Submitted to the Annals of Statistics
arXiv: arXiv:0000.0000

AN MCMC APPROACH TO EMPIRICAL BAYES INFERENCE AND BAYESIAN SENSITIVITY ANALYSIS VIA EMPIRICAL PROCESSES

BY HANI DOSS * AND YEONHEE PARK

University of Florida and MD Anderson Cancer Center

Consider a Bayesian situation in which we observe $Y \sim p_\theta$, where $\theta \in \Theta$, and we have a family $\{\nu_h,\, h \in \mathcal{H}\}$ of potential prior distributions on Θ. Let g be a real-valued function of θ, and let $I_g(h)$ be the posterior expectation of $g(\theta)$ when the prior is ν_h. We are interested in two problems: (i) selecting a particular value of h, and (ii) estimating the family of posterior expectations $\{I_g(h),\, h \in \mathcal{H}\}$. Let $m_y(h)$ be the marginal likelihood of the hyperparameter h: $m_y(h) = \int p_\theta(y)\, \nu_h(d\theta)$. The empirical Bayes estimate of h is, by definition, the value of h that maximizes $m_y(h)$. It turns out that it is typically possible to use Markov chain Monte Carlo to form point estimates for $m_y(h)$ and $I_g(h)$ for each individual h in a continuum, and also confidence intervals for $m_y(h)$ and $I_g(h)$ that are valid pointwise. However, we are interested in forming estimates, with confidence statements, of the entire families of integrals $\{m_y(h),\, h \in \mathcal{H}\}$ and $\{I_g(h),\, h \in \mathcal{H}\}$: we need estimates of the first family in order to carry out empirical Bayes inference, and we need estimates of the second family in order to do Bayesian sensitivity analysis. We establish strong consistency and functional central limit theorems for estimates of these families by using tools from empirical process theory. We give two applications, one to Latent Dirichlet Allocation, which is used in topic modelling, and the other is to a model for Bayesian variable selection in linear regression.

1. Introduction. This paper is concerned with two related problems. In the first, there is a function $B\colon \mathcal{H} \to \mathbb{R}$, where $\mathcal{H}$ is a subset of some Euclidean space, and we wish to obtain confidence sets for $\arg\max_{h\in\mathcal{H}} B(h)$. For each h, the expression for $B(h)$ is analytically intractable; however, we have at our disposal a family of functions $\{f_h,\, h \in \mathcal{H}\}$ and a sequence of random variables $\theta_1, \dots, \theta_n$ (these are iid or the initial segment of an ergodic Markov chain) such that the random function $B_n(h) := (1/n)\sum_{i=1}^n f_h(\theta_i)$ satisfies $B_n(h) \xrightarrow{\text{a.s.}} B(h)$ for each h. We are interested in how we can use B_n to form both a point estimate and a confidence set for $\arg\max_{h\in\mathcal{H}} B(h)$.

This problem appears in empirical Bayes analysis and under many forms in likelihood inference. In empirical Bayes analysis, the application that is the focus

*Supported by NSF Grant DMS-11-06395 and NIH grant P30 AG028740

MSC 2010 subject classifications: Primary 62F15, 91-08; secondary 62F12

Keywords and phrases: Donsker class, geometric ergodicity, hyperparameter selection, regenerative simulation, Latent Dirichlet Allocation model

of this paper, it arises as follows. Suppose we are in a standard Bayesian situation in which we observe a data vector Y whose distribution is P_θ (with density p_θ with respect to some dominating measure) for some $\theta \in \Theta$. We have a family of potential prior densities $\{\nu_h,\ h \in \mathcal{H}\}$, and because the hyperparameter h can have a great impact on subsequent inference, we wish to choose it carefully. Selection of h is often guided by the marginal likelihood of the data under the prior ν_h, given by

$$m_y(h) = \int p_\theta(y)\nu_h(\theta)\, d\theta, \qquad h \in \mathcal{H}. \tag{1.1}$$

By definition, the empirical Bayes choice of h is $\arg\max_h m_y(h)$. Unfortunately, analytic calculation of $m_y(h)$ is not feasible except for a few textbook examples, and estimation of $m_y(h)$ via Monte Carlo is notoriously difficult—for example, the "harmonic mean estimator" introduced by Newton and Raftery (1994) typically converges at a rate which is much slower than $n^{1/2}$ (Wolpert and Schmidler, 2012).

It is very interesting to note that if c is a constant, then the information regarding h given by the two functions $m_y(h)$ and $cm_y(h)$ is the same: the same value of h maximizes both functions, and the second derivative matrices of the logarithm of these two functions are identical. In particular, the Hessians of the logarithm of these two functions at the maximum (i.e. the observed Fisher information) are the same and, therefore, the standard point estimates and confidence regions based on $m_y(h)$ and $cm_y(h)$ are identical. This is a very useful observation because it turns out that it is usually easy to estimate the entire family $\{cm_y(h),\ h \in \mathcal{H}\}$ for a suitable choice of c. Indeed, for any $h \in \mathcal{H}$, let $\nu_{h,y}$ denote the posterior corresponding to ν_h, let h_1 be fixed but arbitrary, and suppose that $\theta_1, \dots, \theta_n$ are either independent and identically distributed according to the posterior $\nu_{h_1,y}$, or are the initial segment an ergodic Markov chain with invariant distribution $\nu_{h_1,y}$. Let $\ell_y(\theta) = p_\theta(y)$ be the likelihood function. Note that $m_y(h)$ given by (1.1) is the normalizing constant in the statement "the posterior is proportional to likelihood times the prior," i.e.

$$\nu_{h,y}(\theta) = \ell_y(\theta)\nu_h(\theta)/m_y(h). \tag{1.2}$$

We have

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} \frac{\nu_h(\theta_i)}{\nu_{h_1}(\theta_i)} &\xrightarrow{\text{a.s.}} \int \frac{\nu_h(\theta)}{\nu_{h_1}(\theta)} \nu_{h_1,y}(\theta)\, d\theta \\ &= \frac{m_y(h)}{m_y(h_1)} \int \frac{\nu_{h,y}(\theta)}{\nu_{h_1,y}(\theta)}\, \nu_{h_1,y}(\theta)\, d\theta = \frac{m_y(h)}{m_y(h_1)}, \end{aligned} \tag{1.3}$$

in which the first equality follows from (1.2) and cancellation of the likelihood. Let $B(h) = m_y(h)/m_y(h_1)$. Since $m_y(h_1)$ is a fixed constant, as noted above, the two functions $m_y(h)$ and $B(h)$ give exactly the same information about h. If we let $f_h = \nu_h/\nu_{h_1}$, then $B_n(h) = (1/n)\sum_{i=1}^n(\nu_h(\theta_i)/\nu_{h_1}(\theta_i))$—this quantity is computable, since it involves only the priors and not the posteriors—so we have precisely the situation discussed in the first paragraph of this paper. Other examples of this situation arising in frequentist inference, and in particular in missing data models, are given in Sung and Geyer (2007) and Doss and Tan (2014).

In Bayesian applications it is rare that Monte Carlo estimates of posterior quantities can be based on iid samples; in the vast majority of cases they are based on Markov chain samples, and that is the case that is the focus of this paper. We show that under suitable regularity conditions,

$$\arg\max_h B_n(h) \xrightarrow{\text{a.s.}} \arg\max_h B(h) \tag{1.4}$$

and

$$n^{1/2}(\arg\max_h B_n(h) - \arg\max_h B(h)) \xrightarrow{d} \mathcal{N}(0,\Sigma), \tag{1.5}$$

where Σ can be estimated consistently. Now, in general, almost sure convergence of $B_n(h)$ to $B(h)$ pointwise is not enough to imply that $\arg\max_h B_n(h)$ converges to $\arg\max_h B(h)$ under any mode of convergence, and in fact it is trivial to construct a counterexample in which g_n and g are deterministic functions defined on $[0,1]$, $g_n(h) \overset{n\to\infty}{\longrightarrow} g(h)$ for every $h \in [0,1]$, but $\arg\max_h g_n(h)$ does not converge to $\arg\max_h g(h)$. To obtain results (1.4) and (1.5) above, some uniformity in the convergence is needed. We establish the necessary uniform convergence and show that (1.4) and (1.5) are true under certain regularity conditions on the sequence $\theta_1, \theta_2, \ldots$, the functions f_h, and the function B. Result (1.5) enables us to obtain confidence sets for $\arg\max_h B(h)$.

The second problem we are interested in pertains to the Bayesian framework discussed earlier and is described as follows. Suppose that g is a real-valued function of θ, and consider $I_g(h) = \int g(\theta)\nu_{h,y}(\theta)\,d\theta$, the posterior expectation of $g(\theta)$ given $Y = y$, when the prior is ν_h. Suppose that $h_1 \in \mathcal{H}$ is fixed but arbitrary, and that $\theta_1, \theta_2, \ldots$ is an ergodic Markov chain with invariant distribution $\nu_{h_1,y}$. A very interesting and well-known fact, which we review in Section 2.3, is that for any $h \in \mathcal{H}$, if we define

$$w_i^{(h)} = \frac{[\nu_h(\theta_i)/\nu_{h_1}(\theta_i)]}{\sum_{l=1}^n[\nu_h(\theta_l)/\nu_{h_1}(\theta_l)]},$$

then

$$\hat{I}_g(h) = \sum_{i=1}^n g(\theta_i)w_i^{(h)} \tag{1.6}$$

is a consistent estimate of $I_g(h)$. Clearly $\hat{I}_g(h)$ is a weighted average of the $g(\theta_i)$'s. Under additional regularity conditions on the Markov chain and the function g, we even have a central limit theorem (CLT): $n^{1/2}\big(\hat{I}_g(h) - I_g(h)\big) \xrightarrow{d} \mathcal{N}(0, \sigma^2(h))$, and we can consistently estimate the limiting variance. Thus, with a single Markov chain run, using knowledge of only the priors and not the posteriors, we can estimate and form confidence intervals for $I_g(h)$ for any particular value of h. Now in Bayesian sensitivity analysis applications, we will be interested in viewing $I_g(h)$ for many values of h. For example, in prior elicitation settings, we may wish to find those aspects of the prior that have the biggest impact on the posterior, so that the focus of the effort is spent on those important aspects. We may also want to determine whether differences in the prior opinions of many experts have a significant impact on the conclusions. (For a discussion of Bayesian sensitivity analysis see Berger (1994) and Kadane and Wolfson (1998).) In these cases we will be interested in forming confidence bands for $I_g(\cdot)$ that are valid globally, as opposed to pointwise.

A common feature of the two problems we study in this paper is the need for uniformity in the convergence: to obtain confidence intervals for $\arg\max_{h\in\mathcal{H}} B(h)$ we need some uniformity in the convergence of $B_n(\cdot)$ to $B(\cdot)$, and to obtain confidence bands for $I_g(\cdot)$ we need functional CLT's for the stochastic process $\hat{I}_g(\cdot)$. Empirical process theory is a body of results that can be used to establish uniform almost sure convergence and functional CLT's in very general settings. However, the results hold only under strong regularity conditions; and these conditions are often hard to check in practical settings—indeed the results can easily be false if the conditions are not met. Empirical process theory is fundamentally based on an iid assumption, whereas in our setting, the sequence $\theta_1, \theta_2, \dots$ is a Markov chain. In this paper we show how empirical process methods can be applied to our two problems when the sequence $\theta_1, \theta_2, \dots$ is a Markov chain, and we also show how the needed regularity conditions can be established.

The rest of the paper is organized as follows. In Section 2 we state our theoretical results, the main ones—those that pertain to the Markov chain case—being as follows. Theorem 3 asserts uniform convergence of B_n to B when the sequence $\theta_1, \theta_2, \dots$ is a Harris ergodic Markov chain, under certain regularity conditions on the family $\{f_h,\, h \in \mathcal{H}\}$ (the precise details are spelled out in the statement of the theorem), and we show how these regularity conditions can be checked with relative ease in standard settings. We then give a simple result which says that under a mild regularity assumption on B, the condition $\sup_h |B_n(h) - B(h)| \xrightarrow{\text{a.s.}} 0$ entails $\arg\max_h B_n(h) \xrightarrow{\text{a.s.}} \arg\max_h B(h)$. Theorem 4 establishes that under certain regularity conditions, we have asymptotic normality of $n^{1/2}\big(\arg\max_h B_n(h) - \arg\max_h B(h)\big)$. Theorem 6 establishes almost sure uniform convergence of $\hat{I}_g(\cdot)$ to $I_g(\cdot)$, and also functional weak convergence: the process $\{n^{1/2}\big(\hat{I}_g(h) - I_g(h)\big),\, h \in \mathcal{H}\}$ converges weakly to a mean 0 Gaussian process indexed by $h \in \mathcal{H}$. We also show how this result can be used to construct confidence bands for $I_g(\cdot)$ that are valid globally. A by-product is functional weak convergence of $\{n^{1/2}(B_n(h) - B(h)),\ h \in \mathcal{H}\}$ to a mean 0 Gaussian process indexed by $h \in \mathcal{H}$, and construction of corresponding globally valid confidence bands for $B(\cdot)$. In Section 3 we give two illustrations on Bayesian models in which serious consideration needs to be given to the effect of the hyperparameter and its choice. The first is to the Latent Dirichlet Allocation topic model, where we show how our methodology can be used to do sensitivity analysis, and the second is to a model for Bayesian variable selection in linear regression, where we show how our methodology can be used to select the hyperparameter. In Doss and Park (2016) we provide the proofs of all the theorems except for Theorem 3; additionally, we show how the regularity conditions in Theorem 1 and Theorem 3 would typically be checked, and we verify these conditions in a simple setting.

2. Convergence of $B_n(\cdot)$ as a Process and Convergence of the Empirical Argmax. This section consists of three parts. Section 2.1 deals with uniform convergence of B_n for the iid case, and introduces the framework that will enable us to obtain results for the Markov chain case; this framework will be used in Section 2.1 and in the rest of the paper. Section 2.2 deals with point estimates and confidence sets for $\arg\max_h B(h)$, and Section 2.3 deals with uniform convergence and functional CLT's for estimates of posterior expectations. Throughout, uniformity refers to a class of functions indexed by $h \in \mathcal{H}$.

