Stochastic Convergence Rates and Applications of Adaptive Quadrature in Bayesian Inference

Blair Bilodeau, Alex Stringer, and Yanbo Tang

ABSTRACT
We provide the first stochastic convergence rates for a family of adaptive quadrature rules used to normalize the posterior distribution in Bayesian models. Our results apply to the uniform relative error in the approximate posterior density, the coverage probabilities of approximate credible sets, and approximate moments and quantiles, therefore, guaranteeing fast asymptotic convergence of approximate summary statistics used in practice. The family of quadrature rules includes adaptive Gauss-Hermite quadrature, and we apply this rule in two challenging low-dimensional examples. Further, we demonstrate how adaptive quadrature can be used as a crucial component of a modern approximate Bayesian inference procedure for high-dimensional additive models. The method is implemented and made publicly available in the aghq package for the R language, available on CRAN. Supplementary materials for this article are available online.

1. Introduction

The central challenge of Bayesian inference is computing the posterior distribution, which requires evaluating an integral—the normalizing constant or marginal likelihood—that is intractable in all but the simplest models. Numerical quadrature (or cubature in multiple dimensions) comprises a range of techniques for approximating deterministic integrals via function evaluations at finitely many points, and is a mature field of study in applied mathematics; see Davis and Rabinowitz (1975) for an overview. However, in Bayesian inference, the integrand is necessarily changing with the observed data, and consequently fixed quadrature rules can perform arbitrarily poorly by failing to capture the shifting mass of the integrand. To address this limitation, adaptive quadrature techniques based on shifting and scaling fixed quadrature rules using the mode and curvature of the integrand have been proposed in Bayesian inference since at least Naylor and Smith (1982). More recently, adaptive quadrature has been employed as a fundamental component of approximate Bayesian inference in the popular INLA framework (Rue, Martino, and Chopin 2009), for integrating out random effects (Pinheiro and Bates 1995; Cagnone and Monari 2013), and in the context of Bayesian inversion (Schillings and Schwab 2016). In the present work, we study stochastic convergence rates as the sample size tends to infinity for fixed parameter dimension and number of quadrature points. Theorem 1 and its corollaries provide the first stochastic convergence rates for the error in normalizing Bayesian posterior densities with adaptive quadrature, as well as for computing approximate moments and marginal densities.

Despite its broad applicability and usage by practitioners, relatively little is known about the theoretical properties of adaptive quadrature for Bayesian inference. Naylor and Smith (1982) discuss the practical application of what Tierney and Kadane (1986) call adaptive Gauss–Hermite quadrature (AGHQ) in Bayesian inference, arguing that it is a useful tool for normalizing posterior distributions and computing approximate summary statistics, but do not provide any theoretical guarantees on its accuracy. Tierney and Kadane (1986) use AGHQ to renormalize a Laplace approximation (which itself corresponds to AGHQ with a single quadrature point) of the marginal posterior density, however, they do not discuss the effect that the error introduced by this numerical renormalization may have on their convergence rate. Kass, Tierney, and Kadane (1990) rigorously establish convergence rates for the Laplace approximation in Bayesian inference, but their proof only applies to one-dimensional parameters, and they do not provide results for AGHQ with multiple quadrature points. More recently, for a specific, restricted class of one-dimensional functions that vary only through a scaling parameter \( n \), Jin and Andersson (2020) expand upon the work of Liu and Pierce (1994) to show that AGHQ with \( k \in \mathbb{N} \) quadrature points converges at relative rate \( O(n^{-(k+2)/3}) \), and comment that this rate holds in multiple dimensions using a specific extension of the univariate rule.

In statistical problems, the integrand varies through \( n \) both as a scaling factor and the data observed at time \( n \), which is not captured by the class of functions considered by Liu and Pierce (1994) and Jin and Andersson (2020). A stochastic convergence perspective is needed in order to quantify the behavior of AGHQ when used in fitting any statistical model, including when used...
to approximate the normalizing constant in Bayesian inference. Beyond the univariate Laplace analysis by Kass, Tierney, and Kadane (1990), a stochastic convergence perspective is also taken by Schillings and Schwab (2016) in the context of Bayesian inversion of operator equations, where they study convergence as the variance of the data tends to zero. Dick et al. (2019) study Bayesian PDE inversion using a Quasi-Monte Carlo variant of numerical quadrature. As is standard in the numerical analysis literature, they suppose that the data (rather than the model) satisfies certain regularity conditions, and provide a convergence rate as the number of quadrature points tends to infinity. Convergence limits as the data variability tends to zero and as the number of quadrature points tends to infinity for a fixed data sequence are both distinct from the asymptotic statistics perspective we take of letting the sample size n tend to infinity, and our results explicitly identify the effect of k on the approximation error.

In this work, we quantify the stochastic error of using adaptive quadrature rules to approximate the posterior distribution and summary statistics based on it. We demonstrate that the crucial property of AGHQ is that the underlying quadrature rule, Gauss-Hermite quadrature, exactly integrates the product of the Gaussian density and any polynomial of total order 2k − 1 or less for a well-chosen integer k in p dimensions; we call this property $\mathcal{P}(k, p)$ (see Definition 1). More precisely, under standard regularity assumptions, we prove that if the posterior normalizing constant is approximated using the adaptive form of any quadrature rule that satisfies $\mathcal{P}(k, p)$, then the relative error of the approximate normalizing constant to the true normalizing constant converges in probability at rate $O_p(n^{-(k+2)/3})$, where n is the number of observed data points (Theorem 1). The main technical contribution that enables this result is a precise quantification of how well the posterior distribution locally approximates a Gaussian distribution as a function of the data.

