likelihood, and so finer control of the prior, as embodied in Assumptions 5–6. More exactly, those additional assumptions concern the interplay between the prior and the data-generating process, restricting the amount of prior probability which can be given to hypotheses whose log-likelihoods converge excessively slowly under P. I build to the first result in the next sub-section, then turn to the control of the integrated likelihood and its consequences in the next three sub-sections, and then consider how these results apply to the example.

Theorem 1 *Under Assumptions 1–3, with probability 1, for all θ where $\pi_0(\theta) > 0$,*

$$\limsup_{t\to\infty} \frac{1}{t} \log \pi_t(\theta) \leq -J(\theta)$$

Theorem 2 *Making Assumptions 1–6, for all $\theta \in \Theta$ where $\pi_0(\theta) > 0$,*

$$\lim_{t\to\infty} \frac{1}{t} \log \pi_t(\theta) = -J(\theta)$$

with probability 1.

Theorem 3 *Make Assumptions 1–7. Pick any set $A \in \mathcal{T}$ where $\Pi_0(A) > 0$ and $h(A) > h(\Theta)$. Then $\Pi_t(A) \to 0$ a.s.*

Theorem 4 *Under the conditions of Theorem 3, if $A \in \mathcal{T}$ is such that*

$$-\limsup t^{-1} \log \Pi_0(A \cap G_t^c) = \beta' \geq 2h(A)$$

then

$$\lim_{t\to\infty} \frac{1}{t} \log \Pi_t(A) = h(\Theta) - h(A)$$

In particular, this holds whenever $2h(A) < \beta$ or $A \subset \bigcap_{k=n}^{\infty} G_k$ for some n.

Theorem 5 *Under Assumptions 1–7,*

$$\begin{aligned} \limsup_{t\to\infty} \rho_H^2(P^t, F_\Pi^t) &\leq h(\Theta) \\ \limsup_{t\to\infty} \rho_{TV}^2(P^t, F_\Pi^t) &\leq 4h(\Theta) \end{aligned}$$

where ρ_H and ρ_{TV} are, respectively, the Hellinger and total variation metrics.

3.1. *Upper Bound on the Posterior Density.* The primary result of this section is a pointwise upper bound on the growth rate of the posterior density. To establish it, I use some subsidiary lemmas, which also recur in later proofs. Lemma 2 extends the almost-sure convergence of the likelihood (Assumption 3) from holding pointwise in Θ to holding on a (possibly random) set of Π_0-measure 1. Lemma 3 shows that the prior-weighted likelihood ratio, $\langle R_t \rangle$ tends to be at least $\exp\{-th(\Theta)\}$. (Both assertions are made more precise in the lemmas themselves.)

Begin with a proposition about exchanging the order of universal quantifiers (with almost-sure caveats).

Lemma 1 *Let $Q \subset \Theta \times \Omega$ be jointly measurable, with sections $Q_\theta = \{\omega : (\omega, \theta) \in Q\}$ and $Q_\omega = \{\theta : (\omega, \theta) \in Q\}$. If, for some probability measure P on Ω,*

$$\forall \theta P\left(Q_\theta\right) = 1 \tag{2}$$

then for any probability measure Π on Θ

$$P\left(\{\omega : \Pi\left(Q_\omega\right) = 1\}\right) = 1 \tag{3}$$

In words, if, for all θ, some property holds a.s., then a.s. the property holds *simultaneously* for almost all θ.

PROOF: Since Q is measurable, for all ω and θ, the sections are measurable, and the measures of the sections, $P(Q_\theta)$ and $\Pi(Q_\omega)$, are measurable functions of θ and ω, respectively. Using Fubini's theorem,

$$\begin{aligned}
\int_\Theta P(Q_\theta) d\Pi(\theta) &= \int_\Theta \int_\Omega \mathbf{1}_Q(\omega, \theta) dP(\omega) d\Pi(\theta) \\
&= \int_\Omega \int_\Theta \mathbf{1}_Q(\omega, \theta) d\Pi(\theta) dP(\omega) \\
&= \int_\Omega \Pi(Q_\omega) dP(\omega)
\end{aligned}$$

By hypothesis, however, $P(Q_\theta) = 1$ for all θ. Hence it must be the case that $\Pi(Q_\omega) = 1$ for P-almost-all ω. (In fact, the set of ω for which this is true must be a measurable set.) □

Lemma 2 *Under Assumptions 1–3, there exists a set $C \subseteq \Xi^\infty$, with $P(C) = 1$, where, for every $y \in C$, there exists a $Q_y \in \mathcal{T}$ such that, for every $\theta \in Q_y$, Eq. 1 holds. Moreover, $\Pi_0(Q_y) = 1$.*

PROOF: Let the set Q consist of the θ, ω pairs where Eq. 1 holds, i.e., for which

$$\lim \frac{1}{t} \log R_t(\theta, \omega) = -h(\theta) \ ,$$

being explicit about the dependence of the likelihood ratio on ω. Assumption 3 states that $\forall \theta P(Q_\theta) = 1$, so applying Lemma 1 just needs the verification that Q is jointly measurable. But, by Assumptions 1 and 2, $h(\cdot)$ is $\mathcal{T}$-measurable, and $R_t(\theta)$ is $\sigma\left(X_1^t\right) \times \mathcal{T}$-measurable for each t, so the set Q where the convergence holds are $\sigma\left(X_1^\infty\right) \times \mathcal{T}$-measurable. Everything then follows from the preceding lemma. □

Remark: Lemma 2 generalizes Lemma 3 in [4]. Lemma 1 is a specialization of the quantifier-reversal lemma used in [36] to prove PAC-Bayesian theorems for learning classifiers. Lemma 1 could be used to extend any of the results below which hold a.s. for each θ to ones which a.s. hold *simultaneously* almost everywhere in Θ. This may seem too good to be true, like an alchemist's recipe for turning the lead of pointwise limits into the gold of uniform convergence. Fortunately or not, however, the lemma tells us nothing about the *rate* of convergence, and is compatible with its varying across Θ from instantaneous to arbitrarily slow, so uniform laws need stronger assumptions.

Lemma 3 *Under Assumptions 1–3, for every $\epsilon > 0$, it is almost sure that the ratio between the integrated likelihood and the true probability density falls below $\exp\{-t(h(\Theta)+\epsilon)\}$ only finitely often:*

$$P\left\{x_1^\infty : \ \langle R_t \rangle \leq \exp\{-t(h(\Theta)+\epsilon)\}, \ \text{i.o.}\right\} = 0 \tag{4}$$

and as a corollary, with probability 1,

$$\liminf_{t\to\infty} \frac{1}{t}\log\langle R_t\rangle \geq -h(\Theta) \tag{5}$$

PROOF: It's enough to show that Eq. 4 holds for all x_1^∞ in the set B from the previous lemma, since that set has probability 1.

Let $N_{\epsilon/2}$ be the set of all θ in the support of Π_0 such that $h(\theta) \leq h(\Theta) + \epsilon/2$. Since $x_1^\infty \in B$, the previous lemma tells us there exists a set $Q_{x_1^\infty}$ of θ for which Eq. 1 holds under the sequence x_1^∞.

$$\begin{aligned}
\exp\{t(\epsilon+h(\Theta))\}\langle R_t\rangle &= \int_\Theta R_t(\theta)\exp\{t(\epsilon+h(\Theta))\}d\Pi_0(\theta)\\
\geq & \int_{N_{\epsilon/2}\cap Q_{x_1^\infty}} R_t(\theta)\exp\{t(\epsilon+h(\Theta))\}d\Pi_0(\theta)\\
= & \int_{N_{\epsilon/2}\cap Q_{x_1^\infty}} \exp\left\{t\left[\epsilon+h(\Theta)+\frac{\log R_t(\theta)}{t}\right]\right\}d\Pi_0(\theta)
\end{aligned}$$