2.1. *Uniform Convergence of $B_n(\cdot)$.* Let Θ be a measurable subset of $\mathbb{R}^d$ for some $d \geq 1$, and let P be a probability measure on $(\Theta, \mathcal{B})$, where $\mathcal{B}$ is the Borel sigma-field on Θ. We assume that $\theta_1, \dots, \theta_n$ are independent and identically distributed according to P, and we let P_n be the empirical measure that they induce. We assume that $\mathcal{H}$ is a convex compact subset of $\mathbb{R}^k$ for some $k \geq 1$, and that for each $h \in \mathcal{H}$, $f_h \colon \Theta \to \mathbb{R}$ is measurable. The strong law of large numbers (SLLN) states that

$$\frac{1}{n}\sum_{i=1}^{n} f_h(\theta_i) \xrightarrow{\text{a.s.}} \int f_h\, dP \qquad \text{if} \int |f_h|\, dP < \infty. \tag{2.1}$$

Since we will be interested in versions of (2.1) that are uniform in h, there will exist measurability difficulties, so we have to be careful in dealing with measurability issues. Before proceeding, we review some terminology and standard facts from the theory of empirical processes. We will use the following standard empirical process notation: for a signed measure μ on Θ and a μ-integrable function $f \colon \Theta \to \mathbb{R}$, $\mu(f)$ denotes $\int f\, d\mu$. Let Q be an arbitrary probability measure on Θ, suppose that $\xi_1, \xi_2, \dots$ are independent and identically distributed according to Q, and let Q_n be the empirical measure induced by $\xi_1, \dots, \xi_n$. If $\mathcal{V}$ is a class of functions mapping Θ

to $\mathbb{R}$, and μ is a signed measure on Θ, we use the notation $\|\mu\|_{\mathcal{V}} = \sup_{v \in \mathcal{V}} |\mu(v)|$. We say that $\mathcal{V}$ is *Glivenko-Cantelli* if $\|Q_n - Q\|_{\mathcal{V}}$ converges to 0 almost surely; sometimes we will say $\mathcal{V}$ is *Q-Glivenko-Cantelli*, to emphasize the dependence on Q. Let $\mathcal{F} = \{f_h,\, h \in \mathcal{H}\}$. Our goal is to establish that $\mathcal{F}$ is P-Glivenko-Cantelli, which is exactly equivalent to the statement that the convergence in (2.1) holds uniformly in h.

The IID Case.

THEOREM 1 (Theorem 6.1 and Lemma 6.1 in Wellner (2005)) *Suppose that $\theta_1, \theta_2, \ldots$ are independent and identically distributed according to P. Suppose that $f_\cdot(\cdot)\colon \mathcal{H} \times \Theta \to \mathbb{R}$ is continuous in h for P-almost all θ. If $\sup_h |f_h|$ is measurable and satisfies $\int \sup_h |f_h| \, dP < \infty$, then the class $\mathcal{F}$ is P-Glivenko-Cantelli.*

Let $B_n(h) = (1/n) \sum_{i=1}^n f_h(\theta_i)$ and $B(h) = E_P(f_h(\theta))$ (the subscript to the expectation indicates that $\theta \sim P$). Then the conclusion of the theorem is the statement $\sup_{h \in \mathcal{H}} |B_n(h) - B(h)| \xrightarrow{\text{a.s.}} 0$.

The integrability condition $\int \sup_h |f_h| \, dP < \infty$ seems strong, and an even stronger integrability condition is imposed in Theorem 3. We discuss this issue in Remark 1 following the statement of Theorem 3, where we explain that in fact the two conditions are fairly easy to check in practice.

The next theorem also establishes that the class $\mathcal{F}$ is Glivenko-Cantelli. In the theorem, the integrability condition on $\sup_h |f_h|$ is replaced by an integrability condition on $\sup_h \|\nabla_h f_h\|$ (here, $\nabla_h f_h$ is the gradient vector of f_h with respect to h, and $\|\cdot\|$ is Euclidean norm). The condition on the gradient is sometimes easier to check. We include the theorem in part because a component of its proof is a key element in the proofs of Theorems 5 and 6 of this paper.

THEOREM 2 *Suppose that $\theta_1, \theta_2, \ldots$ are independent and identically distributed according to P, and that for each $h \in \mathcal{H}$, $\int |f_h| \, dP < \infty$. Assume also that for P-almost all $\theta \in \Theta$, $\nabla_h f_h$ exists and is continuous on $\mathcal{H}$. If $\sup_h \|\nabla_h f_h\|$ is measurable and satisfies $\int \sup_h \|\nabla_h f_h\| \, dP < \infty$, then the class $\mathcal{F}$ is P-Glivenko-Cantelli.*

The Markov Chain Case. Suppose now that the sequence $\theta_1, \theta_2, \ldots$ is a Markov chain with invariant distribution P, and that it is Harris ergodic (that is, it is irreducible, aperiodic, and Harris recurrent; see Meyn and Tweedie (1993, chapter 17) for definitions). Suppose also that $\int |f_h| \, dP < \infty$ for all $h \in \mathcal{H}$. The best way to deal with the family of averages $(1/n) \sum_{i=1}^n f_h(\theta_i)$, $h \in \mathcal{H}$, is through the use of "regenerative simulation." A *regeneration* is a random time at which a stochastic process probabilistically restarts itself; therefore, the "tours" made by the process in between such random times are iid. For example, if the stochastic process is a Markov chain on a discrete state space Θ, and if $\theta_0 \in \Theta$ is any point to which the chain returns infinitely often with probability one, then the times of return to θ_0 form a sequence of regenerations. This iid structure will enable us to establish uniform convergence of the family $(1/n)\sum_{i=1}^n f_h(\theta_i)$, $h \in \mathcal{H}$. Before we explain this, we first note that for most of the Markov chains used in MCMC algorithms, the state space is continuous, and there is no point to which the chain returns infinitely often with probability one. Fortunately, Mykland et al. (1995) provided a general technique for identifying a sequence of regeneration times $1 = \tau_0 < \tau_1 < \tau_2 < \cdots$ that is based on the construction of a *minorization condition*. This construction is reviewed at the end of this subsection, and gives rise to regeneration times with the property that

$$E(\tau_r - \tau_{r-1}) < \infty. \tag{2.2}$$

Suppose now that there exists a regeneration sequence $1 = \tau_0 < \tau_1 < \tau_2 < \cdots$ which satisfies (2.2). Such a Markov chain will be called regenerative. For any $h \in \mathcal{H}$, consider $(1/n)\sum_{i=1}^n f_h(\theta_i)$. Let

$$S_r^{(h)} = \sum_{i=\tau_{r-1}}^{\tau_r - 1} f_h(\theta_i), \qquad r = 1, 2, \ldots \tag{2.3}$$

be the sum of f_h over the r^{th} tour. Also, let $N_r = \tau_r - \tau_{r-1}$, $r = 1, 2, \ldots$, denote the length of the r^{th} tour. The N_r's do not involve h. Note that the pairs $\{(N_r, S_r^{(h)})\}_{r=1}^\infty$ are iid. If we run the chain for R regenerations, then the total number of cycles is given by

$$n = \sum_{r=1}^R N_r = \tau_R.$$

Also, $\sum_{i=1}^n f_h(\theta_i) = \sum_{r=1}^R S_r^{(h)}$. We have

$$E_P(f_h(\theta)) \xleftarrow{\text{a.s.}} \frac{\sum_{i=1}^n f_h(\theta_i)}{n} = \frac{\sum_{r=1}^R S_r^{(h)}}{\sum_{r=1}^R N_r} = \frac{(\sum_{r=1}^R S_r^{(h)})/R}{(\sum_{r=1}^R N_r)/R} \xrightarrow{\text{a.s.}} \frac{E(S_1^{(h)})}{E(N_1)}. \tag{2.4}$$

In (2.4), the convergence statement on the left follows from Harris ergodicity of the chain. The convergence statement on the right follows from two applications of the SLLN: By (2.2), $(1/R)\sum_{r=1}^R N_r \xrightarrow{\text{a.s.}} E(N_1)$ and this, together with the convergence statement on the left, entails convergence of $(1/R)\sum_{r=1}^R S_r^{(h)}$. The SLLN then implies that $E(|S_1^{(h)}|) < \infty$ (if $E(|S_1^{(h)}|) = \infty$ then the SLLN implies that $\limsup(1/R)\sum_{r=1}^R S_r^{(h)} = \infty$ with probability one). We conclude that

$E(S_1^{(h)}) = E_P(f_h(\theta))E(N_1)$. Note that continuity in h of $S_1^{(h)}$ for almost all sequences $\theta_1, \theta_2, \ldots$ follows from continuity in h of f_h for almost all $\theta \in \Theta$, since with probability one, $S_1^{(h)}$ is a finite sum. Suppose in addition that $\sup_h |S_1^{(h)}|$ is measurable and satisfies $E\bigl(\sup_h |S_1^{(h)}|\bigr) < \infty$. Then by Theorem 1 we have $\sup_h \bigl|(\sum_{r=1}^R S_r^{(h)})/R - E(S_1^{(h)})\bigr| \xrightarrow{\text{a.s.}} 0$. Since $(\sum_{r=1}^R N_r)/R \xrightarrow{\text{a.s.}} E(N_1)$, we obtain

$$\sup_h \left| \frac{(\sum_{r=1}^R S_r^{(h)})/R}{(\sum_{r=1}^R N_r)/R} - \frac{E(S_1^{(h)})}{E(N_1)} \right| \xrightarrow{\text{a.s.}} 0,$$

i.e.

$$\sup_h \left| \frac{\sum_{i=1}^n f_h(\theta_i)}{n} - E_P(f_h(\theta)) \right| \xrightarrow{\text{a.s.}} 0. \tag{2.5}$$

We summarize this in the following theorem.

THEOREM 3 *Suppose that $\theta_1, \theta_2, \ldots$ is a Harris ergodic Markov chain with invariant distribution P for which there exists a regeneration sequence $1 = \tau_0 < \tau_1 < \tau_2 < \cdots$ satisfying $E(\tau_1 - \tau_0) < \infty$. Suppose also that $f_{\cdot}(\cdot)\colon \mathcal{H} \times \Theta \to \mathbb{R}$ is continuous in h for P-almost all θ. For each $h \in \mathcal{H}$, let $S_r^{(h)}$, $r = 1, 2, \ldots$ be defined by (2.3). If $\sup_h |S_1^{(h)}|$ is measurable and satisfies $E\bigl(\sup_h |S_1^{(h)}|\bigr) < \infty$, then (2.5) holds.*

REMARK 1 We now discuss the integrability condition $E\bigl(\sup_h |S_1^{(h)}|\bigr) < \infty$, and our discussion encompasses the weaker condition $\int \sup_h |f_h| \, dP < \infty$ assumed in Theorem 1. Suppose that $\int |f_h| \, dP < \infty$ for all $h \in \mathcal{H}$. In Doss and Park (2016) we show that, because $\mathcal{H}$ is assumed to be compact, it is often possible to prove that for some $d \geq 1$,

$$\begin{gathered} \text{there exist } h_1, \ldots, h_d \in \mathcal{H} \text{ and constants } c_1, \ldots, c_d \text{ such that} \\ \sup_h |f_h(\theta)| \leq \sum_{j=1}^d c_j |f_{h_j}(\theta)| \qquad \text{for all } \theta \in \Theta. \end{gathered} \tag{2.6}$$

In this case, since $|S_1^{(h)}| \leq \sum_{i=\tau_0}^{\tau_1 - 1} |f_h(\theta_i)|$, we obtain

$$\sup_h |S_1^{(h)}| \leq \sum_{i=\tau_0}^{\tau_1 - 1} \sup_h |f_h(\theta_i)| \leq \sum_{i=\tau_0}^{\tau_1 - 1} \sum_{j=1}^d c_j |f_{h_j}(\theta_i)|.$$

Hence,

$$E\Bigl(\sup_h |S_1^{(h)}|\Bigr) \leq \sum_{j=1}^d E\left(\sum_{i=\tau_0}^{\tau_1 - 1} c_j |f_{h_j}(\theta_i)| \right) = \sum_{j=1}^d c_j E_P(|f_{h_j}(\theta)|) E(N_1),$$

which is finite. Thus, checking that $E\big(\sup_h |S_1^{(h)}|\big) < \infty$ reduces to establishing (2.6). In Doss and Park (2016) we consider the Bayesian framework discussed in Section 1, in which $f_h = \nu_h/\nu_{h_*}$, where $\{\nu_h,\, h \in \mathcal{H}\}$ is a family of priors, and $P = \nu_{h_*,y}$, the posterior distribution corresponding to the prior ν_{h_*}, where $h_* \in \mathcal{H}$ is fixed. We show that if $\{\nu_h,\, h \in \mathcal{H}\}$ is an exponential family, then condition (2.6) holds. Therefore, the integrability condition $E\big(\sup_h |S_1^{(h)}|\big) < \infty$ is satisfied in a large class of examples. Moreover, the method we use for establishing (2.6) can be applied to other examples as well.

REMARK 2 The idea to transform results for the iid case to the Markov chain case via regeneration has been around for many decades. Levental (1988) also obtained a Glivenko-Cantelli theorem for the Markov chain setting. In essence, the difference between his approach and ours is that his starting point is a Glivenko-Cantelli theorem for the iid case which requires a condition involving the minimum number of balls of radius ϵ in $L_1(P)$ that are needed to cover $\mathcal{F}$—he is using metric entropy. This condition is very hard to check. By contrast, our starting point is a Glivenko-Cantelli theorem for the iid case which is based on bracketing entropy—in brief, the main regularity condition is implied by the continuity condition in Theorem 3. This continuity condition is trivial to verify: the parametric families that we are working with in our Bayesian setting satisfy it automatically.

The Minorization Construction. We now describe a minorization condition that can sometimes be used to construct regeneration sequences. Let $K_\theta(A)$ be the transition function for the Markov chain $\theta_1, \theta_2, \ldots$. The construction described in Mykland et al. (1995) requires the existence of a function $s\colon \Theta \to [0,1)$, whose expectation with respect to P is strictly positive, and a probability measure Q on $(\Theta, \mathcal{B})$, such that K satisfies

$$K_\theta(A) \geq s(\theta)Q(A) \qquad \text{for all } \theta \in \Theta \text{ and } A \in \mathcal{B}. \tag{2.7}$$

This is called a minorization condition and, as we describe below, it can be used to introduce regenerations into the Markov chain driven by K. Define the Markov transition function $G.(\cdot)$ by

$$G_\theta(A) = \frac{K_\theta(A) - s(\theta)Q(A)}{1 - s(\theta)}.$$

Note that for fixed $\theta \in \Theta$, G_θ is a probability measure. We may therefore write

$$K_\theta = s(\theta)Q + (1 - s(\theta))G_\theta,$$

which gives a representation of K_θ as a mixture of two probability measures, Q and G_θ. This provides an alternative method of simulating from K. Suppose that the current state of the chain is θ_n. We generate $\delta_n \sim \text{Bernoulli}(s(\theta_n))$. If $\delta_n = 1$, we draw $\theta_{n+1} \sim Q$; otherwise, we draw $\theta_{n+1} \sim G_{\theta_n}$. Note that if $\delta_n = 1$, the next state of the chain is drawn from Q, which does not depend on the current state. Hence the chain "forgets" the current state and we have a regeneration. To be more specific, suppose we start the Markov chain with $\theta_1 \sim Q$ and then use the method described above to simulate the chain. Each time $\delta_n = 1$, we have $\theta_{n+1} \sim Q$ and the process stochastically restarts itself; that is, the process regenerates. Mykland et al. (1995) provided a very widely applicable method, the so-called "distinguished point technique", for constructing a pair (s, Q) that can be used to form a minorization scheme which satisfies (2.2).