We further describe how to approximate marginal posterior densities and moments using a second application of AGHQ, and show that the convergence rate is preserved for the error of these approximate summary statistics (Corollaries 2 and 3). For approximate quantiles and credible sets, we show that the rate is preserved in the ideal case where one can exactly integrate the AGHQ approximation (Corollary 4 in the supplementary materials), and provide a computational method to approximate these. We include a simulation study (Section S.6 in the supplementary materials) that illustrates the stochastic nature of the convergence rates and demonstrates an example of a simple model in which our stated rate is achieved in finite samples empirically.

To illustrate the breadth of models for which adaptive quadrature provides a useful tool for inference, in Section 4 we directly apply the method to two challenging examples that satisfy the conditions of our theoretical results. For a distance-based, individual-level model for the spread of infectious disease, we show that the results obtained using AGHQ are nearly identical to those obtained via traditional sampling-based approaches by Almutiry, Warriyar, and Deardon (2020), at a substantial reduction in computational time. Additionally, we apply AGHQ to estimate the mass of the Milky Way galaxy using multivariate position and velocity measurements of star clusters and a complex astrophysical model.

In addition to the low-dimensional examples of Section 4, in Section 5 we demonstrate the applicability of adaptive quadrature for high-dimensional models by employing it within a broader method to fit a complex zero-inflated geostatistical binomial regression model, which we use to infer the spatial risk of contracting a certain tropical disease in West Africa, and for which Bayesian inferences had not previously been made. This example is particularly challenging and is not compatible with INLA or (to our knowledge) any other existing (non-MCMC) framework for making approximate Bayesian inferences, and we discuss some observed difficulties with applying MCMC to it as well.

AGHQ and the corresponding high-dimensional method are implemented in the aghq package in the R statistical programming language, made publicly available on the CRAN package repository. All code for the examples and simulations in this article is available in the supplementary materials.

2. Preliminaries
2.1. Bayesian Inference

Suppose that we observe a dataset $Y^{(n)} = (Y_1, \ldots, Y_n) \subseteq \mathbb{R}^d$ generated from some unknown probability distribution. We fix a model for the data defined by a parameter space $\Theta \subseteq \mathbb{R}^p$ and likelihood $\pi(Y^{(n)} \mid \theta)$. Often, the model assumes iid data, in which case the likelihood factors into $\pi(Y^{(n)} \mid \theta) = \prod_{i=1}^n \pi(Y_i \mid \theta)$, but we do not require this restriction. Further, we do not require that the model be well-specified; the only constraint on the data-generating distribution is that it satisfies certain regularity assumptions for the chosen model.

For some prior density $\pi(\cdot)$ on $\Theta$, the object of inferential interest is the posterior density

$$
\pi(\theta \mid Y^{(n)}) = \frac{\pi(\theta, Y^{(n)})}{\int_\Theta \pi(\theta', Y^{(n)}) d\theta'},
$$

where $\pi(\theta, Y^{(n)}) = \pi(\theta)\pi(Y^{(n)} \mid \theta)$ is the unnormalized posterior density. Inference for $\theta$ is based upon point and interval estimates computed from $\pi(\theta \mid Y^{(n)})$.

All posterior summaries (marginals, moments, quantiles, etc.) require knowledge of the normalized posterior distribution, which requires computing the denominator of Equation (1), referred to as the marginal likelihood or normalizing constant,

$$
\pi(Y^{(n)}) = \int_\Theta \pi(\theta', Y^{(n)}) d\theta'.
$$

The computation of this integral—as well as the further integration required to compute posterior summary statistics—is typically not analytically tractable, so inference is instead conducted using integral approximations.

2.2. Numerical Quadrature

A quadrature rule for approximating an integral $F = \int f(\theta) d\theta$ of a function $f : \Theta \rightarrow \mathbb{R}$ is a collection of points $Q \subseteq \Theta$ and
a weight function $\omega : Q \rightarrow \mathbb{R}$, and is denoted by $\mathcal{R}(Q, \omega)$. The approximate integral under such a rule is denoted by

$$\widetilde{F}(Q, \omega) = \sum_{z \in Q} f(z) \omega(z),$$

which we denote by just $\widetilde{F}$ when $\mathcal{R}(Q, \omega)$ is clear. Often, quadrature rules are designed to be exact for specific functions of interest, most often polynomials. A $p$-dimensional polynomial $P$, defined by

$$P(x_1, \ldots, x_p) = \sum_{i=1}^{T} a_i \prod_{d=1}^{p} j_{d}^{i_d}$$

for some $T \in \mathbb{N}$, $(a_1, \ldots, a_T) \in \mathbb{R}^T$, and $(j_1, \ldots, j_p) \in \mathbb{N}^p$, is said to have total order $\max_{i \in [T]} \sum_{d=1}^{p} j_{d}^{i_d}$ (Heiss and Winschel 2008).

Definition 1. For any $k, p \in \mathbb{N}$, a quadrature rule $\mathcal{R}(Q, \omega)$ satisfies $\mathcal{R}(k, p)$ if for all $p$-dimensional polynomials $P : \Theta \rightarrow \mathbb{R}$ of total order $2k - 1$ or less,

$$\int_{\Theta} \phi(\theta; 0, I_p) P(\theta) d\theta = \sum_{z \in Q} \phi(z; 0, I_p) P(z) \omega(z), \quad (2)$$

where $\phi(\theta; 0, I_p)$ is the standard $p$-dimensional Gaussian density.