By Assumption 3,

$$\lim_{t\to\infty}\frac{1}{t}\log R_t(\theta) = -h(\theta)$$

and for all $\theta \in N_{\epsilon/2}$, $h(\theta) \leq h(\Theta) + \epsilon/2$, so

$$\liminf_{t\to\infty}\exp\left\{t\left[\epsilon+h(\Theta)+\frac{1}{t}\log R_t(\theta)\right]\right\} = \infty$$

a.s., for all $\theta \in N_{\epsilon/2} \cap Q_{x_1^\infty}$. We must have $\Pi_0(N_{\epsilon/2}) > 0$, otherwise $h(\Theta)$ would not be the essential infimum, and we know from the previous lemma that $\Pi_0(Q_{x_1^\infty}) = 1$, so $\Pi_0(N_{\epsilon/2} \cap Q_{x_1^\infty}) > 0$. Thus, Fatou's lemma gives

$$\lim_{t\to\infty} \int_{N_{\epsilon/2}\cap Q_{x_1^\infty}} \exp\left\{t\left[\epsilon + h(\Theta) + \frac{1}{t}\log R_t(\theta)\right]\right\} d\Pi_0(\theta) = \infty$$

so

$$\lim_{t\to\infty} \exp\left\{t(\epsilon + h(\Theta))\right\} \langle R_t \rangle = \infty$$

and hence

$$\langle R_t \rangle > \exp\left\{-t(\epsilon + h(\Theta))\right\} \tag{6}$$

for all but finitely many t. Since this holds for all $x_1^\infty \in B$, and $P(B) = 1$, Equation 6 holds a.s., as was to be shown. The corollary statement follows immediately. $\square$

Theorem 1 *Under Assumptions 1–3, with probability 1, for all θ where $\pi_0(\theta) > 0$,*

$$\limsup_{t\to\infty} \frac{1}{t}\log \pi_t(\theta) \leq -J(\theta) \tag{7}$$

Proof: As remarked,

$$\pi_t(\theta) = \pi_0(\theta)\frac{R_t(\theta)}{\langle R_t \rangle}$$

so

$$\frac{1}{t}\log \pi_t(\theta) = \frac{1}{t}\log \pi_0(\theta) + \frac{1}{t}\log R_t(\theta) - \frac{1}{t}\log \langle R_t \rangle$$

By Assumption 3, for each $\epsilon > 0$, it's almost sure that

$$\frac{1}{t}\log R_t(\theta) \leq -h(\theta) + \epsilon/2$$

for all sufficiently large t, while by Lemma 3, it's almost sure that

$$\frac{1}{t}\log \langle R_t \rangle \geq -h(\Theta) - \epsilon/2$$

for all sufficiently large t. Hence, with probability 1,

$$\frac{1}{t}\log \pi_t(\theta) \leq h(\Theta) - h(\theta) + \epsilon + \frac{1}{t}\log \pi_0(\theta)$$

for all sufficiently large t. Hence

$$\limsup_{t\to\infty} \frac{1}{t} \log \pi_t(\theta) \leq h(\Theta) - h(\theta) = -J(\theta)$$

□

Lemma 3 gives a lower bound on the integrated likelihood ratio, showing that in the long run it has to do roughly as well as $\exp\{-th(\Theta))\}$. (More precisely, it does significantly worse than that on vanishingly few occasions.) It does not, however, rule out coming closer. Ideally, we would be able to match this lower bound with an upper bound of the same form, since $h(\Theta)$ is the best attainable divergence rate, and, by Lemma 2, log likelihood ratios per unit time are converging to divergence rates for Π_0-almost-all θ, so values of θ for which $h(\theta)$ are close to $h(\Theta)$ should come to dominate the integral in $\langle R_t \rangle$. It would then be fairly straightforward to show convergence of the posterior distribution.

Unfortunately, additional assumptions are required for such an upper bound, because (as earlier remarked) Lemma 2 does not give *uniform* convergence, merely universal convergence; with a large enough space of hypotheses, the slowest pointwise convergence rates can be pushed arbitrarily low. For instance, let $\overline{x_1^t}$ be the distribution on Ξ^∞ which assigns probability 1 to endless repetitions of x_1^t; clearly, under this distribution, seeing $X_1^t = x_1^t$ is almost certain. If such measures fall within the support of Π_0, they will dominate the likelihood, even though $h(\overline{x_1^t}) = \infty$ under all but very special circumstances (e.g., $P = \overline{x_1^t}$). Generically, then, the likelihood and the posterior weight of $\overline{x_1^t}$ will rapidly plumme. To ensure convergence of the posterior, overly-flexible measures like the family of $\overline{x_1^t}$'s must be either excluded from the support of Π_0 (possibly because they are excluded from Θ), or be assigned so little prior weight that they do not end up dominating the integrated likelihood, or the posterior must stably concentrate on them.

3.2. *Convergence of Posterior Density via Control of the Integrated Likelihood.* The next lemma tells us that sets in Θ of exponentially-small prior measure make vanishingly small contributions to the integrated likelihood. It does not require assumptions beyond those used so far, but its application will.

Lemma 4 *Make Assumptions 1–3, and chose a sequence of sets $B_n \subset \Theta$ such that, for all sufficiently large n, $\Pi_0(B_n) \leq \alpha \exp\{-n\beta\}$ for some $\alpha, \beta > 0$. Then, almost surely,*

$$\int_{B_n} R_n(\theta) d\Pi_0(\theta) \leq \exp\{-n\beta/2\} \tag{8}$$

for all but finitely many n.

Proof: By Markov's inequality. First, use Fubini's theorem and the chain rule for Radon-Nikodym derivatives to calculate the expectation value of the ratio.

$$\begin{aligned}
\mathbf{E}\left[\int_{B_n} R_n(\theta) d\Pi_0(\theta)\right] &= \int_{\mathcal{X}^n} dP(x_1^n) \int_{B_n} d\Pi_0(\theta) R_n(\theta) \\
&= \int_{B_n} d\Pi_0(\theta) \int_{\mathcal{X}^n} dP(x_1^n) \frac{dF_\theta}{dP}(x_1^n) \\
&= \int_{B_n} d\Pi_0(\theta) \int_{\mathcal{X}^n} dF_\theta(x_1^n) \\
&= \int_{B_n} d\Pi_0(\theta) \\
&= \Pi_0(B_n)
\end{aligned}$$

Now apply Markov's inequality:

$$\begin{aligned}
P\left\{x_1^n : \int_{B_n} R_n(\theta) d\Pi_0(\theta) > \exp\{-n\beta/2\}\right\} &\leq \exp\{n\beta/2\} \mathbf{E}\left[\int_{B_n} R_n(\theta) d\Pi_0(\theta)\right] \\
&= \exp\{n\beta/2\} \Pi_0(B_n) \\
&\leq \alpha \exp\{-n\beta/2\}
\end{aligned}$$

for all sufficiently large n. Since these probabilities are summable, the Borel-Cantelli lemma implies that, with probability 1, Eq. 8 holds for all but finitely many n. $\square$

The next lemma asserts that a sequence of exponentially-small sets makes a (logarithmically) negligible contribution to the integrated likelihood, provided the exponent is large enough compared to $h(\Theta)$.

Lemma 5 *Let B_n be as in the previous lemma. If $\beta > 2h(\Theta)$, then*

$$\frac{\int_{B_n^c} d\Pi_0(\theta) R_n(\theta)}{\langle R_n \rangle} \rightarrow 1 \tag{9}$$

Proof: Begin by looking at the likelihood integrated over B_n rather than its complement, and apply Lemmas 3 and 4: for any $\epsilon > 0$

$$\frac{\int_{B_n} d\Pi_0(\theta) R_n(\theta)}{\langle R_n \rangle} \leq \frac{\exp\{-n\beta/2\}}{\exp\{-n(h(\Theta) + \epsilon)\}} \tag{10}$$

$$= \exp\{n(\epsilon + h(\Theta) - \beta/2)\} \tag{11}$$

provided n is sufficiently large. If $\beta > 2h(\Theta)$, this bound can be made to go to zero as $n \to \infty$ by taking ϵ to be sufficiently small. Since

$$\langle R_n \rangle = \int_{B_n^c} d\Pi_0(\theta) R_n(\theta) + \int_{B_n} d\Pi_0(\theta) R_n(\theta)$$

it follows that

$$\frac{\int_{B_n^c} d\Pi_0(\theta) R_n(\theta)}{\langle R_n \rangle} \to 1$$

□

Lemma 6 *Make Assumptions 1–3, and take any set G on which the convergence in Eq. 1 is uniform and where $\Pi_0(G) > 0$. Then, P-a.s.,*

$$\limsup_{t \to \infty} \frac{1}{t} \log \int_G d\Pi_0(\theta) R_t(\theta) \leq -h(G) \tag{12}$$