For any fixed $h \in \mathcal{H}$, consider now the expression

$$\frac{(\sum_{r=1}^R S_r^{(h)})/R}{(\sum_{r=1}^R N_r)/R}$$

in (2.4). The bivariate CLT gives

$$R^{1/2} \begin{pmatrix} (\sum_{r=1}^R S_r^{(h)})/R - E_P(f_h(\theta))E(N_1) \\ (\sum_{r=1}^R N_r)/R - E(N_1) \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, \Sigma_h), \tag{2.8}$$

where $\Sigma_h = \text{Cov}\big((S_1^{(h)}, N_1)^\top\big)$. (We have ignored the moment conditions on $S_1^{(h)}$ and N_1 that are needed, but we will return to these conditions in Section 2.3, where we give a rigorous development of a functional version of the CLT (2.8), in which the left side of (2.8) is viewed as a process in h.) The delta method applied to the function $g(x, y) = x/y$ gives the CLT

$$R^{1/2} \left(\frac{(\sum_{r=1}^R S_r^{(h)})/R}{(\sum_{r=1}^R N_r)/R} - E_P(f_h(\theta)) \right) \xrightarrow{d} \mathcal{N}(0, \sigma_h^2),$$

where $\sigma_h^2 = (\nabla g)^\top \Sigma_h \nabla g$ (and ∇g is evaluated at the vector of means in (2.8)). Moreover, σ_h^2 can be estimated in a simple manner using a plug-in estimate. Whether or not this method gives estimates of variance that are useful in the practical sense depends on whether or not the minorization condition we construct yields regenerations which are sufficiently frequent. Successful constructions of minorization conditions have been developed for widely used chains in many papers (we mention in particular Mykland et al. (1995), Roy and Hobert (2007), Tan and Hobert (2009), and Doss et al. (2014)); nevertheless, successful construction of a minorization condition is the exception rather than the norm. In this context, we point out that here regenerative simulation is notable primarily as a device that enables us to prove the theoretical results in the present paper and to arrive at informative expressions for asymptotic variances, but it may be possible to estimate these variances by other methods; this point is discussed further in Section 2.2.

REMARK 3 The main regularity assumption in Theorem 3 is the condition $E(\sup_h |S_1^{(h)}|) < \infty$. Without giving the details, we mention that in analogy with Theorem 2, it is possible to give a version of Theorem 3 in which this condition is replaced with the condition $E(\sup_h \|\nabla_h S_1^{(h)}\|) < \infty$.

2.2. *A Consistent Estimator and Confidence Sets for* $\arg\max_h B(h)$. This section pertains to $\arg\max_h B_n(h)$ as an estimator of $\arg\max_h B(h)$. After establishing that (2.5) entails that $\arg\max_h B_n(h)$ is consistent, we show that under additional regularity conditions, (i) $n^{1/2}(\arg\max_h B_n(h) - \arg\max_h B(h))$ is asymptotically normal, and (ii) we can consistently estimate the asymptotic variance. Results (i) and (ii) enable us to form asymptotically valid confidence sets for $\arg\max_h B(h)$.

LEMMA 1 *Suppose that* H *is a compact subset of Euclidean space, and let* f_n, $n = 1, 2, \dots$ *and* f *be deterministic real-valued functions defined on* H. *Suppose further that* f *is continuous and has a unique maximizer, and that for each* n *the maximizer of* f_n *exists and is unique. If* f_n *converges to* f *uniformly on* H, *then the maximizer of* f_n *converges to the maximizer of* f.

The proof of Lemma 1 is routine and is given in Doss and Park (2016). Consider now $B_n(h) = (1/n)\sum_{i=1}^n f_h(\theta_i)$ and $B(h) = E_P(f_h(\theta))$. By Lemma 1, if B is continuous and its maximizer is unique, then $\sup_h |B_n(h) - B(h)| \xrightarrow{\text{a.s.}} 0$ implies $\arg\max_h B_n(h) \xrightarrow{\text{a.s.}} \arg\max_h B(h)$. Thus, under continuity of B and uniqueness of its maximizer, any conditions that imply (2.5)—in particular the conditions of Theorems 1, 2, or 3—are also conditions that imply strong consistency of $\arg\max_h B_n(h)$ as an estimator of $\arg\max_h B(h)$.

Before stating the next theorem, we need to set some notation and assumptions. We assume that each of B and B_n, $n = 1, 2, \dots$ has a unique maximizer, and we denote $h_0 = \arg\max_h B(h)$ and $h_n = \arg\max_h B_n(h)$. For a function $g\colon \mathcal{H} \to \mathbb{R}$, $\nabla_h g(h)$ denotes the gradient vector and $\nabla_h^2 g(h)$ denotes the Hessian matrix. We will assume that for every θ, $\nabla_h f_h(\theta)$ and $\nabla_h^2 f_h(\theta)$ exist and are continuous for all h. Recall that $S_r^{(h)}$ is defined by (2.3). The Markov chain will be run for R regenerations, and in the asymptotic results below, $R \to \infty$. We will use the notation $\bar{N} = (\sum_{r=1}^R N_r)/R$, $\bar{S}^{(h)} = (\sum_{r=1}^R S_r^{(h)})/R$, $\nabla_h \bar{S}^{(h)} = (\sum_{r=1}^R \nabla_h S_r^{(h)})/R$, etc. For almost any realization $\theta_1, \theta_2, \dots$, the random variable $S_r^{(h)}$ is a finite sum, and therefore $\nabla_h S_r^{(h)} = \sum_{i=\tau_{r-1}}^{\tau_r - 1} \nabla_h f_h(\theta_i)$. Similarly, $\nabla_h^2 S_r^{(h)} = \sum_{i=\tau_{r-1}}^{\tau_r - 1} \nabla_h^2 f_h(\theta_i)$. We will assume that the family $\{f_h,\ h \in \mathcal{H}\}$ is such that the interchange of the order of integration and either first or second order differentiation is permissible, i.e.

$$\nabla_h \int f_h \, dP = \int \nabla_h f_h \, dP \qquad \text{and} \qquad \nabla_h^2 \int f_h \, dP = \int \nabla_h^2 f_h \, dP. \tag{2.9}$$

For $h \in \mathcal{H}$, let

$$J(h) = \nabla_h^2 B(h), \qquad J_n(h) = \nabla_h^2 B_n(h),$$

$$\tau^2(h) = [E(N_1)]^{-2} E\Big(\big[\nabla_h S_1^{(h)} - N_1 E_P(\nabla_h f_h(\theta))\big] \big[\nabla_h S_1^{(h)} - N_1 E_P(\nabla_h f_h(\theta))\big]^\top \Big),$$

and

$$\tau_n^2(h) = \frac{1}{R \bar{N}^2} \sum_{r=1}^R \big(\nabla_h S_r^{(h)} - N_r \nabla_h \bar{S}^{(h)} / \bar{N}\big) \big(\nabla_h S_r^{(h)} - N_r \nabla_h \bar{S}^{(h)} / \bar{N}\big)^\top.$$

Suppose that $X_1, X_2, \ldots$ is a Markov chain on the measurable space $(\mathsf{X}, \mathcal{B})$ and has π as an invariant probability measure. Let $K^n(x, A)$ be the n-step Markov transition function. Recall that the chain is called *geometrically ergodic* if there exist a constant $c \in [0, 1)$ and a function $M \colon \mathsf{X} \to [0, \infty)$ such that for $n = 1, 2, \ldots,$

$$\sup_{A \in \mathcal{B}} |K^n(x, A) - \pi(A)| \le M(x) c^n \qquad \text{for all } x \in \mathsf{X}.$$

If $Q(\theta)$ is a $k \times k$ matrix, then a statement of the sort $E(|Q(\theta)|) < \infty$ will mean $E(|Q_{i,j}(\theta)|) < \infty$ for $i, j = 1, \ldots, k$. We will refer to the following conditions.

A1 The chain $\{\theta_i\}_{i=0}^\infty$ is geometrically ergodic.
A2 For every $h \in \mathcal{H}$, there exists $\epsilon > 0$ such that $E_P\big(\|\nabla_h f_h(\theta)\|^{2+\epsilon}\big) < \infty$.
A3 The function B is twice continuously differentiable and the $k \times k$ matrix $J(h_0)$ is nonsingular.
A4 $\sup_h |S_1^{(h)}|$ is measurable and $E\big(\sup_h |S_1^{(h)}|\big) < \infty$.
A5 $\sup_h |\nabla_h^2 S_1^{(h)}|$ is measurable and $E\big(\sup_h |\nabla_h^2 S_1^{(h)}|\big) < \infty$.
A6 $\sup_h |\nabla_h f_h|$ is measurable and $E\big(\sup_h |\nabla_h f_h|\big) < \infty$.
A7 $\big(\sup_h |\nabla_h S_1^{(h)}|\big)\big(\sup_h |\nabla_h S_1^{(h)}|\big)^\top$ is measurable and has finite expectation.

THEOREM 4 *Suppose that $\theta_1, \theta_2, \ldots$ is a regenerative Markov chain with invariant distribution P. Let*

$$v^2 = J(h_0)^{-1} \tau^2(h_0) J(h_0)^{-1}. \tag{2.10}$$

1. *Under A1–A5*

$$R^{1/2}(h_n - h_0) \xrightarrow{d} \mathcal{N}(0, v^2) \qquad \text{as } R \to \infty, \tag{2.11}$$

and consequently

$$n^{1/2}(h_n - h_0) \xrightarrow{d} \mathcal{N}\big(0, E(N_1) v^2\big) \qquad \text{as } R \to \infty. \tag{2.12}$$

2. *Under A1–A7, for large R the matrix $J_n(h_n)$ is invertible, and the variance estimate*
$$v_n^2 = \big[J_n(h_n)\big]^{-1}\tau_n^2(h_n)\big[J_n(h_n)\big]^{-1}$$
is a strongly consistent estimate of v^2.

REMARK 4 In the expression for the asymptotic variance given by (2.10), the term $\tau^2(h_0)$ is the variance of a certain function of the Markov chain, and the term $J(h_0)^{-1}$ measures the inverse of the curvature of B at its maximum (B is a deterministic function and does not involve the Markov chain): the flatter the surface B at its maximum, the higher is the asymptotic variance.

REMARK 5 The integrability condition in Assumption A4 was discussed in Remark 1, where we showed that it is satisfied whenever there exist $h_1, \dots, h_d \in \mathcal{H}$ such that $\sup_h |f_h(\theta)| \le \sum_{j=1}^d |f_{h_j}(\theta)|$ for all $\theta \in \Theta$ (cf. (2.6), in which without loss of generality we take the constants c_j to be equal to 1.) The integrability conditions in A5–A7 are satisfied under (2.13) and (2.14) below, which are very similar to (2.6). To make our explanation notationally less cumbersome and easier to follow, we will assume that $\dim(\mathcal{H}) = 1$, so that $\nabla_h S_1^{(h)}$, $\nabla_h f_h(\theta)$, $\nabla_h^2 S_1^{(h)}$, and $\nabla_h^2 f_h(\theta)$ are all scalars. Assume that there exist $h_1, \dots, h_d \in \mathcal{H}$ and constants $c_1, \dots, c_d$ such that

$$\sup_h |\nabla_h f_h(\theta)| \le \sum_{j=1}^d c_j |\nabla_h f_{h_j}(\theta)| \qquad \text{for all } \theta \in \Theta, \tag{2.13}$$

$$\sup_h |\nabla_h^2 f_h(\theta)| \le \sum_{j=1}^d c_j |\nabla_h^2 f_{h_j}(\theta)| \qquad \text{for all } \theta \in \Theta. \tag{2.14}$$

The integrability condition in A5, $E\big(\sup_h |\nabla_h^2 S_1^{(h)}|\big) < \infty$, follows from (2.14) using an argument identical to the one we used to show that the integrability condition in A4 follows from (2.6). Clearly, A6 follows immediately from (2.13).

We now deal with A7 and consider $\big(\sup_h |\nabla_h S_1^{(h)}|\big)^2 = \sup_h \big(\nabla_h S_1^{(h)}\big)^2$. Let $F(\theta) = \sum_{j=1}^d c_j |\nabla_h f_{h_j}(\theta)|$, and let $\mathcal{T}_1$ denote the set of indices that comprise the first tour. Since $\nabla_h S_1^{(h)} = \sum_{i \in \mathcal{T}_1} \nabla_h f_h(\theta_i)$, we have

$$|\nabla_h S_1^{(h)}| \le \sum_{i \in \mathcal{T}_1} |\nabla_h f_h(\theta_i)| \le \sum_{i \in \mathcal{T}_1} F(\theta_i),$$

where the second inequality is from (2.13). Therefore $\big(\nabla_h S_1^{(h)}\big)^2 \le \big(\sum_{i \in \mathcal{T}_1} F(\theta_i)\big)^2$, and hence

$$\sup_h \big(\nabla_h S_1^{(h)}\big)^2 \le \big(\textstyle\sum_{i \in \mathcal{T}_1} F(\theta_i)\big)^2. \tag{2.15}$$

Now by A2 and the Minkowski inequality, $E_P\big(F^{2+\epsilon}(\theta)\big) < \infty$. This integrability condition, together with geometric ergodicity of the chain (cf. A1), enables us to apply Theorem 2 of Hobert et al. (2002) to conclude that $E\big[\big(\sum_{i\in\mathcal{T}_1} F(\theta_i)\big)^2\big] < \infty$ which, by (2.15), implies that $E\big[\sup_h \big(\nabla_h S_1^{(h)}\big)^2\big] < \infty$, which is the integrability condition in A7.

REMARK 6 To see why convergence statement (2.12) is a consequence of (2.11), note that $n = \sum_{r=1}^R N_r$, so $n/R = \big(\sum_{r=1}^R N_r\big)/R \xrightarrow{\text{a.s.}} E(N_1)$. So from (2.11) and Slutsky's theorem, we have $(n/R)^{1/2} R^{1/2}(h_n - h_0) \xrightarrow{d} \mathcal{N}(0, E(N_1)v^2)$, which is statement (2.12).