The choice of the multivariate Gaussian density in Definition 1 is strategic, since in many parametric models the posterior distribution asymptotically (in sample size) converges to a Gaussian distribution by the Bernstein von Mises theorem (van der Vaart 1998, chap. 10). For a nonstandard limit, another baseline density can be used in place of the Gaussian in Definition 1, but we do not consider such models at present.

Consider the univariate case (i.e., $p = 1$) with $\Theta = \mathbb{R}$. For $k \in \mathbb{N}$, let $H_k$ be the $k$th Hermite polynomial, defined for all $z \in \mathbb{R}$ by

$$H_k(z) = (-1)^k e^{z^2/2} \frac{d^k}{dz^k} e^{-z^2/2}, \quad (3)$$

denote its zeroes by $z_1^k, \ldots, z_n^k$. These zeroes are distinct, symmetric about 0, and include 0 if and only if $k$ is odd. The Gauss–Hermite quadrature (GHQ) rule uses the points $Q = \{z_j^k\}_{j \in [k]}$ and the weight function

$$\omega(z_j^k) = \frac{k!}{[H_{k+1}(z_j^k)]^2 \phi(z_j^k)}, \quad (4)$$

where $\phi(\cdot)$ denotes the standard normal density. It is well known (e.g., eq. (3.6.11) of Davis and Rabinowitz 1975) that GHQ satisfies $\mathcal{R}(k, 1)$ using the smallest number of points $k$ possible. Hence, a naive application of GHQ for integrating $f$ may be expected to perform well in the case that $f$ is centered at $\theta = 0$ and $f(\theta) / \phi(\theta)$ is well-approximated by a univariate polynomial of total order $2k - 1$ or less. For a strict subset $\Theta \subset \mathbb{R}$, there exist finite interval Gauss quadrature rules (Section 2.7 of Davis and Rabinowitz, 1975; Theorem 1 of Bojanov and Petrov, 2001) that satisfy $\mathcal{R}(k, p)$.

Quadrature rules in $p$ dimensions are formed by combining a univariate quadrature rule for each dimension. The most common combination technique is the product rule. If $Q \subseteq \mathbb{R}$ is a collection of univariate quadrature points with weight function $\omega : \mathbb{R} \rightarrow \mathbb{R}$, the product rule induced multivariate quadrature rule is defined by $Q = Q^p$ and $\omega(z) = \prod_{i=1}^{p} \omega(z_i)$ for any $z = (z_1, \ldots, z_p) \in Q$. GHQ with the product rule satisfies $\mathcal{R}(k, p)$ (see Section 5.6 of Davis and Rabinowitz 1975), and requires $k^p$ quadrature points. Sparse rules are also available that satisfy $\mathcal{R}(k, p)$, see Section 3.

In addition to $\mathcal{R}(k, p)$, a commonly desired property of quadrature rules is that they are symmetric, which we formalize now in the context of our problem. We note that GHQ with both product rule and the sparse rule we consider is symmetric, and hence restricting ourselves to symmetric rules is benign.

Definition 2. $\mathcal{R}(Q, \omega)$ is symmetric if for all $z \in Q$, $-z \in Q$ and $\omega(z) = \omega(-z)$.

For Bayesian inference, we are interested in integrating functions that depend on $Y^{(n)}$, such as $f(\theta) = \pi(\theta, Y^{(n)})$. A limitation of numerical quadrature rules is that the points and weights remain fixed regardless of the shape and location of $\pi(\theta, Y^{(n)})$. As $n \rightarrow \infty$, standard Bayesian asymptotic theory (van der Vaart 1998, chap. 10) guarantees that with high probability over $Y^{(n)}$, the posterior mode concentrates to some $\hat{\theta}^*$ and the variance at the mode tends to 0. Consequently, any fixed rule will miss most of the mass of $\pi(\theta, Y^{(n)})$, and this problem worsens as $n \rightarrow \infty$. A procedure that explicitly adapts to the changing location and shape of the posterior density is necessary to obtain statistical performance guarantees that hold assuming only standard regularity conditions on the model.

2.3. Adaptive Quadrature

Naylor and Smith (1982) introduced a technique that was eventually named adaptive Gauss–Hermite quadrature (AGHQ) by Tierney and Kadane (1986), which we extend here to adaptive quadrature in general.

Given a function $f$ to integrate (which one expects is well-approximated by a Gaussian density), define the mode, curvature at the mode, and Cholesky decomposition of the inverse curvature by

$$\hat{\theta} = \text{arg max}_{\theta \in \Theta} f(\theta); \quad H(\hat{\theta}) = -\sigma^2 \log f(\hat{\theta}); \quad [H(\hat{\theta})]^{-1} = \hat{L}^T \hat{L}. \quad (5)$$

While we focus on the Cholesky decomposition for concreteness, any other matrix decomposition of this form could be used in our results. For any quadrature rule $\mathcal{R}(Q, \omega)$, the adapted integral approximation under this rule is

$$\widetilde{F}_{(Q, \omega)} = \sum_{z \in Q} f(\hat{L} z + \hat{\theta}) \omega(z).$$