Proof: Pick any $\epsilon > 0$. By the hypothesis of uniform convergence, there almost surely exists a $T(\epsilon)$ such that, for all $t \geq T(\epsilon)$ and for all $\theta \in G$, $t^{-1} \log R_t(\theta) \leq -h(\theta) + \epsilon$. Hence

$$\begin{aligned} t^{-1} \log \int_G d\Pi_0(\theta) R_t(\theta) &= t^{-1} \log \int_G d\Pi_0(\theta) \exp\{\log R_t(\theta)\} && (13) \\ &\leq t^{-1} \log \int_G d\Pi_0(\theta) \exp\{-t[-h(\theta) + \epsilon]\} && (14) \\ &= \epsilon + t^{-1} \log \int_G d\Pi_0(\theta) \exp\{-th(\theta)\} && (15) \end{aligned}$$

Let $\Pi_{0|G}$ denote the probability measure formed by conditioning Π_0 to be in the set G. Then

$$\int_G d\Pi_0(\theta) z(\theta) = \Pi_0(G) \int_G d\Pi_{0|G}(\theta) z(\theta)$$

for any integrable function z. Apply this to the last term from Eq. 15.

$$\log \int_G d\Pi_0(\theta) \exp\{-th(\theta)\} = \log \Pi_0(G) + \log \int_G d\Pi_{0|G}(\theta) \exp\{-th(\theta)\}$$

The second term on the right-hand side is the cumulant generating function of $-h(\theta)$ with respect to $\Pi_{0|G}$, which turns out to have exactly the right behavior as $t \to \infty$.

$$\begin{aligned}
\frac{1}{t}\log \int_G d\Pi_{0|G}(\theta)\exp\{-th(\theta)\} &= \frac{1}{t}\log \int_G d\Pi_{0|G}(\theta)|\exp\{-h(\theta)\}|^t \\
&= \frac{1}{t}\log\left(\left(\int_G d\Pi_{0|G}(\theta)|\exp\{-h(\theta)\}|^t\right)^{1/t}\right)^t \\
&= \frac{1}{t}\left[t\log\|\exp\{-h(\theta)\}\|_{t,\Pi_{0|G}}\right] \\
&= \log\|\exp\{-h(\theta)\}\|_{t,\Pi_{0|G}} \qquad (16)
\end{aligned}$$

Since $h(\theta) \geq 0$, $\exp\{-h(\theta)\} \leq 1$, and the L_p norm of the latter will grow towards its L_∞ norm as p grows. Hence, for sufficiently large t,

$$\begin{aligned}
\log\|\exp\{-h(\theta)\}\|_{t,\Pi_{0|G}} &\leq \log\|\exp\{-h(\theta)\}\|_{\infty,\Pi_{0|G}} + \epsilon \\
&= -\operatorname*{ess\,inf}_{\theta\in G} h(\theta) + \epsilon \\
&= -h(G) + \epsilon \qquad (17)
\end{aligned}$$

where the next-to-last step uses the monotonicity of log and exp.

Putting everything together, we have that, for any $\epsilon > 0$ and all sufficiently large t,

$$t^{-1}\log\int_G d\Pi_0(\theta)R_t(\theta) \leq -h(G) + 2\epsilon + \frac{\log \Pi_0(G)}{t}$$

Hence the limit superior of the left-hand side is at most $-h(G)$. □

Lemma 7 *Making Assumption 1–6,*

$$\limsup_{n\to\infty}\frac{1}{n}\log\langle R_n\rangle \leq -h(\Theta) \qquad (18)$$

Proof: By Lemma 5,

$$\lim_{n\to\infty}\frac{\int_\Theta R_n d\Pi_0(\theta)}{\int_{G_n} R_n d\Pi_0(\theta)} = 1$$

implying that

$$\lim_{n\to\infty}\log\int_\Theta R_n(\theta)d\Pi_0(\theta) - \log\int_{G_n} R_n(\theta)d\Pi_0(\theta) = 0$$

so for every $\epsilon > 0$, for n large enough

$$\log\int_\Theta R_n(\theta)d\Pi_0(\theta) \leq \epsilon/3 + \log\int_{G_n} R_n(\theta)d\Pi_0(\theta)$$

Consequently, again for large enough n,

$$\frac{1}{n}\log\int_{\Theta} R_n(\theta)d\Pi_0(\theta) \leq \epsilon/3n + \frac{1}{n}\log\int_{G_n} R_n(\theta)d\Pi_0(\theta)$$

Now, for each G_n, for every $\epsilon > 0$, if $t \geq T(G_n, \epsilon/3)$ then

$$\frac{1}{t}\log\int_{G_n} R_t(\theta)d\Pi_0(\theta) \leq -h(G_n) + \epsilon/3$$

by Lemma 6. If $n \geq T(G_n, \epsilon/3)$ (which, by Assumption 6, is true for all sufficiently large n), then

$$\frac{1}{n}\log\int_{\Theta} R_n(\theta)d\Pi_0(\theta) \leq -h(G_n) + \epsilon/3n + \epsilon/3$$

for all $\epsilon > 0$ and all n sufficiently large. Since, by Assumption 5, $h(G_n) \to h(\Theta)$, for every $\epsilon > 0$, $h(G_n)$ is within $\epsilon/3$ of $h(\Theta)$ for large enough n, so

$$\frac{1}{n}\log\int_{\Theta} R_n(\theta)d\Pi_0(\theta) \leq -h(\Theta) + \epsilon/3n + \epsilon/3 + \epsilon/3$$

Thus, for every $\epsilon > 0$, then we have that

$$\frac{1}{n}\log\int_{\Theta} R_n(\theta)d\Pi_0(\theta) \leq -h(\Theta) + \epsilon$$

for large enough n, or, in short,

$$\limsup_{n\to\infty}\frac{1}{n}\log\int_{\Theta} R_n(\theta)d\Pi_0(\theta) \leq -h(\Theta)$$

□

Lemma 8 *Making Assumptions 1–6, if* $\Pi_0(I) = 0$*, then*

$$\frac{1}{t}\log\langle R_t\rangle \to -h(\Theta) \tag{19}$$

almost surely.

PROOF: Combining Lemmas 3 and 7,

$$-h(\Theta) \leq \liminf_{t\to\infty}\frac{1}{t}\log\langle R_t\rangle \leq \limsup_{t\to\infty}\frac{1}{t}\log\langle R_t\rangle \leq -h(\Theta)$$

□

The standard version of Egorov's theorem concerns sequences of finite measurable functions converging pointwise to finite measurable limiting functions. However, the proof is easily adapted to the case where the limiting is infinite.

Lemma 9 *Let $f_t(\theta)$ be a sequence of finite, measurable functions, converging to ∞ almost everywhere (Π_0) on I. Then for each $\epsilon > 0$, there exists a possibly-empty $B \subset I$ such that $\Pi_0(B) < \epsilon$, and the convergence is uniform on $I \setminus B$.*

PROOF: Parallel to the usual proof of Egorov's theorem. Begin by removing the measure-zero set of points on which pointwise convergence fails; for simplicity, keep the name I for the remaining set. For each natural number t and k, let $B_{t,k} \equiv \{\theta \in I : f_t(\theta) < k\}$ — the points where the function fails to be at least k by step t. Since the limit of f_t is ∞ everywhere on I, each θ has a last t such that $f_t(\theta) < k$, no matter how big k is. Hence $\bigcap_{t=1}^{\infty} B_{t,k} = \emptyset$. By continuity of measure, for any $\delta > 0$, there exists an n such that $\Pi_0(B_{t,k}) < \delta$ if $t \geq n$. Fix ϵ as in the statement of the lemma, and set $\delta = \epsilon 2^{-k}$. Finally, set $B = \bigcup_{k=1}^{\infty} B_{n,k}$. By the union bound, $\Pi_0(B) \leq \epsilon$, and by construction, the rate of convergence to ∞ is uniform on $I \setminus B$. □

Lemma 10 *The conclusion of Lemma 8 is unchanged if $\Pi_0(I) > 0$.*

PROOF: The integrated likelihood ratio can be divided into two parts, one from integrating over I and one from integrating over its complement. Previous lemmas have established that the latter is upper bounded, in the long run, by a quantity which is $O(\exp\{-h(\Theta)t\})$. We can use Lemma 9 to divide I into a sequence of sub-sets, on which the convergence is uniform, and hence on which the integrated likelihood shrinks faster than any exponential function, and remainder sets, of prior measure no more than $\alpha \exp\{-n\beta\}$, on which the convergence is less than uniform (i.e., slow). If we ensure that $\beta > 2h(\Theta)$, however, by Lemma 5 the remainder sets' contributions to the integrated likelihood is negligible in comparison to that of $\Theta \setminus I$. Said another way, if there are alternatives which a consistent test would rule out at a merely exponential rate, those which would be rejected at a supra-exponential rate end up making vanishingly small contributions to the integrated likelihood. □

Theorem 2 *Making Assumptions 1–6, for all $\theta \in \Theta$ where $\pi_0(\theta) > 0$,*

$$\lim_{t\to\infty} \frac{1}{t} \log \pi_t(\theta) = -J(\theta) \tag{20}$$

with probability 1.