REMARK 7 We now step back and put Theorem 4 in the context of frequentist inference. We do not require that the number of components of our data vector Y goes to infinity, or even that the components are iid. We observe $Y = y$, which induces a marginal likelihood surface $m_y(\cdot)$, and Theorem 4 pertains to estimation of this surface and its argmax, with the asymptotics referring to the Markov chain length n going to infinity. In this regard, it is natural to ask what are the frequentist properties of inference based on this argmax. A very general result, known as the Bernstein-von Mises Theorem, asserts that under certain regularity conditions, if $Y_1, Y_2, \ldots$ are iid with distribution Q_{θ_0}, and if $\hat{\theta}_m$ is the maximum likelihood estimate of θ based on $Y_{(m)} = (Y_1, \ldots, Y_m)$, then for any $h \in \mathcal{H}$, $\big\|\nu_{h,y_{(m)}} - \phi_{\hat{\theta}_m, i^{-1}(\theta_0)/m}\big\|_{\text{TV}} \xrightarrow{m\to\infty} 0$, $[Q_{\theta_0}]$-a.s. Here, $\phi_{a,V}$ denotes the normal distribution with mean vector a and covariance matrix V, $i(\theta)$ is the Fisher information at θ, and the subscript TV denotes total variation norm. In particular, the usual Bayesian 95% credible region coincides with the usual 95% confidence region, and therefore has asymptotic frequentist coverage probability equal to .95. Theorem 1 of Petrone et al. (2014) goes further, and states that the Bernstein-von Mises Theorem holds when we use h_0, the maximum marginal likelihood estimate of h. There are regularity conditions; see Petrone et al. (2014), which also contains references for precise statements of the Bernstein-von Mises Theorem. To conclude, if n is sufficiently large, 95% credible sets based on $\nu_{h_n,y_{(m)}}$ have asymptotic frequentist coverage probability equal to .95.

We now discuss the role of regenerative simulation in our development. Broadly speaking, the *existence* of regenerative sequences is guaranteed under very general conditions—here we note not only the distinguished point technique of Mykland et al. (1995) mentioned earlier, but also the fact that for any chain satisfying our minimal regularity condition of Harris ergodicity, there exists a $j \geq 1$ such that there is a minorizing pair (s, Q) for the j-step Markov transition function K^j (Meyn and Tweedie, 1993, Section 5.2). However, it is often very difficult to construct a *useful* minorization condition, i.e. one that gives rise to regenerations that are frequent enough so that law of large numbers and CLT approximations are valid for reasonable sample sizes. If we do succeed in obtaining a useful regeneration sequence, then we can estimate variances and construct confidence sets using the estimate given in Part 2 of Theorem 4, and it is widely recognized that estimation of variances using regeneration—when it is feasible—outperforms estimation using other methodologies (Flegal and Jones, 2010). Additionally, it has the advantage that because we start the chain at a regeneration point (i.e. $\theta_1 \sim Q$), the issue of burn-in does not even exist.

It is very interesting to note that we have used regenerative simulation in a theoretical manner: our proof of asymptotic normality of $n^{1/2}(h_n - h_0)$ (see (2.12)) requires only the existence of a regeneration sequence, and does not require that we go through a laborious trial and error process to construct one that is useful in the practical sense. Very briefly, to obtain asymptotic results regarding h_n, we need uniformity in the convergence of B_n to B. Empirical process theory gives us results on uniformity, but only in the iid setting, and regenerative simulation bridges the gap between the Markov chain setting and the iid setting. Once we have established the asymptotic normality of $n^{1/2}(h_n - h_0)$, we are free to estimate the asymptotic variance and form confidence sets using other methods, for example batching, which we now discuss.

Batching is implemented by breaking up the sequence $\theta_1, \dots, \theta_n$ into M consecutive pieces of equal lengths called batches. For $m = 1, \dots, M$, batch m is used to produce an estimate $h_n^{[m]}$ in the obvious way. If M is fixed, then under the regularity conditions of Theorem 4, (2.12) states that for each m, $(n/M)^{1/2}(h_n^{[m]} - h_0) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$, where $\sigma^2 = E(N_1)v^2$. If the batch length is large enough relative to the "mixing time" of the chain, then the $h_n^{[m]}$'s are approximately independent. If the independence assumption was exactly true rather than approximately true, then the sample variance of $(n/M)^{1/2}h_n^{[1]}, \dots, (n/M)^{1/2}h_n^{[M]}$ would be a valid estimator of σ^2. Standard theoretical results regarding batching deal with the situation in which g is a P-integrable function, and the Markov chain $\theta_1, \dots, \theta_n$ is used to estimate $\int g\,dP$ via $(1/n)\sum_{i=1}^n g(\theta_i)$. These results, which assume that $n^{1/2}\big((1/n)\sum_{i=1}^n g(\theta_i) - \int g\,dP\big) \xrightarrow{d} \mathcal{N}(0, \sigma^2(g))$, state that under regularity conditions which include $M \to \infty$ at a certain rate, the batch-based estimate of $\sigma^2(g)$ is strongly consistent; see Flegal et al. (2008) and also Jones et al. (2006), who recommend using $M = n^{1/2}$. Our situation is different in that our estimate h_n is not an average, but is the argmax of a function based on $\theta_1, \dots, \theta_n$. Nevertheless, the method applies, with the minor modification that when we form the "sample variance," the centering value is based on h_n rather than on the average of the $h_n^{[m]}$'s. As is clear from the description above, batch-based estimates of variance are very easy to program. However, it is generally acknowledged that they are outperformed by estimates based on regeneration or spectral methods.

2.3. *Convergence of Estimate of Posterior Expectation.* This section concerns the Bayesian framework discussed earlier, in which $\{\nu_h,\, h \in \mathcal{H}\}$ is a family of prior densities on θ; for each h, $\nu_{h,y}$ is the posterior corresponding to ν_h; $h_1 \in \mathcal{H}$ is fixed but arbitrary, and $\theta_1, \theta_2, \dots$ is an ergodic Markov chain with invariant distribution $\nu_{h_1,y}$. Suppose that g is a real-valued function of θ and consider $I_g(h) = \int g(\theta)\nu_{h,y}(\theta)\,d\theta$, the posterior expectation of $g(\theta)$ given $Y = y$, when the prior is ν_h. We have

$$\frac{1}{n}\sum_{i=1}^{n} g(\theta_i)\frac{\nu_h(\theta_i)}{\nu_{h_1}(\theta_i)} \xrightarrow{\text{a.s.}} \int g(\theta)\frac{\nu_h(\theta)}{\nu_{h_1}(\theta)}\nu_{h_1,y}(\theta)\,d\theta = \frac{m_y(h)}{m_y(h_1)} I_g(h), \tag{2.16}$$

in which the first equality follows from (1.2) and cancellation of the likelihood. Therefore,

$$\hat{I}_g(h) := \frac{(1/n)\sum_{i=1}^{n} g(\theta_i)[\nu_h(\theta_i)/\nu_{h_1}(\theta_i)]}{(1/n)\sum_{i=1}^{n}[\nu_h(\theta_i)/\nu_{h_1}(\theta_i)]} \xrightarrow{\text{a.s.}} \frac{[m_y(h)/m_y(h_1)]I_g(h)}{m_y(h)/m_y(h_1)} = I_g(h), \tag{2.17}$$

where the convergence of the numerator and the denominator in the expression for $\hat{I}_g(h)$ follow from (2.16) and (1.3), respectively. In the original expression given in (1.6), $\hat{I}_g(h)$ is a weighted average of the $g(\theta_i)$'s (with weights all equal to $1/n$ if $\nu_h = \nu_{h_1}$, and becoming more disparate as ν_h and ν_{h_1} become more dis-similar). The definition of $\hat{I}_g(h)$ given in (2.17) clearly matches the original expression, so we see that $\hat{I}_g(h)$ may be represented either as a weighted average or as a ratio of two ordinary averages. To establish almost sure uniform convergence and functional weak convergence results for $\hat{I}_g(h)$, we will work with the latter representation, because doing so will enable us to use tools from empirical process theory. With this in mind, recall that in the present framework $f_h = \nu_h/\nu_{h_1}$. We will work with the classes of functions $\mathcal{F} = \{f_h,\, h \in \mathcal{H}\}$ and $\mathcal{G} = \{gf_h,\, h \in \mathcal{H}\}$. We will later assume that the sequence $\theta_1, \theta_2, \dots$ is a Markov chain satisfying certain conditions, and Theorem 6 pertains to that case; however, in order to give an overview of our results, it is convenient to first assume that the θ_i's form an iid sequence: $\theta_i \overset{\text{iid}}{\sim} P := \nu_{h_1,y}$. Recall that P_n is the empirical measure that gives mass $1/n$ to each of $\theta_1, \dots, \theta_n$, and that for a signed measure μ and a function f, $\mu(f)$ denotes $\int f\,d\mu$. In the present specialized Bayesian context, $f_h \geq 0$; thus the $L_1(P)$ norm of f_h is simply $\int f_h\,dP$. Our goal is to establish that under certain conditions:

1. We have the Glivenko-Cantelli results

$$\sup_{h\in\mathcal{H}} |(P_n - P)(f_h)| \xrightarrow{\text{a.s.}} 0 \qquad \text{and} \qquad \sup_{h\in\mathcal{H}} |(P_n - P)(gf_h)| \xrightarrow{\text{a.s.}} 0.$$

2. We have the "Donsker results"

$$n^{1/2}(P_n - P)(f_\cdot) \xrightarrow{d} \mathbb{F}(\cdot) \qquad \text{and} \qquad n^{1/2}(P_n - P)(gf_\cdot) \xrightarrow{d} \mathbb{G}(\cdot), \tag{2.18}$$

where $\mathbb{F}$ and $\mathbb{G}$ are mean 0 Gaussian processes indexed by $\mathcal{H}$.

By applying the delta method to the function $q(u, v) = u/v$, we then obtain the Glivenko-Cantelli and Donsker results

3.
$$\sup_{h\in\mathcal{H}} |\hat{I}_g(h) - I_g(h)| \xrightarrow{\text{a.s.}} 0,$$

4.
$$n^{1/2}(\hat{I}_g(\cdot) - I_g(\cdot)) \xrightarrow{d} \mathbb{I}_g(\cdot), \tag{2.19}$$
where $\mathbb{I}_g$ is a mean 0 Gaussian process indexed by $\mathcal{H}$.

We now give some definitions we will need in order to explain what is meant by (2.18) and (2.19). Define $X_n = n^{1/2}(P_n - P)$. Let $\mathcal{V}$ be any set of real-valued functions defined on Θ and let $l^\infty(\mathcal{V})$ denote the space of bounded functions from $\mathcal{V}$ to $\mathbb{R}$ equipped with the supremum norm. Assume that

$$\sup_{V\in\mathcal{V}} |V(\theta) - P(V)| < \infty \qquad \text{for every } \theta \in \Theta.$$

Under this condition the empirical process $\{X_n(V),\, V \in \mathcal{V}\}$ can be viewed as a map from Θ^n into $l^\infty(\mathcal{V})$. Any measurable function $Z\colon \Theta^n \to l^\infty(\mathcal{V})$ induces a distribution on $l^\infty(\mathcal{V})$. Although the functions we will be working with will in general be measurable, in order to properly state the relevant definitions and theorems from empirical process theory, in our definitions we will deal with functions which are not necessarily measurable. For an arbitrary map M from an arbitrary probability space $(\Omega, \mathcal{E}, \mu)$ to the extended real line $\bar{\mathbb{R}}$, $E^*(M)$ denotes the outer integral of M with respect to μ. (The outer integral is defined by $E^*(M) = \inf\{\int Y\, d\mu\colon Y \text{ is } \mathcal{E}\text{-measurable}, Y \geq M\}$.) Suppose $Z_1, Z_2, \dots$ and Z are maps into $l^\infty(\mathcal{V})$, and that Z is measurable. We say that Z_n converges weakly to Z, and we write $Z_n \xrightarrow{d} Z$, if $E^*(\phi(Z_n)) \to E(\phi(Z))$ for every bounded, continuous, real function ϕ on $l^\infty(\mathcal{V})$.

We now return to the empirical process $X_n = n^{1/2}(P_n - P)$. A class $\mathcal{V}$ is called a Donsker class if $X_n \xrightarrow{d} X$ in $l^\infty(\mathcal{V})$, where the limit X is a mean 0 Gaussian process with covariance function

$$\operatorname{Cov}\bigl(X(V_1), X(V_2)\bigr) = P(V_1V_2) - P(V_1)P(V_2), \qquad V_1, V_2 \in \mathcal{V},$$

and has paths which are uniformly continuous with respect to the semi-metric ρ_P on $l^\infty(\mathcal{V})$ defined by $\rho_P^2(f_1, f_2) = \operatorname{Var}_P\bigl(f_1(\theta) - f_2(\theta)\bigr)$. Sometimes we will say $\mathcal{V}$ is P-Donsker, to emphasize the dependence on P.

We say that a class $\mathcal{V}$ of measurable functions $V\colon \Theta \to \mathbb{R}$ is P-measurable if for every n and every vector $(e_1, \dots, e_n) \in \mathbb{R}^n$, the function

$$(\theta_1, \dots, \theta_n) \mapsto \sup_{V \in \mathcal{V}} \left| \sum_{i=1}^n e_i V(\theta_i) \right|$$

is measurable on the completion of $(\Theta^n, \mathcal{B}^n, P^n)$.

Because $\mathcal{F}$ and $\mathcal{G}$ are simply parametric families indexed by $h \in \mathcal{H}$, we will slightly abuse terminology and take the two convergence statements in (2.18) to mean $X_n \xrightarrow{d} X$ in $l^\infty(\mathcal{F})$ and $X_n \xrightarrow{d} X$ in $l^\infty(\mathcal{G})$, respectively. The limit $\mathbb{F}$ is a mean 0 Gaussian process indexed by $h \in \mathcal{H}$ and covariance function

$$\mathrm{Cov}\big(\mathbb{F}(h'), \mathbb{F}(h'')\big) = P(f_{h'} f_{h''}) - P(f_{h'})P(f_{h''}) \qquad \text{for any } h', h'' \in \mathcal{H}.$$

Similarly, $\mathbb{G}$ is a mean 0 Gaussian process indexed by $h \in \mathcal{H}$ and covariance function

$$\mathrm{Cov}\big(\mathbb{G}(h'), \mathbb{G}(h'')\big) = P(g^2 f_{h'} f_{h''}) - P(g f_{h'})P(g f_{h''}) \qquad \text{for any } h', h'' \in \mathcal{H},$$

and we will discuss the covariance function of the limit $\mathbb{I}_g$ in (2.19) later. For $\delta > 0$, let $\mathcal{F}_\delta = \{\phi - \psi\colon \phi, \psi \in \mathcal{F},\ \|\phi - \psi\|_{P,2} < \delta\}$ and let $\mathcal{F}_\infty^2 = \{\xi^2 : \xi \in \mathcal{F}_\infty\}$.