2.4. Approximate Bayesian Inference

We will use approximate quadrature to approximate three integrals for Bayesian inference: the normalizing constant to obtain an
approximate posterior density, and then the further integration needed to obtain approximate marginal posterior densities and moments. For any approximation \( \tilde{\pi}(Y^{(n)}) \) of \( \pi(Y^{(n)}) \), the approximate posterior distribution is

\[
\tilde{\pi}(\theta \mid Y^{(n)}) = \frac{\pi(\theta, Y^{(n)})}{\tilde{\pi}(Y^{(n)})}.
\]

First, to approximate the normalizing constant, denote the analogous quantities of Equation (5) for the function \( f(\theta) = \pi(\theta, Y^{(n)}) \) by \( \tilde{\theta}, H_n(\tilde{\theta}), \) and \( \tilde{L}_n \). Then, for any quadrature rule \( \mathcal{R}(Q, \omega) \), the adapted approximate normalizing constant under this rule is

\[
\tilde{\pi}_{(Q, \omega)}^A(Y^{(n)}) = |\tilde{L}_n| \sum_{z \in Q} \pi(\tilde{L}_n z + \tilde{\theta}, Y^{(n)}) \omega(z).
\]

When \( \mathcal{R}(Q, \omega) \) is GHQ, the adaptive form is AGHQ by definition, and we denote the approximate normalizing constant by \( \tilde{\pi}^{AGHQ}(Y^{(n)}) \). When \( k = 1, z = 0 \) and \( \omega(z) = (2\pi)^{p/2} \), so the AGHQ approximation is actually a Laplace approximation (Tierney and Kadane 1986) and may be applied to integrals of any dimension without computational difficulties. In Section 5, we provide an example of how to combine AGHQ on low-dimensional parameters of interest with a Laplace approximation for high-dimensional nuisance parameters.

Second, to approximate the marginal posterior distribution, we apply AGHQ twice. Suppose the parameter can be decomposed into \( \theta = (\psi, \lambda) \) where \( \psi \in \mathbb{R}^d \), and we are interested in computing the marginal posterior density at \( \psi = \tilde{\psi}_0 \).

\[
\pi(\psi_0 \mid Y^{(n)}) = \frac{\int \pi(\psi_0, \lambda, Y^{(n)}) d\lambda}{\int \pi(\theta, Y^{(n)}) d\theta}.
\]

Define \( \tilde{\theta}^{\psi}_0 = \arg \max_{\theta \in \Theta} \pi(\theta, \lambda, Y^{(n)}) \) for \( \Theta(\psi_0) = \{ \theta \in \Theta : \theta = (\psi_0, \lambda) \} \), \( H_{\psi}^{\psi_0}(\theta) = -\partial^2 \log \pi(\psi_0, \lambda, Y^{(n)}) \), and \( H_{\psi}^{\psi_0}(\tilde{\theta}_0^{\psi}_0)^{-1} = \tilde{L}_n^{\psi}(\tilde{\theta}_0^{\psi}_0)^{-1} \). Then, given a fixed \( \mathcal{R}(Q, \omega) \) and \( \mathcal{R}(Q', \omega') \) of dimensions \( p \) and \( p - q \), respectively, the approximate marginal density is

\[
\tilde{\pi}(\psi_0 \mid Y^{(n)}) = \frac{1}{|\tilde{L}_n|} \sum_{z \in Q} \pi \left( (0, \tilde{L}_n^{\psi_0} z^T + \tilde{\theta}_0^{\psi_0}, Y^{(n)}) \omega(z) \right),
\]

which we denote by \( \tilde{\pi}^{AGHQ}(\psi_0 \mid Y^{(n)}) \) when both rules correspond to AGHQ.

Third, let \( \phi : \Theta \rightarrow \mathbb{R}_+ \) be any nonnegative function satisfying \( \int \phi(\theta) \pi(\theta) d\theta < \infty \). Denote the analogous quantities of Equation (5) for the function \( f(\theta) = \pi(\theta, Y^{(n)}) \) by \( \phi(\theta), H_n^\phi(\theta), \) and \( \tilde{L}_n^\phi \), when they exist (see Section S.4.3 in the supplementary materials). Then, given a fixed \( \mathcal{R}(Q, \omega) \), define

\[
\tilde{E}[\phi(\theta) \mid Y^{(n)}] = \frac{|\tilde{L}_n|}{|\tilde{L}_n|} \sum_{z \in Q} \pi(\tilde{L}_n z + \tilde{\theta}, Y^{(n)}) \omega(z),
\]

which we denote by \( \tilde{E}^{AGHQ}[\phi(\theta) \mid Y^{(n)}] \) when the rule is AGHQ.

### 3. Convergence Rates

In this section, we provide stochastic convergence rates for adaptive quadrature applied to adaptive quadrature as well as stochastic convergence rates for various summary statistics of inferential interest. All proofs are deferred to the supplementary materials. We denote probability under the true data-generating distribution by \( P_n^* \), and use \( C \) to denote a generic constant in \( n \) that may otherwise depend on \( p, k \), and the universal constants defined in Section S.1 in the supplementary materials.