PROOF: Theorem 1 says that, for all θ,

$$\limsup_{t\to\infty} \frac{1}{t} \log \pi_t(\theta) \leq -J(\theta)$$

a.s., so there just needs to be a matching lim inf. Pick any $\epsilon > 0$. By Assumption 3, it's almost certain that, for all sufficiently large t,

$$\frac{1}{t}\log R_t(\theta) \geq -h(\theta) - \epsilon/2$$

while by Lemma 10, it's almost certain that for all sufficiently large t,

$$\frac{1}{t}\log \langle R_t \rangle \leq -h(\Theta) + \epsilon/2$$

Combining these as in the proof of Theorem 1, it's almost certain that for all sufficiently large t

$$\frac{1}{t}\log \pi_t(\theta) \geq h(\Theta) - h(\theta) - \epsilon$$

so

$$\liminf_{t\to\infty} \frac{1}{t}\log \pi_t(\theta) \geq h(\Theta) - h(\theta) = -J(\theta)$$

□

3.3. *Convergence and Large Deviations of the Posterior Measure.* Adding Assumption 7 to those before it implies that the posterior measure concentrates on sets $A \subset \Theta$ where $h(A) = h(\Theta)$.

Theorem 3 *Make Assumptions 1–7. Pick any set $A \in \mathcal{T}$ where $\Pi_0(A) > 0$ and $h(A) > h(\Theta)$. Then $\Pi_t(A) \to 0$ a.s.*

PROOF:

$$\begin{aligned}\Pi_t(A) &= \Pi_t(A \cap G_t) + \Pi_t(A \cap G_t^c)\\ &\leq \Pi_t(A \cap G_t) + \Pi_t(G_t^c)\end{aligned}$$

The last term is easy to bound. From Eq. 11 in the proof of Lemma 5,

$$\begin{aligned}\Pi_t(G_t^c) &= \frac{\int_{G_t^c} d\Pi_0(\theta)R_t(\theta)}{\langle R_t \rangle}\\ &\leq \exp\{t(\epsilon + h(\Theta) - \beta/2\} \qquad (21)\end{aligned}$$

for any $\epsilon > 0$, for all sufficiently large t, almost surely. Since $\beta > 2h(\Theta)$, the whole expression $\to 0$ as $t \to \infty$.

To bound $\Pi_t(A \cap G_t)$, reasoning as in the proof of Lemma 7, but invoking Assumption 7, leads to the conclusion that, for any $\epsilon > 0$, with probability 1,

$$\frac{1}{t}\log \int_{A\cap G_t} d\Pi_0(\theta)R_t(\theta) \leq -h(A) + \epsilon$$

for all sufficiently large n. Recall that by Lemma 3, for all $\epsilon > 0$ it's almost sure that

$$\frac{1}{t}\log\langle R_t\rangle \geq -h(\Theta) - \epsilon$$

for all sufficiently large n. Hence for every $\epsilon > 0$, it's almost certain that for all sufficiently large t,

$$\Pi_t(A \cap G_t) \leq \exp\{t[h(\Theta) - h(A) + 2\epsilon]\} \tag{22}$$

Since $h(A) > h(\Theta)$, by picking ϵ small enough the right hand side goes to zero. □

The proof of the theorem provides an exponential *upper* bound on the posterior measure of sets where $h(A) > h(\Theta)$. In fact, even without the final assumption needed for the theorem, there is an exponential *lower* bound on that posterior measure.

Lemma 11 *Make Assumption 1–6, and pick a set $A \in \mathcal{T}$ with $\Pi_0(A) > 0$. Then*

$$\liminf_{t\to\infty} \frac{1}{t}\log \Pi_t(A) \geq h(\Theta) - h(A) \tag{23}$$

Proof: Reasoning as in the proof of Lemma 3, it is easy to see that

$$\liminf_{t\to\infty} \frac{1}{t}\log\int_A d\Pi_0(\theta)R_t(\theta) \geq -h(A)$$

and by Lemma 7,

$$\limsup_{t\to\infty} \frac{1}{t}\log\langle R_t\rangle \leq -h(\Theta)$$

hence

$$\begin{aligned}\liminf_{t\to\infty} \frac{1}{t}\log \Pi_t(A) &= \liminf_{t\to\infty} \frac{1}{t}\log\frac{\int_A d\Pi_0(\theta)R_t(\theta)}{\langle R_t\rangle} \\ &\geq -h(A) + h(\Theta)\end{aligned}$$

□

Theorem 4 *Under the conditions of Theorem 3, if $A \in \mathcal{T}$ is such that*

$$-\limsup t^{-1}\log \Pi_0(A \cap G_t^c) = \beta' \geq 2h(A) \tag{24}$$

then

$$\lim_{t\to\infty} \frac{1}{t}\log \Pi_t(A) = h(\Theta) - h(A) \tag{25}$$

In particular, this holds whenever $2h(A) < \beta$ or $A \subset \bigcap_{k=n}^{\infty} G_k$ for some n.

PROOF: Trivially,

$$\frac{1}{t}\log \Pi_t(A) = \frac{1}{t}\log \Pi_t(A \cap G_t) + \Pi(A \cap G_t^c)$$

From Eq. 22 from the proof of Theorem 3, we know that, for any $\epsilon > 0$,

$$\Pi_t(A \cap G_t) \leq \exp\left\{t[h(\Theta) - h(A) + \epsilon]\right\}$$

a.s. for sufficiently large t. On the other hand, under the hypothesis of the theorem, the proof of Eq. 21 can be imitated for $\Pi_t(A \cap G_t^c)$, with the conclusion that, for all $\epsilon > 0$,

$$\Pi_t(A \cap G_t^c) \leq \exp\left\{t[h(\Theta) - \beta'/2 + \epsilon]\right\}$$

again a.s. for sufficiently large t. Since $\beta'/2 > h(A)$, Eq. 25 follows.

Finally, to see that this holds for any A where $h(A) < \beta/2$, observe that we can always upper bound $\Pi_t(A \cap G_n^c)$ by $\Pi_t(G_n^c)$, and the latter goes to zero with rate $-\beta/2$. □

Remarks: Because $h(A)$ is the essential infimum of $h(\theta)$ on the set A, as the set shrinks $h(A)$ grows. Sets where $h(A)$ is much larger than $h(\Theta)$ tend accordingly to be small. The difficulty is that the sets G_t^c are also small, and conceivably overlaps so heavily with A that the integral of the likelihood over A is dominated by the part coming from $A \cap G_t^c$. Eventually this will shrink towards zero exponentially, but perhaps only at the comparatively slow rate $h(\Theta) - \beta/2$, rather than the faster rate $h(\Theta) - h(A)$ attained on the well-behaved part $A \cap G_t$.

Theorem 4 is close to, but not quite, a large deviations principle on Θ. We have shown that the posterior probability of any arbitrary set A where $J(A) > 0$ goes to zero with an exponential rate at least equal to

$$\beta/2 \wedge \operatorname*{ess\,inf}_{\theta \in A} J(\theta) = \operatorname*{ess\,inf}_{\theta \in A} \beta/2 \wedge J(\theta) \tag{26}$$

But in a true LDP, the rate would have to be an infimum, not just an essential infimum, of a point-wise rate function. This deficiency could be removed by means of additional assumptions on Π_0 and $h(\theta)$.

Ref. [17] obtains proper large and even moderate deviations principles, but for the location of Π_t in the space $\mathcal{M}_1(\Theta)$ of all distributions on Θ, rather than on Θ itself. Essentially, they use the assumption of IID sampling, which makes the posterior a function of the empirical distribution, to leverage the LDP for the latter into an LDP for the former. This strategy may be more widely applicable but goes beyond the scope of this paper. Papangelou [40], assuming that Θ consists of discrete-valued Markov chains of arbitrary order and P is in the support of the prior, and using methods similar to those in Appendix B, derives a result which is closely related to Theorem 4. In fact, fixing the sets G_n as in Appendix B, Theorem 4 implies the theorem of [40].