Before we state the next theorem, we need to lay down preparations for its fourth part, which regards functional weak convergence of the process $n^{1/2}\big(\hat{I}_g(\cdot) - I_g(\cdot)\big)$. Let $C(\mathcal{H})$ be the space of all continuous functions $x\colon \mathcal{H} \to \mathbb{R}$, with the topology induced by the sup norm metric ρ: for $x, y \in C(\mathcal{H})$, $\rho(x, y) = \|x - y\|_\infty = \sup_h |x(h) - y(h)|$. Clearly, functional weak convergence of $n^{1/2}\big(\hat{I}_g(\cdot) - I_g(\cdot)\big)$ cannot take place in a space of the type $l^\infty(\mathcal{V})$ for some set of functions $\mathcal{V}$, and in fact, as we will see, the weak convergence will take place in the space $C(\mathcal{H})$. (As usual, if μ_n, $n = 1, 2, \dots$ and μ are probability measures on $C(\mathcal{H})$, we say that $\mu_n \xrightarrow{d} \mu$ if $\int \Phi\, d\mu_n \to \int \Phi\, d\mu$ for all functions $\Phi\colon C(\mathcal{H}) \to \mathbb{R}$ which are bounded and continuous.)

We now define the expression for the covariance function and give motivation for its form. For any $h', h'' \in \mathcal{H}$, the multivariate CLT states that

$$\begin{pmatrix} U_1 \\ U_2 \\ U_3 \\ U_4 \end{pmatrix} := n^{1/2} \begin{pmatrix} P_n(g f_{h'}) - P(g f_{h'}) \\ P_n(f_{h'}) - P(f_{h'}) \\ P_n(g f_{h''}) - P(g f_{h''}) \\ P_n(f_{h''}) - P(f_{h''}) \end{pmatrix} \xrightarrow{d} \mathcal{N}\big(0, \Sigma(h', h'')\big), \tag{2.20}$$

where $\Sigma(h', h'')$ is the 4×4 matrix given by $\Sigma(h', h'')_{ij} = \mathrm{Cov}(U_i, U_j)$, $i, j = 1, 2, 3, 4$. Consider the function $\phi\colon \mathbb{R}^4 \to \mathbb{R}^2$ defined by $\phi(u_1, u_2, u_3, u_4) =$

$(u_1/u_2, u_3/u_4)$. Then, if we apply the delta method to (2.20) using ϕ, we get

$$n^{1/2}\begin{pmatrix}\hat{I}_g(h') - I_g(h') \\ \hat{I}_g(h'') - I_g(h'')\end{pmatrix} \xrightarrow{d} \mathcal{N}(0, M(h', h'')), \tag{2.21}$$

where $M(h', h'') = (\nabla\phi)^\top \Sigma(h', h'')\nabla\phi$, and $\nabla\phi$ (viewed as a 4×2 matrix) is evaluated at the vector of means $\big(P(gf_{h'}), P(f_{h'}), P(gf_{h''}), P(f_{h''})\big)$. The matrix $M(h', h'')$ describes the covariance structure for the process $\mathbb{I}_g(\cdot)$. (Expressions for $\nabla\phi$ and $M(h', h'')$ are given in Park (2015).)

THEOREM 5 *Assume that $\theta_1, \dots, \theta_n$ are independent and identically distributed according to P.*

1 (a) *Suppose that $f_\cdot(\cdot)\colon \mathcal{H} \times \Theta \to \mathbb{R}$ is continuous in h for P-almost all θ. If $\sup_{h\in\mathcal{H}} f_h$ is measurable and $\int \sup_{h\in\mathcal{H}} f_h \, dP < \infty$, then $\mathcal{F}$ is P-Glivenko-Cantelli.*

  (b) *Suppose that $(gf_\cdot)(\cdot)\colon \mathcal{H} \times \Theta \to \mathbb{R}$ is continuous in h for P-almost all θ. If $\sup_{h\in\mathcal{H}} |gf_h|$ is measurable and $\int \sup_{h\in\mathcal{H}} |gf_h| \, dP < \infty$, then $\mathcal{G}$ is P-Glivenko-Cantelli.*

2 *Assume the conditions of Part 1 of the theorem, and also that for every $\theta \in \Theta$, $\nabla_h f_h$ exists and is continuous on $\mathcal{H}$. Then*

$$\sup_{h\in\mathcal{H}} |\hat{I}_g(h) - I_g(h)| \xrightarrow{a.s.} 0. \tag{2.22}$$

3 (a) *Suppose that the classes $\mathcal{F}$, $\mathcal{F}_\delta$, $\delta > 0$, and $\mathcal{F}^2_\infty$ are all P-measurable. Assume also that for P-almost all $\theta \in \Theta$, $\nabla_h f_h$ exists and is continuous on $\mathcal{H}$. If (1) $\sup_{h\in\mathcal{H}} \|\nabla_h f_h\|$ is measurable and (2) the functions f_h, $h \in \mathcal{H}$ and $\sup_{h\in\mathcal{H}} \|\nabla_h f_h\|$ are all square integrable with respect to P, then the class $\mathcal{F}$ is P-Donsker.*

  (b) *Suppose that the classes $\mathcal{G}$, $\mathcal{G}_\delta$, $\delta > 0$, and $\mathcal{G}^2_\infty$ are all P-measurable. Assume also that for P-almost all $\theta \in \Theta$, $\nabla_h(gf_h)$ exists and is continuous on $\mathcal{H}$. If (1) $\sup_{h\in\mathcal{H}} \|\nabla_h(gf_h)\|$ is measurable and (2) the functions gf_h, $h \in \mathcal{H}$ and $\sup_{h\in\mathcal{H}} \|\nabla_h(gf_h)\|$ are all square integrable with respect to P, then the class $\mathcal{G}$ is P-Donsker.*

4 *Under the conditions of Part 3 of the theorem, we have*

$$n^{1/2}\big(\hat{I}_g(\cdot) - I_g(\cdot)\big) \xrightarrow{d} \mathbb{I}_g(\cdot) \qquad \text{in } C(\mathcal{H}),$$

where $\mathbb{I}_g$ is a Gaussian process indexed by $\mathcal{H}$ with mean 0 and covariance function

$$\begin{aligned}&\operatorname{Cov}(\mathbb{I}_g(h'), \mathbb{I}_g(h'')) \\ &= \frac{P(g^2 f_{h'} f_{h''}) - P(gf_{h'}f_{h''})\Big(\frac{P(gf_{h''})}{P(f_{h''})} + \frac{P(gf_{h'})}{P(f_{h'})}\Big) + \frac{P(gf_{h'})P(gf_{h''})}{P(f_{h'})P(f_{h''})}P(f_{h'}f_{h''})}{P(f_{h'})P(f_{h''})}.\end{aligned}$$

Part 1(a) is, of course, simply a restatement of Theorem 1; we have repeated it here only to clarify the structure of our results. The P-measurability conditions cannot be omitted. However, in all the problems we have encountered, the relevant functions are not only measurable, but are actually continuous.

In Remark 8, which follows the statement of Theorem 6, we develop a construction of confidence bands for $I_g(\cdot)$ and we explain why Theorem 6 shows that these bands are valid globally. Theorem 6 pertains to Markov chains, but the same construction and arguments can be applied to the iid case—we use Theorem 5 instead of Theorem 6.

The next result is a version of Theorem 5 that applies to Markov chains. Recall that $N_r = \tau_r - \tau_{r-1}$ is the length of the r^{th} tour and that $S_r^{(h)}$ is defined by (2.3). Similarly, define $T_r^{(h)} = \sum_{i=\tau_{r-1}}^{\tau_r - 1} g(\theta_i) f_h(\theta_i)$, $r = 1, 2, \ldots$. Let $\mathscr{F} = \{S_1^{(h)},\, h \in \mathcal{H}\}$ and $\mathscr{G} = \{T_1^{(h)},\, h \in \mathcal{H}\}$. Part 3 of Theorem 6 asserts that under certain conditions the classes $\mathscr{F}$ and $\mathscr{G}$ are Donsker, and before stating the theorem, it is necessary to be very clear regarding what these classes are, and what "Donsker" means. Let P be the distribution of the Markov chain $\theta_1, \theta_2, \ldots$. For any $h \in \mathcal{H}$, $S_1^{(h)}$ is a function mapping the measure space $(\Theta^\infty, \mathcal{B}^\infty, \mathsf{P})$ into $\mathbb{R}_+$. To see this it may be helpful to imagine that we are dealing with the very simple case of a regenerative chain which has an "proper atom" at a singleton. That is, there exists a point $\alpha \in \Theta$ which has positive probability under the invariant measure. Thus, with probability one the chain returns to α infinitely often, and the times of return to α are regeneration times $\tau_0, \tau_1, \tau_2, \ldots$. In this case (with probability one) the sequence $\theta_1, \theta_2, \ldots$ itself determines τ_0 and τ_1. Then, $S_1^{(h)} \colon \Theta^\infty \to \mathbb{R}_+$ is defined by $S_1^{(h)}(\theta_1, \theta_2, \ldots) = \sum_{i=\tau_0}^{\tau_1 - 1} f_h(\theta_i)$, and we have a similar definition for $T_1^{(h)}$. Chains which have a proper atom at a singleton are quite rare, and we consider them only for exposition. We remark on the case of a general regenerative Markov chain at the end of the proof of Theorem 6. To clarify, $\mathscr{F}$ and $\mathscr{G}$ are classes of functions on Θ^∞, in contrast to $\mathcal{F}$ and $\mathcal{G}$, which are classes of functions on Θ. These classes will be P-Donsker, and we note that P is a distribution on the infinite product space Θ^∞, to be distinguished from P, which is a distribution on Θ.

As we will see, Parts 3 and 4 of Theorem 6 are functional CLT's that concerns certain stochastic processes indexed by $h \in \mathcal{H}$. In order to motivate them, we need to first understand the version of these parts of the theorem that pertains to the very simple situation in which we are considering a single value of h. Thus, let $h \in \mathcal{H}$ be fixed. We now consider CLT's for averages formed from the sequences $S_1^{(h)}, S_2^{(h)}, \ldots$ and $T_1^{(h)}, T_2^{(h)}, \ldots$. We have $E(S_1^{(h)}) = E_P(f_h(\theta))E(N_1)$ and $E(T_1^{(h)}) = E_P(g(\theta) f_h(\theta))E(N_1)$ (see (2.4)). Under A1 and the conditions $E_P(f_h^{2+\epsilon}(\theta)) < \infty$ and $E_P[(g f_h)^{2+\epsilon}(\theta)] < \infty$, the expectations $E[(S_1^{(h)})^2]$,

$E\big[(T_1^{(h)})^2\big]$, and $E(N_1^2)$ are all finite (Theorem 2 of Hobert et al. 2002). Therefore, the simple multivariate CLT gives

$$R^{1/2}\begin{pmatrix} (\sum_{r=1}^R T_r^{(h)})/R - E_P(g(\theta)f_h(\theta))E(N_1) \\ (\sum_{r=1}^R S_r^{(h)})/R - E_P(f_h(\theta))E(N_1) \\ (\sum_{r=1}^R N_r)/R - E(N_1) \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, V_h), \tag{2.23}$$

where $V_h = \operatorname{Cov}\big((T_1^{(h)}, S_1^{(h)}, N_1)^\top\big)$. We apply the delta method to (2.23) three times, using the functions $q_1(u,v,w) = v/w$, $q_2(u,v,w) = u/w$, and $q_3(u,v,w) = u/v$ to obtain three CLT's:

$$\begin{aligned} R^{1/2}\bigg(\frac{\sum_{r=1}^R S_r^{(h)}}{\sum_{r=1}^R N_r} - E_P(f_h(\theta))\bigg) &\xrightarrow{d} \mathcal{N}\big(0, (\nabla q_1)^\top V_h \nabla q_1\big), \\ R^{1/2}\bigg(\frac{\sum_{r=1}^R T_r^{(h)}}{\sum_{r=1}^R N_r} - E_P(g(\theta)f_h(\theta))\bigg) &\xrightarrow{d} \mathcal{N}\big(0, (\nabla q_2)^\top V_h \nabla q_2\big), \\ R^{1/2}\bigg(\frac{\sum_{r=1}^R T_r^{(h)}}{\sum_{r=1}^R S_r^{(h)}} - I_g(h)\bigg) &\xrightarrow{d} \mathcal{N}\big(0, (\nabla q_3)^\top V_h \nabla q_3\big). \end{aligned} \tag{2.24}$$

With the relationships $n = \sum_{r=1}^R N_r$, $\sum_{r=1}^R S_r^{(h)} = \sum_{i=1}^n f_h(\theta_i)$, $\sum_{r=1}^R T_r^{(h)} = \sum_{i=1}^n g(\theta_i)f_h(\theta_i)$, and the fact that $n/R \xrightarrow{\text{a.s.}} E(N_1)$, (2.24) may be restated as

$$\begin{aligned} n^{1/2}\bigg(\frac{\sum_{i=1}^n f_h(\theta_i)}{n} - E_P(f_h(\theta))\bigg) &\xrightarrow{d} \mathcal{N}\big(0, E(N_1)(\nabla q_1)^\top V_h \nabla q_1\big), \\ n^{1/2}\bigg(\frac{\sum_{i=1}^n g(\theta_i)f_h(\theta_i)}{n} - E_P(g(\theta)f_h(\theta))\bigg) &\xrightarrow{d} \mathcal{N}\big(0, E(N_1)(\nabla q_2)^\top V_h \nabla q_2\big), \\ n^{1/2}\bigg(\frac{\sum_{i=1}^n g(\theta_i)f_h(\theta_i)}{\sum_{i=1}^n f_h(\theta_i)} - I_g(h)\bigg) &\xrightarrow{d} \mathcal{N}\big(0, E(N_1)(\nabla q_3)^\top V_h \nabla q_3\big) \end{aligned} \tag{2.25}$$

(with the understanding that here, n is random). Of course, under geometric ergodicity and the moment conditions $E_P(f_h^{2+\epsilon}(\theta)) < \infty$ and $E_P[(gf_h)^{2+\epsilon}(\theta)] < \infty$, asymptotic normality of the three quantities on the left side of (2.25) is already known (corollary to Theorem 18.5.3 of Ibragimov and Linnik 1971). The point of obtaining (2.25) as we did above is that the method enables us to get functional versions of the three statements in (2.25) (i.e. weak convergence of the three quantities on the left side of (2.25) as processes in h) if we can show that the classes $\mathscr{F}$ and $\mathscr{G}$ are Donsker. This is precisely what Part 3 of Theorem 6 asserts. The theorem will refer to the following conditions.

B1 For every $h \in \mathcal{H}$, there exists $\epsilon > 0$ such that $E_P(f_h^{2+\epsilon}(\theta)) < \infty$.
B2 For every $h \in \mathcal{H}$, there exists $\epsilon > 0$ such that $E_P[(gf_h)^{2+\epsilon}(\theta)] < \infty$.