#### 3.1. Approximate Posterior

**Theorem 1.** Suppose there exists \( m \geq 4 \) such that the likelihood of the data \( \pi(Y^{(n)} \mid \theta) \) is \( m \)-times differentiable as a function of \( \theta \) and the regularity assumptions of Section S.1 in the supplementary materials. For \( 1 \leq k \leq \lfloor m/2 \rfloor \), if \( \mathcal{R}(Q, \omega) \) is a symmetric quadrature rule satisfying \( \mathcal{S}(k, p) \) then

\[
\lim_{n \to \infty} P_n^* \left( \left| \frac{\pi(Y^{(n)})}{\tilde{\pi}_{(Q, \omega)}^A(Y^{(n)})} - 1 \right| \leq C n^{-\left(1+\frac{k}{2}\right)} \right) = 1.
\]

**Remark 1.** For AGHQ with \( k = 1 \), Theorem 1 recover the known \( O(n^{-1}) \) rate for the Laplace approximation (Kass, Tierney, and Kadane 1990).

**Remark 2.** If \( k > \lfloor m/2 \rfloor \) the rate for \( k = \lfloor m/2 \rfloor \) applies; this can be seen by reproducing the proof with a Taylor expansion of order \( m \) rather than one of order \( 2k \).

The following corollary demonstrates the utility of Theorem 1, and follows immediately from the algebra of Section S.2.1 in the supplementary materials and the definition of \( \tilde{\pi}_{(Q, \omega)}^A(\psi_0 \mid Y^{(n)}) \).

**Corollary 1.** Under the conditions of Theorem 1,

\[
\lim_{n \to \infty} P_n^* \left( \sup_{\theta \in \Theta} \left| \frac{\pi(\theta \mid Y^{(n)})}{\tilde{\pi}_{(Q, \omega)}^A(\theta \mid Y^{(n)})} - 1 \right| \leq C n^{-\left(1+\frac{k}{2}\right)} \right) = 1
\]

and

\[
\lim_{n \to \infty} P_n^* \left( \left\| \pi(\cdot \mid Y^{(n)}) - \tilde{\pi}_{(Q, \omega)}^A(\cdot \mid Y^{(n)}) \right\|_{\text{TV}} \leq C n^{-\left(1+\frac{k}{2}\right)} \right) = 1.
\]

The results of this section apply to any symmetric quadrature rule satisfying \( \mathcal{S}(k, p) \). In our applications, we focus on AGHQ defined using the product rule due to its simplicity and the fact that we have provided a robust implementation in the \( \text{aghq} \) package. It is of interest to compare alternatives empirically, such as the nested rule considered by Genz and Keister (1996), which is a Gaussian extension of the Gauss–Kronrod–Patterson construction.

Further, for multidimensional posteriors, our theoretical results apply to symmetric quadrature rules based on “sparse grids” as long as they satisfy \( \mathcal{S}(k, p) \). For example, Smolyak’s quadrature rule satisfies this criteria (Heiss and Winuschel 2008, Theorem 1), which reduces the dependence on the dimension from exponential to polynomial asymptotically. For specific \( k \) and \( p \), however, sparse rules may actually be more
computationally intensive than the product rule; for example, when \( k = 5 \) and \( p = 2 \), the product rule uses 25 quadrature points while the Smolyak rule uses 55.

The convergence rate depends directly on the number of quadrature points as follows. If one uses a product rule extension, then \( k^p \) quadrature points are needed to satisfy \( \mathcal{P}(k,p) \). Hence, if the model is of dimension \( p \) and one uses a product rule extension with \( |Q| \) quadrature points, the convergence rate will be \( n^{-\left(|Q|/|P_p|+2/3\right)} \).

### 3.2. Approximate Posterior Summaries

We now show that the convergence rate of Theorem 1 is realized for the approximations to marginal distributions and moments. Refer to Section 2.4 for definitions and Section S.5.2 in the supplementary materials for computational details.

**Corollary 2.** Fix the value of \( \psi\). Suppose the conditions in Theorem 1 are satisfied when replacing all instances of \( \theta \) with \( (\psi, \lambda), \theta^\ast \) with some constant \( \theta^\ast = (\psi, \lambda \theta^\ast) \), \( H_n \) with \( H_{n, \psi}^\ast \) and \( b_{\theta^\ast}^k (\cdot) \) with \( b_{\theta^\ast}^{k-1} (\cdot) \). Then

\[
\lim_{n \to \infty} \mathbb{P}_n \left( \frac{\tau(\psi_0 \mid Y^{(n)})}{\tau(\psi^\ast_0 \mid Y^{(n)})} - 1 \right) \leq C n^{-\left(\frac{k-2}{2}\right)} = 1.
\]

**Corollary 3.** Suppose \( g : \Theta 	o \mathbb{R}_+ \) satisfies assumptions (M1) through (M3) from Section S.4.3 in the supplementary materials. Then, if the conditions of Theorem 1 also hold,

\[
\lim_{n \to \infty} \mathbb{P}_n \left( \frac{\mathbb{E}[g(\theta) \mid Y^{(n)}]}{\mathbb{E}[g(\theta) \mid Y^{(n)}]} - 1 \right) \leq C n^{-\left(\frac{k-2}{2}\right)} = 1.
\]

Both Corollaries 2 and 3 require additional assumptions to be verified. In Sections S.4.2 and S.4.3 in the supplementary materials, we show that (a) Corollary 2 applies to all values of \( \psi \) in a \( n^{-1/2} \)-neighborhood of the unrestricted posterior mean (Proposition 1 in the supplementary materials) and (b) Corollary 3 applies to all marginal posterior moments (Proposition 2 in the supplementary materials).

### 3.3. Proof Sketch of Theorem 1

Finally, we provide a brief sketch of the proof of Theorem 1 to highlight the intuition for how convergence rates of the posterior inform the ultimate approximation error rate, and contrast our result with previous analyses of adaptive quadrature rules.