3.4. *Generalization Performance.* Lemma 10 shows that, in hindsight, the Bayesian learner does a good job of matching the data: the log integrated likelihood ratio per time-step approaches $-h(\Theta)$, the limit of values attainable by individual hypotheses within the support of the prior. This leaves open, however, the question of the prospective or generalization performance.

The one-step-ahead predictive distribution of the hypothesis θ is given by $F_\theta\left(X_t|\sigma\left(X_1^{t-1}\right)\right)$, with the convention that $t = 1$ gives the marginal distribution of the first observation. Abbreviate this by F_θ^t. Similarly, let $P^t \equiv P\left(X_t|\sigma\left(X_1^{t-1}\right)\right)$; this is the best probabilistic prediction we could make, did we but know P [31]. The posterior predictive distribution is given by mixing the individual predictive distributions with weights given by the posterior:

$$F_\Pi^t \equiv \int_\Theta F_\theta^t d\Pi_t(\theta)$$

What we want is for F_Π^t to approach P^t, but we cannot in general hope for the convergence to be complete, since our models are mis-specified. The final theorem uses $h(\Theta)$ to put an upper bound on how far the posterior predictive distribution can remain from the true predictive distribution.

Theorem 5 *Under Assumptions 1–7,*

$$\limsup_{t\to\infty} \rho_H^2(P^t, F_\Pi^t) \leq h(\Theta) \tag{27}$$

$$\limsup_{t\to\infty} \rho_{TV}^2(P^t, F_\Pi^t) \leq 4h(\Theta) \tag{28}$$

where ρ_H and ρ_{TV} are, respectively, the Hellinger and total variation metrics.

Proof: Recall the well-known inequalities relating Hellinger distance to to Kullback-Leibler divergence on the one side and to total variation distance on the other: distributions P and Q:

$$\rho_H^2(P,Q) \leq D(P\|Q) \tag{29}$$

$$\rho_{TV}(P,Q) \leq 2\rho_H(P,Q) \tag{30}$$

It's enough to prove Eq. 27, and Eq. 28 then follows from Eq. 30.

Abbreviate $\rho_H(P^t, F^t_\theta)$ by $\rho_H(t,\theta)$. Pick any $\epsilon > h(\Theta)$, and say that $A_\epsilon = \{\theta : \rho^2_H(t,\theta) > \epsilon\}$. By convexity and Jensen's inequality,

$$\begin{aligned}
\rho^2_H(P^t, F^t_\Pi) &\leq \int_\Theta \rho^2_H(t,\theta)d\Pi_n(\theta) \\
&= \int_{A^c_\epsilon} \rho^2_H(t,\theta)d\Pi_n(\theta) + \int_{A_\epsilon} \rho^2_H(t,\theta)d\Pi_n(\theta) \\
&= \epsilon\Pi_t(A^c_\epsilon) + \sqrt{2}\Pi_t(A_\epsilon)
\end{aligned}$$

By Eq. 29, $d(\theta) > \rho^2_H(t,\theta)$. Thus $h(A_\epsilon) \geq \epsilon$, and $\epsilon > h(\theta)$ so, by Theorem 3, $\Pi_t(A_\epsilon) \to 0$ a.s. Hence

$$\rho^2_H(P^t, F^t_\Pi) \leq \epsilon$$

eventually almost surely. Since this holds for any $\epsilon > h(\Theta)$, Eq. 27 follows. □

Remark: It seems like it should be possible to prove a similar result for the divergence rate of the predictive distribution, namely that

$$\limsup_{t\to\infty} h(\Pi_t) \leq h(\Theta)$$

but it would take a different approach, because $h(\cdot)$ has no upper bound, and the posterior weight of the high-divergence regions might decay too slowly to compensate for this.

3.5. *Application of the Results to the Example.* Because $h(\Theta) = 0$, while $h(\theta) > 0$ everywhere, the behavior of the posterior is somewhat peculiar. Every compact set $K \subset \Theta$ has $J(K) > 0$, so by Theorem 3, $\Pi_t(K) \to 0$. On the other hand, $\Pi_t(G_t) \to 1$ — the sequence of good sets contains models of increasingly high order, with increasingly weak constraints on the transition probabilities, and this lets its posterior weight grow, even though every individual compact set within it ultimately loses all weight.

In fact, each G_n is a convex set, and $h(\cdot)$ is a convex function, so there is a unique minimizer of the divergence rate within each good set. Conditional on being within G_n, the posterior probability becomes increasingly concentrated on neighborhoods of this minimizer, but the minimizer itself keeps moving, since it can always be improved upon by increasing the order of the chain and reducing some transition probabilities. (Recall that P gives probability 0 to sequences 010, 01110, etc., where the block of 1's is of odd length, but Θ contains only chains with strictly positive transition probabilities.)

Outside of the good sets, the likelihood is peaked around hypotheses which provide stationary and smooth approximations to the $\overline{x^t_1}$ distribution that endlessly repeats the observed sequence to date. The divergence rates of these hypotheses are however extremely high, so none of them retains its high likelihood for very long. ($\overline{x_1^t}$ is a Markov chain of order t, but it is not in Θ, since it's neither stationary nor does it have strictly positive transition probabilities. It can be made stationary, however, by assigning equal probability to each of its t states; this gives the data likelihood $1/t$ rather than 1, but that still is vastly larger than the $O(\exp\{-ct\})$ likelihoods of better models. (Recall that even the likelihood of the true distribution is only $O(2^{-\frac{2}{3}t})$.) Allowing each of the t states to have a probability $0 < \iota \ll 1$ of not proceeding to the next state in the periodic sequence is easy and leads to only an $O(\iota t)$ reduction in the likelihood up to time t. In the long run, however, it means that the likelihood will be $O(\iota^t)$.) In any case, the total posterior probability of G_t^c is going to zero exponentially.

Despite — or rather, because of — of the fact that no point in Θ is the *ne plus ultra* around which the posterior concentrates, the conditions of Theorem 5 are met, and since $h(\Theta) = 0$, the posterior predictive distribution converges to the true predictive distribution in the Hellinger and total variation metrics. That is, the weird gyrations of the posterior do not prevent us from attaining *predictive* consistency. This is so even though the posterior always gives the wrong answer to such basic questions as "Is $P(X_t^{t+2} = 010) > 0$?" — inferences which in this case can be made correctly through non-Bayesian methods [38, 44].

4. Discussion. The crucial assumptions were 3, 5 and 6. Together, these amount to assuming that the time-averaged log likelihood ratio converges universally; to fashioning a sieve, successively embracing regions of Θ where the convergence is increasingly ill-behaved; and the hope that the prior weight of the remaining bad sets can be bounded exponentially.

Using asymptotic equipartition in place of the law of large numbers is fairly straightforward. Both results belong to the general family of ergodic theorems, which allow us to take sufficiently long sample paths as representative of entire processes. The unique a.s. limit in Eq. 1 can be replaced with a.s. convergence to a distinct limit in each ergodic component of P. However, the notation gets ugly, so the reader should regard $h(\theta)$ as that random limit, and treat all subsequent results as relative to the ergodic decomposition of P. (Cf. [12, 25].) It may be possible to weaken this assumption yet further, but it is hard to see how Bayesian updating can succeed if the past performance of the likelihood is not a guide to future results.

A bigger departure from the usual approach to posterior convergence may be allowing $h(\Theta) > 0$; this rules out posterior consistency, to begin with.

More subtly, it requires $\beta > 2h(\Theta)$. This means that a prior distribution which satisfies the assumptions for one value of P may not satisfy them for another, depending, naturally enough, on just how mis-specified the hypotheses are, and how much weight the prior puts on very bad hypotheses. On the other hand, when $h(\Theta) = 0$, Theorem 5 implies predictive consistency, as in the example.

Assumption 6 is frankly annoying. Its job is to make sure that the log likelihood ratio doesn't just converge, but converges quickly, at least on the good sets, that we can be confident that integrated likelihood of G_n has converged well by the time we want G_n to start dominating the prior. Unfortunately, verifying the assumption in its present form means proving the sub-linear growth rate of sequences of random last entry times, and these times are not generally convenient to work with. (Cf. Appendix B.) It would be nice to replace it with a bracketing or metric entropy condition, as in [4] or (in a non-Bayesian context) [37, 48]. It seems doubtful that a uniformly consistent test condition, of the kind widely employed in Bayesian nonparametrics [24, 51] would work when, in fact, the truth is not in the support of the prior.