THEOREM 6 *Assume that $\theta_1, \theta_2, \ldots$ is a Harris ergodic Markov chain with invariant distribution P for which there exists a regeneration sequence $1 = \tau_0 < \tau_1 < \tau_2 < \cdots$ satisfying $E(\tau_1 - \tau_0) < \infty$.*

1 (a) Suppose that $f.(\cdot)\colon \mathcal{H} \times \Theta \to \mathbb{R}$ is continuous in h for P-almost all θ. Suppose also that $\sup_h S_1^{(h)}$ is measurable and integrable. Then (2.5) holds.

(b) Suppose that $(gf.)(\cdot)\colon \mathcal{H} \times \Theta \to \mathbb{R}$ is continuous in h for P-almost all θ. Suppose also that $\sup_h |T_1^{(h)}|$ is measurable and integrable. Then in analogy with (2.5), we have

$$\sup_h \left| \frac{1}{n} \sum_{i=1}^n g(\theta_i) f_h(\theta_i) - E_P(g(\theta) f_h(\theta)) \right| \xrightarrow{a.s.} 0.$$

2 Assume the conditions of Part 1 of the theorem, and also that for every $\theta \in \Theta$, $\nabla_h f_h$ exists and is continuous on $\mathcal{H}$. Then

$$\sup_{h \in \mathcal{H}} |\hat{I}_g(h) - I_g(h)| \xrightarrow{a.s.} 0. \tag{2.26}$$

3 (a) Suppose that the classes $\mathscr{F}$, $\mathscr{F}_\delta$, $\delta > 0$, and $\mathscr{F}_\infty^2$ are all P-measurable. Suppose also that for almost all $\theta \in \Theta$, $\nabla_h f_h$ exists and is continuous on $\mathcal{H}$. Under A1, B1, and the condition that $\sup_{h\in\mathcal{H}} \|\nabla_h S_1^{(h)}\|$ is measurable and square integrable with respect to P, the class $\mathscr{F}$ is P-Donsker.

(b) Suppose that the classes $\mathscr{G}$, $\mathscr{G}_\delta$, $\delta > 0$, and $\mathscr{G}_\infty^2$ are all P-measurable. Suppose also that for almost all $\theta \in \Theta$, $\nabla_h (gf_h)$ exists and is continuous on $\mathcal{H}$. Under A1, B2, and the condition that $\sup_{h\in\mathcal{H}} \|\nabla_h T_1^{(h)}\|$ is measurable and square integrable with respect to P, the class $\mathscr{G}$ is P-Donsker.

4 Under the conditions of Part 3 of the theorem, we have

$$R^{1/2}(\hat{I}_g(\cdot) - I_g(\cdot)) \xrightarrow{d} \mathbb{I}_g^*(\cdot) \qquad \text{in } C(\mathcal{H}), \tag{2.27}$$

where $\mathbb{I}_g^*$ is a Gaussian process indexed by $\mathcal{H}$ with mean 0 and covariance function

$$\begin{aligned}\operatorname{Cov}(\mathbb{I}_g^*(h'), \mathbb{I}_g^*(h'')) = \Big[\mathsf{P}(S_1^{(h')})\mathsf{P}(S_1^{(h'')})\Big]^{-1} \Bigg[& \mathsf{P}(T_1^{(h')} T_1^{(h'')}) \\ & - \mathsf{P}(S_1^{(h')} T_1^{(h'')}) \left(\frac{\mathsf{P}(T_1^{(h'')})}{\mathsf{P}(S_1^{(h'')})} + \frac{\mathsf{P}(T_1^{(h')})}{\mathsf{P}(S_1^{(h')})} \right) \\ & + \frac{\mathsf{P}(T_1^{(h')})\mathsf{P}(T_1^{(h'')})}{\mathsf{P}(S_1^{(h')})\mathsf{P}(S_1^{(h'')})} \mathsf{P}(S_1^{(h')} S_1^{(h'')}) \Bigg].\end{aligned}$$

Consequently,

$$n^{1/2}\big(\hat{I}_g(\cdot) - I_g(\cdot)\big) \xrightarrow{d} \tilde{\mathbb{I}}_g(\cdot) \qquad \text{in } C(\mathcal{H}), \tag{2.28}$$

where $\tilde{\mathbb{I}}_g$ *is a Gaussian process indexed by* $\mathcal{H}$ *with mean* 0 *and covariance function*

$$\operatorname{Cov}\big(\tilde{\mathbb{I}}_g(h'), \tilde{\mathbb{I}}_g(h'')\big) = E(N_1)\operatorname{Cov}\big(\mathbb{I}_g^*(h'), \mathbb{I}_g^*(h'')\big).$$

In (2.27) $\hat{I}_g(h)$ *is interpreted as* $\hat{I}_g(h) = \big(\sum_{r=1}^R T_r^{(h)}\big)/\sum_{r=1}^R S_r^{(h)}$, *and the limit is as* $R \to \infty$, *whereas in* (2.28) $\hat{I}_g(h)$ *and the limit are interpreted differently:* $\hat{I}_g(h) = \big(\sum_{i=1}^n g(\theta_i) f_h(\theta_i)\big)/\sum_{i=1}^n f_h(\theta_i)$, *and* $n = \sum_{r=1}^R N_r$ *is random.*

REMARK 8 Here we discuss how to form globally valid confidence bands for $I(\cdot)$ (we drop the subscript "g" to lighten the notation). We would like to proceed as follows. Having established that $n^{1/2}\big(\hat{I}(\cdot) - I(\cdot)\big) \xrightarrow{d} \tilde{\mathbb{I}}(\cdot)$, we find the distribution of $\sup_h |\tilde{\mathbb{I}}(h)|$. If s_α is the $(1-\alpha)$-quantile of this distribution, then the band $\hat{I}(h) \pm n^{-1/2} s_\alpha$ has asymptotic coverage probability equal to $1-\alpha$. Unfortunately, except for very unusual cases, the distribution of $\sup_h |\tilde{\mathbb{I}}(h)|$ cannot be obtained analytically. Spectral methods can be used for the problem of forming confidence intervals for $I(h)$ for a single value of h, but not for the problem of forming confidence bands. We know of no way to use regenerative simulation to construct confidence bands. However, the method of batching works, as follows.

For a positive integer M, the sequence $\theta_1, \dots, \theta_n$ is broken up into M consecutive pieces, each of length n/M (we are ignoring divisibility issues). For $m = 1, \dots, M$, let $\hat{I}^{(m)}(h)$ be the estimate of $I(h)$ based on batch m, and let

$$\mathcal{I}_m = \sup_h \Big(\frac{n}{M}\Big)^{1/2} |\hat{I}^{(m)}(h) - \hat{I}(h)|, \qquad \bar{\mathcal{I}}_m = \sup_h \Big(\frac{n}{M}\Big)^{1/2} |\hat{I}^{(m)}(h) - I(h)|.$$

(The difference between $\mathcal{I}_m$ and $\bar{\mathcal{I}}_m$ is that the latter is not computable, because it involves the unknown function $I(\cdot)$.) Let $\bar{\mathcal{I}}_{[1]} \le \bar{\mathcal{I}}_{[2]} \le \cdots \le \bar{\mathcal{I}}_{[M]}$ be the order statistics of the sequence $\bar{\mathcal{I}}_1, \dots, \bar{\mathcal{I}}_M$ and, similarly, let $\mathcal{I}_{[1]} \le \mathcal{I}_{[2]} \le \cdots \le \mathcal{I}_{[M]}$ be the order statistics of the sequence $\mathcal{I}_1, \dots, \mathcal{I}_M$. Now suppose that $M \to \infty$ in such a way that $n/M \to \infty$. Below is the outline of an argument which shows that the band $\hat{I}(h) \pm n^{-1/2}\mathcal{I}_{[(1-\alpha)M]}$ has coverage probability that is asymptotically equal to $1-\alpha$.

1. For every m, we have $\bar{\mathcal{I}}_m \xrightarrow{d} \sup_h |\tilde{\mathbb{I}}(h)|$ by Theorem 6, and if the distribution of $\sup_h |\tilde{\mathbb{I}}(h)|$ is continuous, then $\bar{\mathcal{I}}_{[(1-\alpha)M]}$ converges in distribution to δ_{s_α}, the point mass at s_α.
2. Therefore the (uncomputable) band $\hat{I}(h) \pm n^{-1/2}\bar{\mathcal{I}}_{[(1-\alpha)M]}$ has coverage probability that converges to $1-\alpha$.

3. The difference between $\mathcal{I}_m$ and $\bar{\mathcal{I}}_m$ is small uniformly in m; more precisely, we have $\max_{1\le m\le M} |\mathcal{I}_m - \bar{\mathcal{I}}_m| \xrightarrow{P} 0$. Therefore the band $\hat{I}(h) \pm n^{-1/2}\mathcal{I}_{[(1-\alpha)M]}$ also has coverage probability that converges to $1-\alpha$.

Details are given in Park (2015).

REMARK 9 We have seen that for any $h_1 \in \mathcal{H}$, if $\theta_1, \theta_2, \dots$ is a Markov chain with invariant distribution $\nu_{h_1,y}$ then, under certain regularity conditions, the estimates $B_n(h)$ and $\hat{I}_g(h)$ are consistent and asymptotically normal. These estimates can be unstable, however, if h is far from h_1, and there may not exist a single value of h_1 that gives rise to estimates that are stable for all $h \in \mathcal{H}$. Serial tempering (Marinari and Parisi (1992); Geyer and Thompson (1995); see also Geyer (2011) for a review, and Tan (2014) for recent developments) can be very effective in handling this problem. A very brief description of the method in the present context is as follows. We select m points $h_1, \dots, h_m \in \mathcal{H}$; these should be taken to "cover" $\mathcal{H}$ in the sense that every h in $\mathcal{H}$ is "close" to at least one of the h_j's. Let $\mathcal{L} = \{1, \dots, m\}$; the elements of $\mathcal{L}$ are called "labels." For each $j \in \mathcal{L}$, let Φ_j be a Markov transition function with invariant distribution $\nu_{h_j,y}$. A Markov chain running on the state space $\mathcal{L} \times \Theta$ is generated as follows. If the current state of the chain is (j, θ), a new label j' is generated, and θ' is generated from the distribution $\Phi_{j'}(\theta, \cdot)$. The mechanism for generating the labels is set up in such a way that the θ-sequence has invariant distribution $\sum_{j=1}^m \alpha_j \nu_{h_j,y}$, where the α_j's are all nearly equal to $1/m$. From the θ-sequence, the quantities $B(h)$ and $I_g(h)$ can be estimated in a stable manner for any h which is "close" to at least one of the h_j's, or more precisely, for any h such that ν_h is "close" to at least one of $\nu_{h_1}, \dots, \nu_{h_m}$. *The results of this paper do not require that the sequence $\theta_1, \theta_2, \dots$ have invariant distribution equal to $\nu_{h_1,y}$ for some $h_1 \in \mathcal{H}$, and in fact the invariant distribution can be a mixture $\sum_{j=1}^m \alpha_j \nu_{h_j,y}$, for judiciously chosen $h_1, \dots, h_m$, as described above, for example.*

3. Illustrations. Here we present two illustrations. The first deals with the so-called Latent Dirichlet Allocation model, which is used for organizing and searching electronic documents. The version of the model we discuss is indexed by a two-dimensional hyperparameter. Our focus will be on obtaining globally-valid confidence sets for a certain posterior expectation of interest. For the data set we study, the amount of time it takes to run the Markov chain is a significant issue because each cycle has length 7788. We will use the results of Section 2.3 to determine the minimal Markov chain length that is needed to obtain acceptably narrow confidence regions. The second illustration deals with a model for Bayesian variable selection in linear regression. For this situation our interest will be on hyperparameter selection, and we will use the results of Section 2.2. We will see that for the data set we use, a very modest Markov chain length is all that is needed to produce narrow confidence sets for the empirical Bayes choice of the hyperparameters.

3.1. *Sensitivity Analysis in the Latent Dirichlet Allocation Model.* Probabilistic topic modelling is an area of machine learning that deals with methods for understanding, summarizing, and searching large electronic archives. Traditional keyword-based searches are very fast, but have important deficiencies. Suppose we are interested in searching for all statistical papers that deal with censored data. A search using the keywords "censored data" will not return papers that use the expression "incomplete data". In topic-based searches, we do a search based on a concept or topic. A topic is not an expression; it is, by definition, a distribution over a set of expressions. Thus the topic mentioned above gives a lot of mass to expressions like "Kaplan-Meier", "censored data", and "incomplete data", and little mass to expressions like "spectral decomposition".

Latent Dirichlet Allocation (LDA, Blei et al. 2003) is by far the most used topic model. We will consider the version of the model that deals only with individual words, as opposed to expressions consisting of several words. Suppose we have a corpus of documents, for example a set of articles from *The New York Times*, and these span several different topics, such as sports, medicine, politics, etc. The words in the documents come from a vocabulary $\mathcal{V}$, which is a set consisting of V words $u_1, \dots, u_V$. For each document, the data we have for that document is a sequence of length V consisting of the number of times that word u_v occurs, for $v = 1, \dots, V$. In LDA, we imagine that for each word in each document, there is a latent (i.e. unobserved) variable indicating a topic from which that word is drawn. LDA enables us to make inference on these latent variables, and therefore, on the topics that are covered by each document as a whole. Therefore, LDA enables us to cluster together documents which are similar, i.e. documents which share common topics. By its very nature, LDA is completely automatic in how it defines the topics: these are distributions over the vocabulary, and are themselves latent variables. To be more precise, in LDA there is no such thing as a topic called "sports". Instead, there is a distribution on $\mathcal{V}$ which gives most of its mass to words like "homerun", "marathon", and "NBA". A human is then free to call this distribution "sports" if he/she wishes.