**Proof sketch (informal) of Theorem 1.** The proof quantifies the rate at which the posterior behaves locally Gaussian with polynomial error, combines this with the polynomial exactness property \( \mathcal{P}(k,p) \), and then quantifies that the contribution to the posterior mass outside of this local neighborhood is negligible with high probability. Specifically, the proof of Theorem 1 is composed of demonstrating that the following key facts hold with high probability asymptotically. We refer to the corresponding formal statements by their location in the supplementary materials.

1. There exists a neighborhood centered at a fixed parameter with radius defined by the curvature of the likelihood such that the likelihood is exponentially small outside of the neighborhood. See Lemma 6 in the supplementary materials for the precise statement.
2. Within this neighborhood, there exists a smaller neighborhood with radius decaying at rate \( \sqrt{\log(n)/n} \) such that the likelihood is polynomially small within the annulus outside of the shrinking neighborhood. See Lemma 7 in the supplementary materials for the precise statement.
3. A Taylor series expansion (with order depending on \( k \)) of the unnormalized posterior provides an accurate polynomial approximation within the shrinking neighborhood. See the proof of Lemma 4 in the supplementary materials for details.

The two most relevant works to our result are Kass, Tierney, and Kadane (1990) and Jin and Andersson (2020). We now contrast our proof with the analyses in both to highlight our technical contribution. First, we note that all of these works have only proved results for specifically AGHQ, while we have distilled the rate down to a simpler set of assumptions satisfied by more rules. Second, Kass, Tierney, and Kadane (1990) only prove the \( k = p = 1 \) case. They remark \( k > 1 \) is trivial, however, multivariate Taylor expansions lead to a product of sums rather than simply a sum of products, which consequently must be further upper bounded (see Equation (20) in supplementary materials).

Third, for the \( k > 1 \) case, higher-order derivatives are required, and Jin and Andersson (2020) sketch a proof for a limited class of functions in this setting. However, their analysis requires limiting assumptions that rule out posterior functions. Specifically, Jin and Andersson (2020) (inherit the assumptions of Liu and Pierce 1994) only allow integrands of the form \( \exp(n |\theta|) \) rather than \( \exp(n |\theta|) \), meaning \( \ell \) cannot depend on \( n \) and consequently also cannot depend on data; this eliminates all log-likelihoods.

Finally, neither Jin and Andersson (2020) nor Liu and Pierce (1994) provide an explicit argument for the order of the remainder terms from a Taylor series expansion. Handling these remainder terms is highly nontrivial, as can be seen by the proofs in Kass, Tierney, and Kadane (1990) and the supplementary materials. Specifically, this requires (a) identifying whether the remainder term is odd or even, (b) controlling the higher-order derivatives of the likelihood at various distances from the target parameter (undefined for Jin and Andersson 2020), and (c) applying probabilistic concentration results depending on the order of the remainder (also undefined for Jin and Andersson 2020, who take a deterministic approach).

### 4. Low-Dimensional Parameter Spaces

In the next two sections, we complement the theoretical results of Section 3 through three challenging examples, demonstrating the attractive computational properties of AGHQ for approximate Bayesian inference. In all of the examples, the quadra-
ture rule we use is AGHQ with the product rule extension to multiple dimensions. Because MCMC is arguably the most widely researched method for making approximate Bayesian inferences, and enjoys robust implementation in open-source software, we pay attention to the practical advantages of AGHQ compared to state of the art MCMC methods for the chosen examples.

4.1. Example: Modeling Infectious Disease Spread

We consider the popular Susceptible, Infectious, Removed (SIR) model for infectious disease spread as implemented in the EpiILMCT package in R (Almutiry, Warriyar, and Deardon 2020). Despite the low dimension of the parameter space, MCMC is the methodology of choice for fitting these models, leading to long run times and the need for specialized tuning and practical assessment of convergence. We demonstrate here that AGHQ gives fast and stable results that closely match the output of MCMC in a small fraction of the run time.

Almutiry, Warriyar, and Deardon (2020) consider an outbreak of Tomato Spotted Wilt Virus in \( n = 520 \) plants. Plants were grown on an even grid and checked for the virus every 14 days, a total of seven times. There were \( n_0 = 327 \) plants infected by the end of the study period. For each plant we observe infection times \( I_1 \leq \cdots \leq I_{n_0} \) and \( I_i = \infty \) for \( i = n_0 + 1, \ldots, n \). Plants may infect other plants while they are infected, and we observe associated removal times \( R_i, i \in [n] \) where a plant is no longer infectious. The likelihood function for these observed infection and removal times is given by

\[
\pi(I, R | \alpha, \beta) = \prod_{j=2}^{n_0} \left( \sum_{i: I_i < I_j \leq R_i} \lambda_{ij} \right) \exp \left\{ - \sum_{i=1}^{n_0} \sum_{j=1}^{n} \left[ \min(R_i, I_j) - \min(I_i, I_j) \right] \lambda_{ij} \right\},
\]

where \( \lambda_{ij} = \alpha d_{ij}^{-\beta} \) is the infectivity rate: the rate at which an infectious plant \( i \) passes the disease to a susceptible plant \( j \). Here \( d_{ij} \) is the Euclidean distance between plants \( i \) and \( j \), and \( \alpha, \beta > 0 \) are the parameters of inferential interest. Independent Exponential(0.01) priors are placed on \( \alpha, \beta \). As discussed in Section S.5 in the supplementary materials, we transform the parameters as \( \theta_1 = \log \alpha \) and \( \theta_2 = \log \beta \), perform the quadrature on this transformed scale, and then transform back when reporting results.