These results go some way toward providing a frequentist explanation of the success of Bayesian methods in many practical problems. Under these conditions, the posterior is increasingly weighted towards the parts of Θ which are closest (in the Kullback-Leibler sense) to the data-generating process P. For a $\Pi_t(A)$ to persistently be much more or much less than $\approx \exp\{-tJ(A)\}$, $R(\theta)$ must be persistently far from $\exp\{-th(\theta)\}$, not just for isolated $\theta \in A$, but a whole positive-measure subset of them. With a reasonably smooth prior, this requires a run of bad luck amounting almost to a conspiracy. From this point of view, Bayesian inference amounts to introducing bias so as to reduce variance, and then relaxing the bias. Experience with frequentist non-parametric methods shows this can work if the bias is relaxed sufficiently slowly, which is basically what the assumptions here do. As the example shows, this can succeed as a predictive tactic without supporting substantive inferences about the data-generating process.

When $h(\Theta) > 0$ and all the models are more or less wrong, there is an additional advantage to averaging the models, as is done in the predictive distribution. (I owe the argument which follows to Scott Page; cf. [39].) With a convex loss function ℓ, such as squared error, Kullback-Leibler divergence, Hellinger distance, etc., the loss of the predictive distribution $\ell(\Pi_t)$ will be no larger than the posterior-mean loss of the individual models $\langle \ell(\theta) \rangle$. For squared error loss, the difference is equal to the variance of the models' predictions [32]. For divergence, a little algebra shows that

$$h(\Pi_t) = \langle h(\theta) \rangle + \left\langle \mathbf{E}\left[\log \frac{dF_\theta}{dF_\Pi}\right]\right\rangle \tag{31}$$

where the second term on the RHS is again an indication of the diversity of the models; the more different their predictions are, on the kind of data generated by P, the smaller the error of made by the mixture. Having a diversity of wrong answers can be as important as reducing the average error itself. The way to accomplish this is to give more weight to models which make mostly good predictions, but make different mistakes. This suggests that there may actually be predictive benefits to having the posterior concentrate on a set containing multiple hypotheses.

Finally, it is worth remarking on the connection between these results and prediction with "mixtures of experts" [2, 7]. Formally, the role of the negative log-likelihood and of Bayes's rule in this paper was to provide a loss function and a multiplicative scheme for updating the weights. All but one of the main results (the exception Theorem 5, which uses the connection between the Kullback-Leibler divergence and Hellinger distance) would carry over to multiplicative weight training using a different loss function, provided the accumulated loss per unit time converged.

APPENDIX A: BAYESIAN UPDATING AS REPLICATOR DYNAMICS

Replicator dynamics are one of the fundamental models of evolutionary biology; they represent the effects of natural selection in large populations, without (in their simplest form) mutation, sex, or other sources of variation. [28] provides a thorough discussion. They also arise as approximations to many other adaptive processes, such as reinforcement learning [5, 6, 42]. In this appendix, I show that Bayesian updating also follows the replicator equation.

We have a set of *replicators* — phenotypes, species, reproductive strategies, etc. — indexed by $\theta \in \Theta$. The population density at type θ is $\pi(\theta)$. We denote by $\phi_t(\theta)$ the *fitness* of θ at time t, i.e., the average number of descendants left by each individual of type θ. The fitness of θ may in fact be a function of π_t, in which case it is said to be *frequency-dependent*. Many applications assume the fitness function to be deterministic, rather than random, and further assume that it is not an explicit function of t, but these restrictions are inessential.

The discrete-time *replicator dynamic* [28] is the dynamical system given by the map

$$\pi_t(\theta) = \pi_{t-1}(\theta)\frac{\phi_t(\theta)}{\langle \phi_t \rangle} \tag{32}$$

where $\langle \phi_t \rangle$ is the population mean fitness at t, i.e.,

$$\langle \phi_t \rangle \equiv \int_{\Theta} \phi_t(\theta) d\pi_t(\theta)$$

The effect of these dynamics is to re-weight the population towards replicators with above-average fitness.

It is immediate that Bayesian updating has the same form as Eq. 32, as soon as we identify the distribution of replicators with the posterior distribution, and the fitness with the conditional likelihood. In fact, Bayesian updating is an extra simple case of the replicator equation, since the fitness function is frequency-independent, though stochastic. Updating corresponds to the action of natural selection, without variation, in a fluctuating environment. The results in the main text assume (Assumption 3) that, despite the fluctuations, the long-run fitness is nonetheless a determinate function of θ. The theorems assert that selection can then be relied upon to drive the population to the peaks of the long-run fitness function, at the cost of reducing the diversity of the population, rather as in Fisher's fundamental theorem of natural selection [18, 28].

Corollary 1 *Define the relative fitness $\tilde{\phi}_t(\theta) \equiv L_t(\theta) / \langle L_t \rangle$. Under the conditions of Theorem 2, the time average of the log relative fitness converges a.s.*

$$\frac{1}{t} \sum_{n=1}^{t} \log \tilde{\phi}_n(\theta) \to -J(\theta) + o(1) \tag{33}$$

Proof: Unrolling Bayes's rule over multiple observations,

$$\pi_t(\theta) = \pi_0(\theta) \prod_{n=1}^{t} \tilde{\phi}_n(\theta)$$

Take log of both sides, divide through by t, and invoke Theorem 2. □

Remark: Theorem 2 implies that

$$H_t \equiv |\log \pi_t(\theta) + tJ(\theta)|$$

is a.s. $o(t)$. To strengthen Eq. 33 from convergence of the time average or Cesàro mean to plain convergence requires forcing $H_t - H_{t-1}$ to be $o(1)$, which it generally isn't.

It is worth noting that Haldane [27] defined the *intensity of selection* on a population as, in the present notation,

$$\log \frac{\pi_t(\hat{\theta})}{\pi_0(\hat{\theta})}$$

where $\hat{\theta}$ is the "optimal" (i.e., most selected-for) value of θ. For us, this intensity of selection is just $R_t(\hat{\theta})/\langle R_t\rangle$ where $\hat{\theta}$ is the (or a) MLE.

APPENDIX B: VERIFICATION OF ASSUMPTIONS 5–7 FOR THE EXAMPLE

Since the X_1^∞ process is a function of the S_1^∞ process, and the latter is an aperiodic Markov chain, both are ψ-mixing (see [35, 46] for the definition of ψ-mixing and demonstrations that aperiodic Markov chains and their functions are ψ-mixing). Let $\widehat{P^{(k,t)}}$ be the empirical distribution of sequences of length k obtained from x_1^t. For a Markov chain of order k, the likelihood is a function of $\widehat{P^{(k+1,t)}}$ alone; we will use this and the ergodic properties of the data-generating process to construct sets on which the time-averaged log-likelihood converges uniformly. Doing this will involve constraining both the order of the Markov chains and their transition probabilities, and gradually relaxing the constraints.

It will simplify notation if from here on all logarithms are taken to base 2.

Pick $\epsilon > 0$ and let $k(t)$ be an increasing positive-integer-valued function of t, $k(t) \to \infty$, subject to the limit $k(t) \le \frac{\log t}{h_P+\epsilon}$, where h_P is the Shannon entropy rate of P, which direct calculation shows is 2/3. The ψ-mixing property of X_1^∞ implies [46, p. 179] that

$$P(p_{TV}(\widehat{P^{(k(t),t)}}, P^{(k(t))}) > \delta) \le \frac{\log t}{h+\epsilon} 2(n+1)^{t^{\gamma_1}} 2^{-nC_1\delta^2} \tag{34}$$

where ρ_{TV} is total variation distance, $P^{(k(t))}$ is the P's restriction to sequences of length $k(t)$, $n = \lfloor t/k(t)\rfloor - 1$, $\gamma_1 = (h_P + \epsilon/2)/(h_P + \epsilon)$ and C_1 is a positive constant specific to P (the exact value of which is not important).

The log-likelihood per observation of a Markov chain $\theta \in \Theta_k$ is

$$t^{-1} \log f_\theta(x_1^t) = t^{-1} \log f_\theta(x_1^k) + \sum_{w\in\Xi^k} \sum_{a\in\Xi} \widehat{P^{(k+1,t)}}(wa) \log f_\theta(a|w)$$

where $f_\theta(a|w)$ is of course the probability, according to θ, of producing a after seeing w. By asymptotic equipartition, this is converging a.s. to its expected value, $-h_P - h(\theta)$.