We now give more detail. The vocabulary $\mathcal{V}$ is taken to be the union of all the words in all the documents of the corpus, after removing uninformative words (like "the" and "of"). There are D documents in the corpus, and for $d = 1, \dots, D$, document d has n_d words, $w_{d1}, \dots, w_{dn_d}$. The order of the words is viewed as uninformative, so is neglected. Each word is represented as an index $1 \times V$ vector with a 1 at the s^{th} element, where s denotes the term selected from the vocabulary. Thus, document d is represented by the vector $\boldsymbol{w}_d = (w_{d1}, \dots, w_{dn_d})$ and the corpus is represented by the vector $\boldsymbol{w} = (\boldsymbol{w}_1, \dots, \boldsymbol{w}_D)$. The number of topics, K, is finite and known. By definition, a topic is a point in $\mathbb{S}_V$, the $(V-1)$-dimensional simplex. For $d = 1, \dots, D$, for each word w_{di}, z_{di} is an index $1 \times K$ vector which represents the latent variable that denotes the topic from which w_{di} is drawn. The distribution of $z_{d1}, \dots, z_{dn_d}$ will depend on a document-specific variable θ_d which indicates a distribution on the topics for document d. We will use $\mathrm{Dir}_L(a_1, \dots, a_L)$ to denote the finite-dimensional Dirichlet distribution on the L-dimensional simplex. Also, we will use $\mathrm{Mult}_L(b_1, \dots, b_L)$ to denote the multinomial distribution with number of trials equal to 1 and probability vector $(b_1, \dots, b_L)$. We will form a $K \times V$ matrix $\boldsymbol{\beta}$, whose t^{th} row is the t^{th} topic (how $\boldsymbol{\beta}$ is formed will be described shortly). Thus, $\boldsymbol{\beta}$ will consist of vectors $\beta_1, \dots, \beta_K$, all lying in $\mathbb{S}_V$. Formally, LDA is described by the following hierarchical model, in which $\eta, \alpha \in (0, \infty)$ are hyperparameters:

1. $\beta_t \stackrel{\text{iid}}{\sim} \mathrm{Dir}_V(\eta, \dots, \eta),\ t = 1, \dots, K.$
2. $\theta_d \stackrel{\text{iid}}{\sim} \mathrm{Dir}_K(\alpha, \dots, \alpha),\ d = 1, \dots, D$, and the θ_d's are independent of the β_t's.
3. Given $\theta_1, \dots, \theta_D$, $z_{di} \stackrel{\text{iid}}{\sim} \mathrm{Mult}_K(\theta_d),\ i = 1, \dots, n_d,\ d = 1, \dots, D$, and the D vectors $(z_{11}, \dots, z_{1n_1}), \dots, (z_{D1}, \dots, z_{Dn_D})$ are independent.
4. Given $\boldsymbol{\beta}$ and the z_{di}'s, w_{di} are independently drawn from the row of $\boldsymbol{\beta}$ indicated by $z_{di},\ i = 1, \dots, n_d,\ d = 1, \dots, D.$

From the model statement, we see that there is a latent topic variable for every word that appears in the corpus. Thus it is possible that a document spans several topics. However, because there is a single θ_d for document d, the model encourages different words in the same document to have the same topic. Also note that the hierarchical nature of LDA encourages different documents to share the same topics. This is because $\boldsymbol{\beta}$ is chosen once, at the top of the hierarchy, and is shared among the D documents. Let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_D)$, $\boldsymbol{z}_d = (z_{d1}, \dots, z_{dn_d})$ for $d = 1, \dots, D$, $\boldsymbol{z} = (\boldsymbol{z}_1, \dots, \boldsymbol{z}_D)$, and let $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z})$. The model is indexed by the hyperparameter vector $h = (\eta, \alpha)$. For any given h, lines 1–3 induce a prior distribution on $\boldsymbol{\psi}$, which we denote by ν_h. Line 4 gives the likelihood. The words $\boldsymbol{w}$ are observed, and we are interested in $\nu_{h,\boldsymbol{w}}$, the posterior distribution of $\boldsymbol{\psi}$ given $\boldsymbol{w}$ corresponding to ν_h.

The hyperparameter h has a strong effect on the distribution of the parameters of the model. For example, when η is large, the topics tend to be probability vectors which spread their mass evenly among many words in the vocabulary, whereas when η is small, the topics tend to put most of their mass on only a few words. Also, when α is large, each document tends to involve many different topics; on the other hand, in the limiting case where $\alpha \to 0$, each document involves a single topic, and this topic is randomly chosen from the set of all topics.

In the literature, the following choices for $h = (\eta, \alpha)$ have been presented: $h_{\text{GS}} = (0.1, 50/K)$, used in Griffiths and Steyvers (2004); $h_{\text{A}} = (0.1, 0.1)$, used in Asuncion et al. (2009); and $h_{\text{RS}} = (1/K, 1/K)$, used in the `Gensim` topic modelling package (Řehůřek and Sojka, 2010), a well-known package used in the topic modelling community. These choices are ad-hoc, and not based on any principle; nevertheless, they do get used. Blei et al. (2003) propose $h_0 = \arg\max_h m_{\boldsymbol{w}}(h)$, as we do, but their approach for estimating h_0 is quite a bit different from ours, and involves a combination of the EM algorithm and "variational inference." Very briefly, $\boldsymbol{w}$ is viewed as "observed data," and $\boldsymbol{\psi}$ is viewed as "missing data." Because the "complete data likelihood" $p_h(\boldsymbol{\psi}, \boldsymbol{w})$ is available, the EM algorithm is a natural candidate for estimating $\arg\max_h m_{\boldsymbol{w}}(h)$, since $m_{\boldsymbol{w}}(h)$ is the "incomplete data likelihood." But the E-step in the algorithm is infeasible because it requires calculating an expectation with respect to the intractable distribution $\nu_{h,\boldsymbol{w}}$. Blei et al. (2003) substitute an approximation to this expectation. Unfortunately, because there are no useful bounds on the approximation, and because the approximation is used at every iteration of the algorithm, there are no results regarding the theoretical properties of this method. Determination of the hyperparameter is currently an open problem in LDA modelling (Wallach et al., 2009).

We illustrate our methodology on a corpus of documents from the English Wikipedia, originally created by George (2015). When a Wikipedia article is created, it is typically tagged to one or more categories, one of which is the "primary category." The corpus consists of 8 documents from the category *Leopardus*, 8 from the category *Lynx*, and 7 from *Prionailurus*, and we took $K = 3$, as in George (2015). There are 303 words in the vocabulary, and the total number of words in the corpus is 7788. The data set is relatively small. However, it is challenging to analyze because the topics are very close to each other, so in the posterior distribution there is a great deal of uncertainty regarding the latent topic indicator variables, and this is why we chose this data set.

A reader of a given article may wish to look at related articles, so a question of interest is whether the topics for two given documents are nearly the same. One way to word this question precisely is to ask what is the posterior probability that $\|\theta_i - \theta_j\| \le \epsilon$, where i and j are the indices of the documents in question and ϵ is some user-specified small number. Here, $\|\cdot\|$ denotes ordinary Euclidean distance. This posterior probability will of course depend on h, and we would like to view the estimates of the posterior probability as h varies, together with (simultaneous) error margins.

To this end, we used the methodology developed in Section 2.3 for simultaneous estimation of posterior expectations (here the posterior expectations of the indicator of a set). The warning given in Remark 9 regarding the high variance of the simple single-chain estimate (1.3) applies, and we use instead a serial tempering chain (cf. Remark 9), the details of which are given in the next paragraph. We consider documents 7 and 8, which are the articles "Pampas cat" and "Pantanal cat" under the Wikipedia category Leopardus, and we are interested in the posterior probability of the event $\|\theta_7 - \theta_8\| \le .05$. Our estimate of $\arg\max_h m_{\boldsymbol{w}}(h)$ is $h_n = (\eta_n, \alpha_n) = (.915, .245)$, and the estimate of the posterior probability under the empirical Bayes choice of h is $\nu_{h_n,\boldsymbol{w}}(\|\theta_7 - \theta_8\| \le .05) = .7039$. For the other choices of h we have $\nu_{h_{\text{GS}},\boldsymbol{w}}(\|\theta_7 - \theta_8\| \le .05) = .1619$, $\nu_{h_{\text{A}},\boldsymbol{w}}(\|\theta_7 - \theta_8\| \le .05) = .1498$, and $\nu_{h_{\text{RS}},\boldsymbol{w}}(\|\theta_7 - \theta_8\| \le .05) = .1298$, and we see that all three are far from the estimate based on the empirical Bayes choice of h. We also calculated the ratio of the marginal likelihood of h_n to the marginal likelihood of each of h_{GS}, h_{A}, and h_{RS} and noted that each ratio is astronomically large. Therefore, none of these values of h are deemed even remotely plausible, and as these choices of h do not have any theoretical basis, there is no credibility to posterior probability estimates based on them. Figure 1 gives a plot of the estimate of $\nu_{h,\boldsymbol{w}}(\|\theta_7 - \theta_8\| \le .05)$, together with a globally valid confidence set of level .95 over a relatively small region centered at h_n. The figure shows that the posterior probabilities vary greatly with h, ranging from .553 to .972, even over a small h-region, underscoring the fact that the choice of hyperparameter should be made carefully.

Our serial tempering chain is based on the "augmented collapsed Gibbs sampler" developed in George (2015), and which runs on the entire set of latent variables $(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z})$. A single cycle of this Markov chain runs over 7788 nodes. To form the confidence region we used the construction described in Remark 8. We took the grid size for the chain ("m" in Remark 8) to be 105, with the 105 reference values evenly spaced over the h-region. With this choice the chain gives very stable estimates. The length of the chain was 500,000, and the number of batches was 707 (roughly the square root of the chain length). With this chain length the confidence region is adequately narrow, and with a length of only 50,000 it was not.

3.2. *Hyperparameter Choice for Bayesian Variable Selection in Linear Regression.* The most commonly used setup for variable selection in Bayesian linear regression is described as follows. We have a response vector $Y = (Y_1, \dots, Y_m)^\top$ and a set of potential predictors $X_1, \dots, X_q$, each a vector of length m. Every subset of predictors is identified with a binary vector $\gamma = (\gamma_1, \dots, \gamma_q)^\top \in \{0,1\}^q$, where $\gamma_j = 1$ if X_j is included in the model and $\gamma_j = 0$ otherwise. For every γ, we have a model given by

$$Y = 1_m \beta_0 + X_\gamma \beta_\gamma + \epsilon,$$

where 1_m is the vector of m 1's, X_γ is the design matrix whose columns consist of the predictor vectors corresponding to γ, β_γ is the vector of coefficients for that subset, and $\epsilon \sim \mathcal{N}_m(0, \sigma^2 I)$. For this setup, the unknown parameter is $\theta = (\gamma, \sigma, \beta_0, \beta_\gamma)$, which includes the indicator of the subset of variables that go into the regression model. The prior on θ is a hierarchy in which we first select the variables that go into the regression model, then a "non-informative prior" is

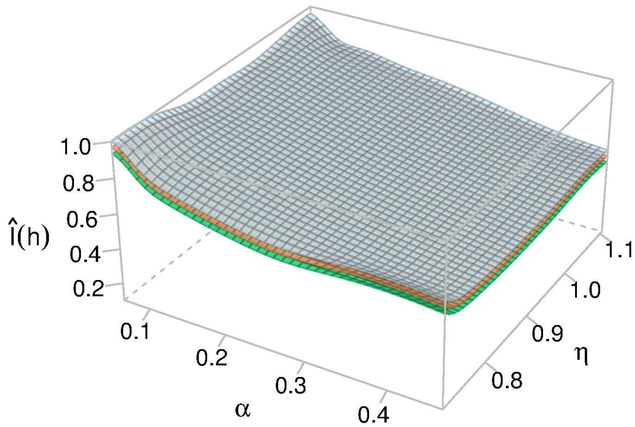

FIG 1. *Estimates with confidence region for $I(h) = \nu_{h,\boldsymbol{w}}(\|\theta_7 - \theta_8\| \leq .05)$, the posterior probability that the topics for documents 7 and 8 of the Wikipedia corpus are "very close." The plot shows that this posterior probability varies considerably with h, and suggests that care be taken in choosing the hyperparameter.*

given to (σ^2, β_0), and given γ and σ, we choose β_γ from some proper distribution. The specific instance of this model that we will consider is indexed by two hyperparameters, $w \in (0,1)$ and $g > 0$, and is given in detail as follows:

$$\text{given } \gamma, \sigma, \beta_0, \beta_\gamma, \qquad Y \sim \mathcal{N}_m(1_m\beta_0 + X_\gamma\beta_\gamma, \sigma^2 I), \tag{3.1a}$$

$$\text{given } \gamma, \sigma, \qquad \beta_\gamma \sim \mathcal{N}_{q_\gamma}(0, g\sigma^2(X_\gamma^\top X_\gamma)^{-1}), \tag{3.1b}$$

$$(\sigma^2, \beta_0) \sim p(\beta_0, \sigma^2) \propto 1/\sigma^2, \tag{3.1c}$$

$$\gamma \sim p(\gamma) = w^{q_\gamma}(1-w)^{q-q_\gamma}. \tag{3.1d}$$

The prior on γ given by (3.1d) is the so-called independence Bernoulli prior, in which every variable goes into the model with probability w, independently of all the other variables. In (3.1b), $q_\gamma = \sum_{j=1}^q \gamma_j$ is the number of predictors that go in the regression, and the prior on β_γ is Zellner's g-prior (Zellner, 1986). Because (σ^2, β_0) is given an improper prior (line (3.1c)), the prior on θ is improper; however, it turns out that the posterior distribution of θ is proper. Models of the type (3.1) were introduced by Mitchell and Beauchamp (1988) and have been studied in dozens of papers; see Liang et al. (2008) for a review.

The hyperparameter $h = (w, g)$ plays a critical role: if w is small and g is large, the prior ν_h concentrates its mass on models with few variables and large coefficients, while if w is large and g is small, ν_h concentrates its mass on models with many variables and small coefficients. (To appreciate the importance of the role played by h, note that George and Foster (2000) have shown that for the slightly different version of (3.1) in which σ^2 is assumed known, h can be chosen so that the highest posterior probability model is exactly the best model under the AIC/C_p, BIC, or RIC criteria.) Thus, h effectively determines the method that is used to carry out variable selection, so it is important to choose it properly.

Unless q is relatively small (q less than 20 or 25), the posterior distribution of $\theta = (\gamma, \sigma, \beta_0, \beta_\gamma)$ is intractable, because to compute it we need to calculate 2^q integrals (George and Foster, 2000). Smith and Kohn (1996) developed a Markov chain algorithm which runs only on γ, the other variables being integrated out. Their chain is a simple Gibbs sampler which runs on the vector $(\gamma_1, \ldots, \gamma_q)^\top$, updating one component at a time. This chain does not fit into our framework, which requires a Markov chain that runs on $\theta = (\gamma, \sigma, \beta_0, \beta_\gamma)$. Buta (2010) developed a Markov chain, based on the Smith and Kohn (1996) chain, which runs over $(\gamma, \sigma, \beta_0, \beta_\gamma)$. (She proved that for her Markov chain, the rate of convergence to the posterior distribution of θ is exactly the same as the rate of convergence to the posterior distribution of γ for the Smith and Kohn (1996) chain, where convergence is in terms of the absolute deviation norm.) We will use the chain developed by Buta (2010) for the analysis below.