Figure 1 shows the posterior density estimates obtained using \( k = 3, 5, 7 \), and a comparison to a long MCMC run. The \( k = 5 \) and 7 results are visually indistinguishable from the density obtained through MCMC (and each other). Table 1 makes this more precise, with comparisons of posterior summaries of interest for \( k = 3, 5, 7, 9, 11, 13 \).

Table 2 shows the dramatic improvement in run time of AGHQ, as measured by the number of MCMC iterations (not including any time spent tuning the sampler) that could have been run in the same amount of time as it took to run the
Table 1. Posterior summaries are reported for AGHQ applied to the infectious disease data of Section 4.1.

| $k$ | $\alpha$ | $\beta$ | $\alpha$ | $\beta$ | $\alpha$ | $\beta$ | $\alpha$ | $\beta$ | $\alpha$ | $\beta$ |
|-----|------|------|------|------|------|------|------|------|------|------|
| 3   | 1.21 | 1.31 | 0.239 | 0.148 | 0.829 | 1.06 | 1.70 | 1.63 | 0.0326 | 0.0362 |
| 5   | 1.20 | 1.30 | 0.232 | 0.152 | 0.750 | 0.982 | 1.67 | 1.59 | 0.0234 | 0.0258 |
| 7   | 1.20 | 1.30 | 0.232 | 0.153 | 0.759 | 0.985 | 1.66 | 1.58 | 0.0132 | 0.0129 |
| 9   | 1.20 | 1.30 | 0.233 | 0.153 | 0.758 | 0.984 | 1.66 | 1.58 | 0.0126 | 0.0158 |
| 11  | 1.20 | 1.30 | 0.233 | 0.153 | 0.757 | 0.984 | 1.66 | 1.58 | 0.0168 | 0.0157 |
| 13  | 1.20 | 1.30 | 0.233 | 0.153 | 0.756 | 0.986 | 1.65 | 1.58 | – | – |
| MCMC | 1.20 | 1.30 | 0.228 | 0.151 | 0.761 | 0.984 | 1.65 | 1.58 | – | – |

NOTE: Comparison with MCMC is reported using the Kolmogorov-Smirnov (KS) distance. Mean, SD, and quantiles for $\alpha$ are multiplied by 100.

Table 2. Median (over 100 replications) run times to compute the marginal posteriors of both $\alpha$ and $\beta$ using AGHQ for the infectious disease data of Section 4.1.

| $k$ | Time (seconds) | Effective iterations |
|-----|---------------|---------------------|
| 3   | 0.101         | 12                  |
| 5   | 0.160         | 19                  |
| 7   | 0.224         | 27                  |
| 9   | 0.285         | 34                  |
| 11  | 0.357         | 42                  |
| 13  | 0.441         | 52                  |

NOTE: Effective iterations are the number of MCMC iterations that could have been performed in the same time it took to run AGHQ, calculated based on a maximum run time of 84 seconds for 10,000 MCMC iterations using 4 parallel chains in `tmbstan`.

full AGHQ procedure. Running MCMC for the maximum such number of iterations resulted in all such iterations being marked as divergent and NaN estimates for the number of effective parameters. This demonstrates the substantial computational gains attained by AGHQ in this simple example when compared to MCMC.

In practice, choosing $k$ remains an open question. As helpfully suggested by a referee, one strategy is to fit the model with successively increasing $k$ until inferences no longer change with $k$. Table 1 shows this occurring for the infectious disease example, where up to $k = 13$ was fit, with similar estimates from about $k = 7$ or so. Adding up the first row of Table 2, we see that the total time for this entire strategy is about 1.568 sec, or 186 total MCMC iterations, still a dramatic computational gain.

4.2. Example: Estimating the Mass of the Milky Way

Estimating the mass of the Milky Way Galaxy (hereafter the “Galaxy”) is of importance to astrophysicists interested in determining the amount of Dark Matter in the universe, among other things. Eadie and Harris (2016) describe a probabilistic model for estimating and, importantly, quantifying uncertainty in the mass of the Galaxy using Bayesian inference. They use three-dimensional observed position and velocity measurements of star clusters in orbit of the Galaxy within a probabilistic physical model whose parameters determine the mass of the Galaxy at any radial distance from its center. The parameters are subject to nonlinear constraints and are found to have strongly correlated, highly skewed posteriors with mode lying on or near the boundary of the parameter space (Eadie and Harris 2016). Care is required in implementing AGHQ for this problem.

The choice of priors was observed to have a substantial effect on inference in this problem (Eadie and Harris 2016), and a large body of knowledge on how to do this is available from the underlying physics. Eadie and Harris (2016) consider many different choices of priors and subsets of their data and the effect that this has on the estimated mass of the Galaxy. They use MCMC for inference, where each new model fit in their application requires careful tuning and assessment of the chains as well as potentially inconvenient run times. We find that AGHQ exhibits fast and stable performance in this challenging problem, although we note that our present implementation with `tmbstan` (software that was not available at the time Eadie and Harris 2016, was written) seems to avoid some of the reported challenges with MCMC as well. Nonetheless, this example serves to illustrate the application of AGHQ in a challenging applied problem.