Let $z(\theta) = \max_{w,a} |\log f_\theta(a|w)|$. If $z(\theta) \le z_0$ and $\rho_{TV}(\widehat{P^{(k+1,t)}}, P^{(k+1)}) \le \delta$, then $t^{-1} \log f_\theta(x_1^t)$ is within $z_0\delta$ of $-h_P - h(\theta)$. Meanwhile, $t^{-1} \log p(x_1^t)$ is converging a.s. to $-h_P$, and again [46]

$$P(|t^{-1} \log p(X_1^t) - h_P| > \delta) \le q(t, \delta) 2^{-tC_2\delta} \tag{35}$$

for some $C_2 > 0$ and sub-exponential $q(t,\delta)$. (The details are unilluminating in the present context and thus skipped.)

Define $G(n, z_0)$ as the set of all Markov models whose order is less than or equal to $k(n) - 1$ and whose log transition probabilities do not exceed z_0, in symbols

$$G(n, z_0) = \{\theta : z(\theta) \leq z_0\} \cap \left(\bigcup_{j=1}^{k(n)-1} \Theta_j \right)$$

Combining the deviation-probability bounds 34 and 35, for all $\theta \in G(n, z_0)$

$$P\left(\left| \frac{\log R_t(\theta)}{t} - h(\theta) \right| > \delta \right) \leq \frac{\log t}{h+\epsilon} 2(n+1)^{t^{\gamma_1}} 2^{-\frac{nC_1\delta^2}{4z_0}} + q(t,\delta) 2^{-\frac{tC_2\delta}{2}} \tag{36}$$

These probabilities are clearly summable as $t \to \infty$, so by the Borel-Cantelli lemma, we have uniform almost-sure convergence of $t^{-1} \log R_t(\theta)$ to $-h(\theta)$ for all $\theta \in G(n, z_0)$.

The sets $G(n, z_0)$ eventually expand to include Markov models of arbitrarily high order, but maintain a constant bound on the transition probabilities. To relax this, let z_t be an increasing function of t, $z(t) \to \infty$, subject to $z_t \leq C_3 t^{\gamma_2}$ for positive $\gamma_2 < \gamma_1$. Then the deviation probabilities remain summable, and for each n, the convergence of $t^{-1} \log R_t(\theta)$ is still uniform on $G(n, z_n)$. Set $G_n = G(n, z_n)$, and turn to verifying the remaining assumptions.

Start with Assumption 5; take its items in reverse order. So far, the only restriction on the prior Π_0 has been that its support should be the whole of Θ, and that it should have the "Kullback-Leibler rate property", giving positive weight to every set $N_\epsilon = \{\theta : d(\theta) < \epsilon\}$. This, together with the fact that $\lim_n G_n = \Theta$, means that $h(G_n) \to h(\Theta)$, which is item (3) of the assumption. The same argument also delivers Assumption 7. Item (2), uniform convergence on each G_n, is true by construction. Finally (for this assumption), since $h(\Theta) = 0$, any $\beta > 0$ will do, and there are certainly probability measures where $\Pi_0(G_n^c) \leq \alpha \exp\{-\beta n\}$ for some $\alpha, \beta > 0$. So, Assumption 5 is satisfied.

Only Assumption 6 remains. Since Assumptions 1–3 have already been checked, we can apply Eq. 61 from the proof of Lemma 6 and see that, for any $\epsilon > 0$, for all sufficiently large t,

$$t^{-1} \log \int_{G_n} R_t(\theta) d\Pi_0 \leq -h(G_n) + \epsilon + t^{-1} \log \Pi_0(G_n) \text{ a.s.}$$

This shows that $T(G_n, \delta)$ is almost surely finite for all n and δ, but still leaves open the question of whether for every δ, $n \geq T(G_n, \delta)$ for all sufficiently large n (a.s.). Reformulating a little, the desideratum is that for each δ, with probability 1, $n < T(G_n, \delta)$ only finitely often. By the Borel-Cantelli lemma, this will happen if $\sum_n P(T(G_n, \delta) > n) \leq \infty$. However, if $T(G_n, \delta) > n$, it must be equal to some particular $t > n$, so there is a union bound:

$$\sum_n P(T(G_n, \delta) > n) \leq \sum_n \sum_{t=n+1}^{\infty} P\left(\frac{\log \int_{G_n} R_t(\theta) d\Pi_0}{t} > \delta - h(G_n)\right) \tag{37}$$

From the proof of Lemma 6 (specifically from Eqs. 15, 16 and 17), we can see that by making n large enough, the only way to have the event $t^{-1} \log \int_{G_n} R_t(\theta) d\Pi_0 > \delta - h(G_n)$ is to have $\left|t^{-1} \log R_t(\theta) - h(\theta)\right| > \delta/2$ everywhere on a positive-measure subset of G_n. But we know from Eq. 36 not only that the inner sum can be made arbitrarily small by taking n sufficiently large, but that the whole double sum is finite. So $T(G_n, \delta) > n$ only finitely often (a.s.).

ACKNOWLEDGMENTS

Portions of this work were done while at the Center for the Study of Complex Systems at the University of Michigan, supported by a grant from the James S. McDonnell Foundation, and while at the "Science et Gastronomie 2003" workshop organized by the Institut des Systèmes Complexes and the Laboratoire de l'Informatique du Paralléisme of the École Normale Supériere de Lyon. Thanks to Michel Morvan and Cristopher Moore for organizing that workshop, and to Carl Bergstrom, Anthony Brockwell, Peter Hoff, Giles Hooker, Martin Nilsson Jacobi, Jay Kadane, Kevin T. Kelly, David Krakauer, Michael Lachmann, Scott E. Page, Mark Schervish, Eric Smith, Surya Tokdar and Larry Wasserman for discussion and suggestions; and above all to Kristina Klinkner.

REFERENCES

[1] Algoet, Paul H. and Thomas M. Cover (1988). "A Sandwich Proof of the Shannon-McMillan-Breiman Theorem." *Annals of Probability*, **16**: 899–909. URL `http://projecteuclid.org/euclid.aop/1176991794`.

[2] Arora, Sanjeev, Elad Hazan and Satyen Kale (2005). "The Multiplicative Weights Update Method: a Meta Algorithm and Applications." URL `http://www.cs.princeton.edu/~arora/pubs/MWsurvey.pdf`.

[3] Badii, Remo and Antonio Politi (1997). *Complexity: Hierarchical Structures and Scaling in Physics*. Cambridge, England: Cambridge University Press.

[4] Barron, Andrew, Mark J. Schervish and Larry Wasserman (1999). "The Consistency of Posterior Distributions in Nonparametric Problems." *Annals of Statistics*, **27**: 536–561. URL `http://projecteuclid.org/euclid.aos/1018031206`.

[5] Börgers, Tilman and Rajiv Sarin (1997). "Learning Through Reinforcement and Replicator Dynamics." *Journal of Economic Theory*, **77**: 1–14.