To implement the methods of this paper, we need a "ratio of densities ν_{h_1}/ν_{h_2}" (cf. equation (1.3)). Note that the prior distributions are not absolutely continuous with respect to the product of counting measure on $\{0,1\}^q$ and Lebesgue measure on $(0,\infty) \times \mathbb{R}_+ \times \mathbb{R}^{q+1}$ (the dimension of β_γ is not fixed). The "ratio of densities ν_{h_1}/ν_{h_2}" then needs to be replaced by the Radon-Nikodym derivative. To be precise, let $\bar{\nu}_h$ be the distribution on θ induced by (3.1d), (3.1c), and (3.1b). Then (1.3) becomes

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{d\bar{\nu}_h}{d\bar{\nu}_{h_1}}\right](\theta_i) \xrightarrow{\text{a.s.}} \int \left[\frac{d\bar{\nu}_h}{d\bar{\nu}_{h_1}}\right](\theta)\, \bar{\nu}_{h_1,y}(d\theta) = \frac{m_y(h)}{m_y(h_1)}.$$

The Radon-Nikodym derivative was obtained in Doss (2007) and is given by

$$\left[\frac{d\bar{\nu}_{h_1}}{d\bar{\nu}_{h_2}}\right](\theta) = \left(\frac{w_1}{w_2}\right)^{q_\gamma}\left(\frac{1-w_1}{1-w_2}\right)^{q-q_\gamma} \times \frac{\phi_{q_\gamma}\big(\beta_\gamma; 0, g_1\sigma^2(X_\gamma' X_\gamma)^{-1}\big)}{\phi_{q_\gamma}\big(\beta_\gamma; 0, g_2\sigma^2(X_\gamma' X_\gamma)^{-1}\big)},$$

where $\phi_d(u; a, V)$ is the density of the d-dimensional normal distribution with mean a and covariance V, evaluated at u.

For our illustration we consider the ragweed data of Stark et al. (1997), who were interested in determining how meteorological variables can be used to forecast ragweed pollen levels. The response variable is the ragweed level (grains/m^3) for 335 days in Kalamazoo, Michigan, USA. Although the data set contains other predictors, we restrict our analysis to two: day (day number in the current ragweed pollen season) and wind (wind speed forecast in knots for following day).

Following Ruppert et al. (2003), we take the square root of the ragweed level as the response. Figure 2 gives separate plots of the response versus each of the two predictors. From the figure we see that the effect of day is certainly nonlinear, but whether wind acts nonlinearly is not clear.

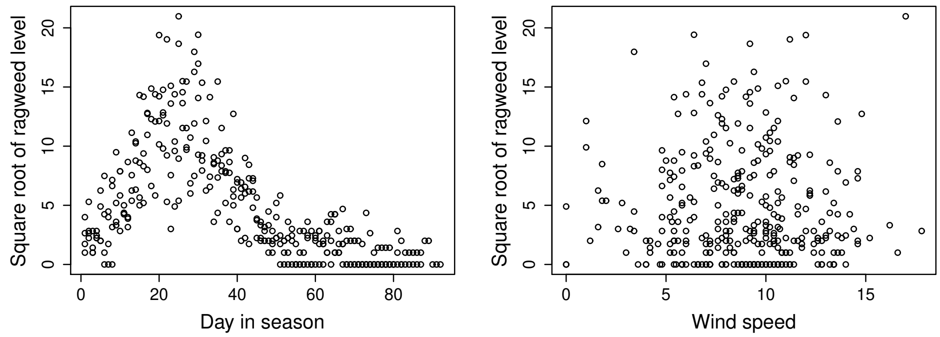

FIG 2. *Scatterplots of response against each of two predictors for the ragweed data set.*

We fit each of the two predictors nonparametrically via cubic regression splines involving 10 equally spaced knots. Hence the model we use has the form

$$Y_i = \beta_0 + \alpha_1 \mathsf{day}_i + \alpha_2 \mathsf{day}_i^2 + \alpha_3 \mathsf{day}_i^3 + \textstyle\sum_{t=1}^{10} \alpha_{t+3}(\mathsf{day}_i - \tilde{d}_t)_+^3$$
$$+ \beta_1 \mathsf{wind}_i + \beta_2 \mathsf{wind}_i^2 + \beta_3 \mathsf{wind}_i^3 + \textstyle\sum_{t=1}^{10} \beta_{t+3}(\mathsf{wind}_i - \tilde{w}_t)_+^3 + \epsilon_i,$$

for $i = 1, \dots, 335$, where $\tilde{d}_1 < \cdots < \tilde{d}_{10}$ represent the knots for the day explanatory variable, $\tilde{w}_1 < \cdots < \tilde{w}_{10}$ the knots for the wind explanatory variable, and $(x)_+ = \max\{0, x\}$. Note that there are 26 coefficients that could be set to 0, of which 20 correspond to knots along the domain of the two predictors. Our plan is to carry out the following two steps:

1. We form a point estimate and confidence region for $\arg\max_h m_y(h)$ by running a Markov chain.
2. We estimate the posterior distribution of θ when the prior is ν_{h_n}, where h_n is the estimate of $\arg\max_h m_y(h)$ obtained in Step 1, by running another Markov chain.

For Step 1 we ran a Markov chain of length $40{,}000$, using $h_1 = (.3, 100)$, from which we formed the surface $B_n(h)$, shown on the left panel of Figure 3. The argmax of the surface is $(.23, 176)$, and the 95% confidence region for $\arg\max_h m_y(h)$ is the ellipse shown in the right panel of Figure 3. For Step 2, we ran a new Markov chain, of length 10^5. For this chain, the highest probability model is the model which selects the variables wind, day^2, day^3, $(\mathsf{day} - \tilde{d}_3)_+^3$, and

$(\mathsf{day} - \tilde{d}_5)_+^3$. Interestingly, this model is the same as the model selected by the lasso, when we choose the tuning parameter by cross-validation.

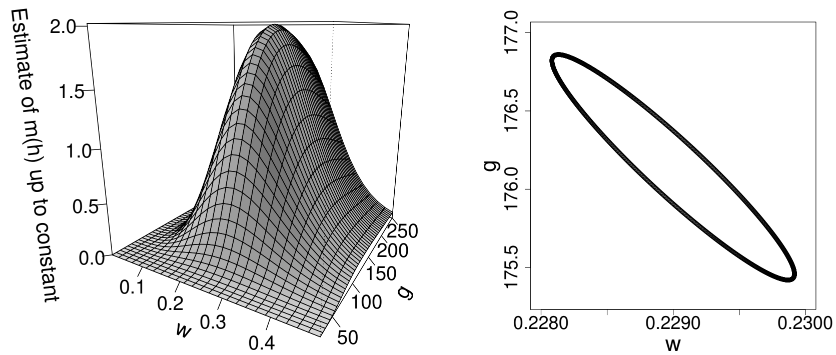

FIG 3. *Left Panel: Estimate of the marginal likelihood $m_y(h)$ (up to a multiplicative constant). The argmax is $(w_n, g_n) = (.23, 176)$, and the small value of w_n suggests a sparse model. Right Panel: Confidence region for $\arg\max_h m_y(h)$. The tight region indicates that the small Markov chain length used is adequate.*

Let $\mathcal{E}$ denote the ellipse. Our theory tells us that we are 95% confident that $\arg\max_h B(h) \in \mathcal{E}$, so we should run chains with posterior distributions $\nu_{h,y}$, $h \in \mathcal{E}$, and determine the highest posterior probability models for all $h \in \mathcal{E}$. By checking a few points on the boundary of the ellipse, we saw that the ellipse is narrow enough so that the highest probability model is the same for all $h \in \mathcal{E}$. Had this not been the case, we would have run the Step 1 chain for more cycles, getting a ellipse that is more narrow.

The value of w that is selected is small, which reflects sparsity: a small model is adequate for fitting the data. We now put our approach in the context of the existing literature. Liang et al. (2008) review methods for selecting g in the version of model (3.1) in which w is fixed at $1/2$. The literature has several data-independent choices (e.g. $g = \max(m, q^2)$), but these generally do not perform well. As a data-dependent choice, they propose $\hat{g} = \arg\max_g m_y(g)$, and to obtain it suggest an EM algorithm in which the model indicator γ is viewed as missing data. Unfortunately, the M-step in the algorithm involves a sum of 2^q terms. Unless q is relatively small, complete enumeration is not possible, and Liang et al. (2008) propose summing only over the most significant terms. However, determining which terms these are may be very difficult in some problems. Our approach provides a feasible way of obtaining the maximizer of the likelihood, and this for the model in which both w and g are unknown.

Acknowledgments. We thank the referees for their helpful comments.

SUPPLEMENTARY MATERIAL

Supplement to "An MCMC approach to empirical Bayes inference and Bayesian sensitivity analysis via empirical processes"
(doi: COMPLETED BY THE TYPESETTER; .pdf). In the supplement Doss and Park (2016) we provide proofs of Theorems 2, 4, 5, and 6, and Lemma 1. We also show that the key regularity condition (2.6), which is needed in the theorems in this paper, is satisfied in a large class of examples.

References.

ASUNCION, A., WELLING, M., SMYTH, P. and TEH, Y. W. (2009). On smoothing and inference for topic models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. UAI '09, AUAI Press, Arlington, Virginia, United States.

BERGER, J. O. (1994). An overview of robust Bayesian analysis (with discussion). *Test* **3** 5–124.

BLEI, D. M., NG, A. Y. and JORDAN, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research* **3** 993–1022.

BUTA, E. (2010). *Computational Approaches for Empirical Bayes Methods and Bayesian Sensitivity Analysis*. Ph.D. thesis, University of Florida.

DOSS, C., FLEGAL, J. M., JONES, G. L. and NEATH, R. C. (2014). Markov chain Monte Carlo estimation of quantiles. *Electronic Journal of Statistics* **8** 2448–2478.

DOSS, H. (2007). Bayesian model selection: Some thoughts on future directions. *Statistica Sinica* **17** 413–421.

DOSS, H. and PARK, Y. (2016). Supplement to "An MCMC approach to empirical Bayes inference and Bayesian sensitivity analysis via empirical processes".

DOSS, H. and TAN, A. (2014). Estimates and standard errors for ratios of normalizing constants from multiple Markov chains via regeneration. *Journal of the Royal Statistical Society,* Series B **76** 683–712.

FLEGAL, J. M., HARAN, M. and JONES, G. L. (2008). Markov chain Monte Carlo: Can we trust the third significant figure? *Statistical Science* **23** 250–260.

FLEGAL, J. M. and JONES, G. L. (2010). Batch means and spectral variance estimators in Markov chain Monte Carlo. *The Annals of Statistics* **38** 1034–1070.

GEORGE, C. P. (2015). *Latent Dirichlet Allocation: Hyperparameter Selection and Applications to Electronic Discovery*. Ph.D. thesis, University of Florida.

GEORGE, E. I. and FOSTER, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87** 731–747.

GEYER, C. J. (2011). Importance sampling, simulated tempering, and umbrella sampling. In *Handbook of Markov Chain Monte Carlo* (S. P. Brooks, A. E. Gelman, G. L. Jones and X. L. Meng, eds.). Chapman & Hall/CRC, Boca Raton, 295–311.

GEYER, C. J. and THOMPSON, E. A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association* **90** 909–920.

GRIFFITHS, T. L. and STEYVERS, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences* **101** 5228–5235.

HOBERT, J. P., JONES, G. L., PRESNELL, B. and ROSENTHAL, J. S. (2002). On the applicability of regenerative simulation in Markov chain Monte Carlo. *Biometrika* **89** 731–743.

IBRAGIMOV, I. A. and LINNIK, Y. V. (1971). *Independent and Stationary Sequences of Random Variables*. Wolters-Noordhoff, Groningen.

JONES, G. L., HARAN, M., CAFFO, B. S. and NEATH, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *Journal of the American Statistical Association* **101** 1537–1547.

KADANE, J. and WOLFSON, L. J. (1998). Experiences in elicitation. *Journal of the Royal Statistical Society: Series D (The Statistician)* **47** 3–19.

LEVENTAL, S. (1988). Uniform limit theorems for Harris recurrent Markov chains. *Probability Theory and Related Fields* **80** 101–118.

LIANG, F., PAULO, R., MOLINA, G., CLYDE, M. A. and BERGER, J. O. (2008). Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association* **103** 410–423.

MARINARI, E. and PARISI, G. (1992). Simulated tempering: A new Monte Carlo scheme. *Europhysics Letters* **19** 451–458.

MEYN, S. P. and TWEEDIE, R. L. (1993). *Markov Chains and Stochastic Stability*. Springer-Verlag, New York, London.

MITCHELL, T. and BEAUCHAMP, J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association* **83** 1023–1036.

MYKLAND, P., TIERNEY, L. and YU, B. (1995). Regeneration in Markov chain samplers. *Journal of the American Statistical Association* **90** 233–241.

NEWTON, M. and RAFTERY, A. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion). *Journal of the Royal Statistical Society,* Series B **56** 3–48.

PARK, Y. (2015). *A Markov Chain Monte Carlo Approach to Empirical Bayes Inference and Bayesian Sensitivity Analysis via Empirical Processes*. Ph.D. thesis, University of Florida.

PETRONE, S., ROUSSEAU, J. and SCRICCIOLO, C. (2014). Bayes and empirical Bayes: do they merge? *Biometrika* **101** 285–302.

ŘEHŮŘEK, R. and SOJKA, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta.

ROY, V. and HOBERT, J. P. (2007). Convergence rates and asymptotic standard errors for MCMC algorithms for Bayesian probit regression. *Journal of the Royal Statistical Society,* Series B **69** 607–623.

RUPPERT, D., WAND, M. and CARROLL, R. (2003). *Semiparametric Regression*. Cambridge University Press, Cambridge.

SMITH, M. and KOHN, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics* **75** 317–343.

STARK, P. C., RYAN, L. M., MCDONALD, J. L. and BURGE, H. A. (1997). Using meteorologic data to model and predict daily ragweed pollen levels. *Aerobiologia* **13** 177–184.

SUNG, Y. J. and GEYER, C. J. (2007). Monte Carlo likelihood inference for missing data models. *The Annals of Statistics* **35** 990–1011.

TAN, A. and HOBERT, J. P. (2009). Block Gibbs sampling for Bayesian random effects models with improper priors: convergence and regeneration. *Journal of Computational and Graphical Statistics* **18** 861–878.

TAN, Z. (2014). Self-adjusted mixture sampling and locally weighted histogram analysis. Tech. rep., Technical Report, Department of Statistics, Rutgers University.

WALLACH, H. M., MURRAY, I., SALAKHUTDINOV, R. and MIMNO, D. (2009). Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM.

WELLNER, J. (2005). Empirical processes: Theory and applications.
URL https://www.stat.washington.edu/people/jaw/RESEARCH/TALKS/Delft/emp-proc-delft-big.pd

WOLPERT, R. L. and SCHMIDLER, S. C. (2012). α-stable limit laws for harmonic mean estimators of marginal likelihoods. *Statistica Sinica* **22** 1233–1251.

ZELLNER, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti* (P. K. Goel and A. Zellner, eds.). Elsevier, New York.

DEPARTMENT OF STATISTICS
UNIVERSITY OF FLORIDA
GAINESVILLE, FL 32611
USA
E-MAIL: doss@stat.ufl.edu

DEPARTMENT OF BIOSTATISTICS
MD ANDERSON CANCER CENTER
HOUSTON, TX 77030
USA
E-MAIL: ypark3@mdanderson.org