[6] Borkar, Vivek S. (2002). "Reinforcement Learning in Markovian Evolutionary Games." *Advances in Complex Systems*, **5**: 55–72. doi:10.1142/S0219525902000535.
[7] Cesa-Bianchi, Nicolò and Gábor Lugosi (2006). *Prediction, Learning, and Games*. Cambridge, England: Cambridge University Press.
[8] Chamley, Christophe (2004). *Rational Herds: Economic Models of Social Learning*. Cambridge, England: Cambridge University Press.
[9] Choi, Taeryon and R. V. Ramamoorthi (2008). "Remarks on Consistency of Posterior Distributions." In *Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh* (Bertrand Clarke and Subhashis Ghosal, eds.), pp. 170–186. Beechwood, Ohio: Institute of Mathematical Statistics. URL `http://arxiv.org/abs/0805.3248`. doi:10.1214/074921708000000138.
[10] Crutchfield, James P. (1992). "Semantics and Thermodynamics." In *Nonlinear Modeling and Forecasting* (Martin Casdagli and Stephen Eubank, eds.), pp. 317–359. Reading, Massachusetts: Addison-Wesley.
[11] Daw, C. S., C. E. A. Finney and E. R. Tracy (2003). "A review of symbolic analysis of experimental data." *Review of Scientific Instruments*, **74**: 916–930. URL `http://www-chaos.engr.utk.edu/abs/abs-rsi2002.html`.
[12] Dębowski, Łukasz (2006). *Ergodic decomposition of excess entropy and conditional mutual information*. Tech. Rep. 993, Institute of Computer Science, Polish Academy of Sciences (IPI PAN). URL `http://www.ipipan.waw.pl/~ldebowsk/docs/raporty/ee_report.pdf`.
[13] Diaconis, Persi and David Freedman (1986). "On the Consistency of Bayes Estimates." *Annals of Statistics*, **14**: 1–26. URL `http://projecteuclid.org/euclid.aos/1176349830`.
[14] Doob, Joseph L. (1949). "Application of the theory of martingales." In *Colloques Internationaux du Centre National de la Recherche Scientifique*, vol. 13, pp. 23–27. Paris: Centre National de la Recherche Scientifique.
[15] Dynkin, E. B. (1978). "Sufficient statistics and extreme points." *Annals of Probability*, **6**: 705–730. URL `http://projecteuclid.org/euclid.aop/1176995424`.
[16] Earman, John (1992). *Bayes or Bust? A Critical Account of Bayesian Confirmation Theory*. Cambridge, Massachusetts: MIT Press.
[17] Eichelsbacher, Peter and Ayaladi Ganesh (2002). "Moderate Deviations for Bayes Posteriors." *Scandanavian Journal of Statistics*, **29**: 153–167. doi:10.1111/1467-9469.00278.
[18] Fisher, Ronald Aylmer (1958). *The Genetical Theory of Natural Selection*. New York: Dover, 2nd edn. First edition published Oxford: Clarendon Press, 1930.
[19] Fraser, Andrew M. (2008). *Hidden Markov Models and Dynamical Systems*. Philadelphia: SIAM Press.
[20] Ghosal, Subhashis, Jayanta K. Ghosh and R. V. Ramamoorthi (1999). "Consistency issues in Bayesian nonparametrics." In *Asymptotics, Nonparametrics and Time Series: A Tribute to Madan Lal Puri* (Subir Ghosh, ed.), pp. 639–667. Marcel Dekker. URL `http://www4.stat.ncsu.edu/~sghosal/papers/review.pdf`.
[21] Ghosal, Subhashis, Jayanta K. Ghosh and Aad W. van der Vaart (2000). "Convergence Rates of Posterior Distributions." *Annals of Statistics*, **28**: 500–531. URL `http://projecteuclid.org/euclid.aos/1016218228`.
[22] Ghosal, Subhashis and Yongqiang Tang (2006). "Bayesian Consistency for Markov Processes." *Sankhya*, **68**: 227–239.
[23] Ghosal, Subhashis and Aad van der Vaart (2007). "Convergence Rates of Posterior Distributions for Non-IID Observations." *Annals of Statistics*, **35**: 192–223. doi:10.1214/009053606000001172.

[24] Ghosh, J. K. and R. V. Ramamoorthi (2003). *Bayesian Nonparametrics*. New York: Springer Verlag.

[25] Gray, Robert M. (1988). *Probability, Random Processes, and Ergodic Properties*. New York: Springer-Verlag. URL `http://ee.stanford.edu/~gray/arp.html`.

[26] — (1990). *Entropy and Information Theory*. New York: Springer-Verlag. URL `http://ee.stanford.edu/~gray/it.html`.

[27] Haldane, J. B. S. (1954). "The Measurement of Natural Selection." In *Proceedings of the 9th International Congress of Genetics*, vol. 1, pp. 480–487.

[28] Hofbauer, Josef and Karl Sigmund (1998). *Evolutionary Games and Population Dynamics*. Cambridge, England: Cambridge University Press.

[29] Kallenberg, Olav (2002). *Foundations of Modern Probability*. New York: Springer-Verlag, 2nd edn.

[30] Kitchens, Bruce P. (1998). *Symbolic Dynamics: One-sided, Two-sided and Countable State Markov Shifts*. Berlin: Springer-Verlag.

[31] Knight, Frank B. (1975). "A Predictive View of Continuous Time Processes." *Annals of Probability*, **3**: 573–596. URL `http://projecteuclid.org/euclid.aop/1176996302`.

[32] Krogh, Anders and Jesper Vedelsby (1995). "Neural Network Ensembles, Cross Validation, and Active Learning." In *Advances in Neural Information Processing 7 [NIPS 1994]* (Gerald Tesauro and David Tourtetsky and Todd Leen, eds.), pp. 231–238. Cambridge, Massachusetts: MIT Press. URL `http://books.nips.cc/papers/files/nips07/0231.pdf`.

[33] Lijoi, Antonio, Igor Prünster and Stephen G. Walker (2007). "Bayesian Consistency for Stationary Models." *Econometric Theory*, **23**: 749–759. doi:10.1017/S0266466607070314.

[34] Lind, Douglas and Brian Marcus (1995). *An Introduction to Symbolic Dynamics and Coding*. Cambridge, England: Cambridge University Press.

[35] Marton, Katalin and Paul C. Shields (1994). "Entropy and the Consistent Estimation of Joint Distributions." *Annals of Probability*, **22**: 960–977. URL `http://projecteuclid.org/euclid.aop/1176988736`. Correction, *Annals of Probability*, **24** (1996): 541–545.

[36] McAllister, David A. (1999). "Some PAC-Bayesian Theorems." *Machine Learning*, **37**: 355–363.

[37] Meir, Ron (2000). "Nonparametric Time Series Prediction Through Adaptive Model Selection." *Machine Learning*, **39**: 5–34. URL `http://www.ee.technion.ac.il/~rmeir/Publications/MeirTimeSeries00.pdf`.

[38] Ornstein, Donald S. and Benjamin Weiss (1990). "How Sampling Reveals a Process." *Annals of Probability*, **18**: 905–930. URL `http://projecteuclid.org/euclid.aop/1176990729`.

[39] Page, Scott E. (2007). *The Difference: How the Power of Diveristy Creates Better Groups, Firms, Schools, and Societies*. Princeton, New Jersey: Princeton University Press.

[40] Papangelou, F. (1996). "Large Deviations and the Bayesian Estimation of Higher-Order Markov Transition Functions." *Journal of Applied Probability*, **33**: 18–27. URL `http://www.jstor.org/stable/3215260`.

[41] Ryabko, Daniil and Boris Ryabko (2008). "Testing Statistical Hypotheses About Ergodic Processes." E-print, arxiv.org, 0804.0510. URL `http://arxiv.org/abs/0804.0510`.

[42] Sato, Yuzuru and James P. Crutchfield (2003). "Coupled replicator equations for the dynamics of learning in multiagent systems." *Physical Review E*, **67**: 015206. URL

http://arxiv.org/abs/nlin.AO/0204057.

[43] Shalizi, Cosma Rohilla and James P. Crutchfield (2001). "Computational Mechanics: Pattern and Prediction, Structure and Simplicity." *Journal of Statistical Physics*, **104**: 817–879. URL http://arxiv.org/abs/cond-mat/9907176.

[44] Shalizi, Cosma Rohilla and Kristina Lisa Klinkner (2004). "Blind Construction of Optimal Nonlinear Recursive Predictors for Discrete Sequences." In *Uncertainty in Artificial Intelligence: Proceedings of the Twentieth Conference (UAI 2004)* (Max Chickering and Joseph Y. Halpern, eds.), pp. 504–511. Arlington, Virginia: AUAI Press. URL http://arxiv.org/abs/cs.LG/0406011.

[45] Shen, Xiaotong and Larry Wasserman (2001). "Rates of convergence of posterior distributions." *Annals of Statistics*, **29**: 687–714. URL http://projecteuclid.org/euclid.aos/1009210686.

[46] Shields, Paul C. (1996). *The Ergodic Theory of Discrete Sample Paths*. Providence, Rhode Island: American Mathematical Society.

[47] Strelioff, Christopher C., James P. Crutchfield and Alfred W. Hübler (2007). "Inferring Markov Chains: Bayesian Estimation, Model Comparison, Entropy Rate, and Out-of-class Modeling." *Physical Review E*, **76**: 011106. URL http://arxiv.org/math.ST/0703715. doi:10.1103/PhysRevE.76.011160.

[48] Vidyasagar, M. (2003). *Learning and Generalization: With Applications to Neural Networks*. Berlin: Springer-Verlag, 2nd edn.

[49] Walker, Stephen (2004). "New Approaches to Bayesian Consistency." *Annals of Statistics*, **32**: 2028–2043. URL http://arxiv.org/abs/math.ST/0503672. doi:10.1214/009053604000000409.

[50] Weiss, Benjamin (1973). "Subshifts of Finite Type and Sofic Systems." *Monatshefte für Mathematik*, **77**: 462–474.

[51] Xing, Yang and Bo Ranneby (2008). "Both Necessary and Sufficient Conditions for Bayesian Exponential Consistency." URL http://arxiv.org/abs/0812.1084.

Statistics Department
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA
15213-3890 USA E-mail: cshalizi@cmu.edu