Let $Y_i = (y_{i1}, y_{i2}, y_{i3})$ denote the three measurements for each star cluster: position, radial velocity, and tangential velocity relative to the center of the Galaxy (referred to as galactocentric measurements), and let the full matrix of data be $Y^{(n)} = \{Y_i : i \in [n]\}$. There are $n = 70$ clusters with complete measurements. The probability density for $Y_i$ is

$$f(Y_i; \Psi_0, \gamma, \alpha, \beta) = \frac{L_i^{-2\beta} e^{\beta(y_{i1} - 1)^2 + \beta(y_{i2} - 0)^2 + \beta(y_{i3} - 0)^2}}{8\pi^{3/2} 2^{-2\beta} \Psi_0^n \Gamma(\alpha - 2\beta + 1) \Gamma(\beta(2 - \gamma) + \gamma - 1)}$$

where $L_i = y_{i1} y_{i3}, E_i = \Psi_0 y_{i1}^{\beta - 1}, \theta = (\Psi_0, \gamma, \alpha, \beta)$ determine the mass of the Galaxy at radial distance $r$ kiloparsecs (kpc) from its center according to $M(r) = \Psi_0 r^{1 - \gamma}$. While $M(r)$ only directly depends on $\Psi_0$ and $\gamma$, its posterior will depend indirectly on all four parameters due to correlation between them.

Eadie and Harris (2016) consider many different strongly informative priors for the four model parameters. We choose one configuration of theirs: $\Psi_0 \sim \text{Unif}(1, 200), \gamma \sim \text{Unif}(0.3, 0.7), \alpha - 3 \sim \text{Gamma}(1, 4.6), \beta \sim \text{Unif}(-0.5, 1)$. The parameters are further subject to nonlinear constraints $\alpha > \gamma, \alpha > \beta(2 - \gamma) + \gamma / 2, \text{and } E_i > 0, i \in [n]$.

We find the following transformations convenient in this example:

$$\theta_j = \log \left(- \log \left[ \frac{\Xi_j - a_j}{b_j - a_j} \right] \right), j = 1, 2, 4, \quad \theta_3 = \log(\alpha - 3),$$

where $(a_j, b_j)$ are the endpoints of the uniform prior for $\Xi_j$. We let $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$ and normalize the posterior $\pi(\theta | Y^{(n)})$ using AGHQ with $k = 5$. We emphasize that these transformations are not required to apply the theoretical results of Section 3, and refer the reader to Section 5.5 in the supplementary materials for further discussion about implementation. To find the posterior mode accounting for the remaining nonlinear constraints, we perform a constrained optimization using the IPOPT software (Wachter and Biegler 2006) with derivatives of
adaptive quadrature is an increasingly popular technique in modern Bayesian statistics as one important component of more complicated methods for approximate posterior inference in models with high-dimensional parameter spaces. In this section we demonstrate the use of one such type of method, based on the INLA method of Rue, Martino, and Chopin (2009), through fitting a spatial model for zero-inflated counts, for which MCMC-based inference is observed to be challenging.

The methods described in this section have no known convergence theory, and their usefulness in applied Bayesian statistics makes development of such theory a topic of substantial current interest. Theorem 1, which describes the convergence properties of the adaptive quadrature rules used at the core of these methods, is a first step in this direction.

5.1. High-Dimensional Approximation Method

Consider a parameter vector \((w, \theta)\) where \(\theta \in \mathbb{R}^p\) and \(w \in \mathbb{R}^m\) with \(p \ll m\). Bayesian inferences for these parameters are made using the posterior distributions:

\[
\pi(\theta \mid Y^{(n)}) = \frac{\int \pi(w, \theta, Y^{(n)})dw}{\int \int \pi(w, \theta, Y^{(n)})dwd\theta}
\]

\[
\pi(w \mid Y^{(n)}) = \int \pi(w \mid \theta, Y^{(n)})\pi(\theta \mid Y^{(n)})d\theta.
\]

It is assumed that \(p\) is small enough to make it computationally feasible to directly apply adaptive quadrature to \(d\theta\) integrals, but that \(m\) is large enough for this to be infeasible for \(dw\) integrals, even using sparse grids or other nonproduct rule extensions to

---

**Table 3.** Comparison of AGHQ with \(k = 5\) to MCMC using the KS distance for the astronomy data of Section 4.2.

| Param. | \(\Psi_0\) | \(\gamma\) | \(\alpha\) | \(\beta\) |
|--------|-------------|-------------|-------------|-------------|
| KS(AGHQ,MCMC) | 0.00872 | 0.00844 | 0.0358 | 0.00739 |

---
multiple dimensions. This occurs, for example, in hierarchical models (Kass and Steffey 1989; Rue, Martino, and Chopin 2009; Wood, Pya, and Säfken 2016; Geirsson et al. 2020; Section 5.2) where \( w \) typically relate to the mean response, and \( \theta \) are variance components.

For any fixed \( \theta \), Tierney and Kadane (1986) suggest approximating \( \pi(\theta | Y^{(n)}) \approx \pi_{LA}^{AGHQ}(\theta | Y^{(n)}) \) by first approximating \( \int \pi(w, \theta, Y^{(n)})dw \approx \pi_{LA}(\theta, Y^{(n)}) \) using AGHQ with \( k = 1 \) (a Laplace approximation), and then renormalizing the result using numerical integration, for which they also use AGHQ in their experiments. Stringer, Brown, and Stafford (2021) combine this approximation with a Gaussian approximation \( \pi(w | \theta, Y^{(n)}) \approx \pi_{G}(w | \theta, Y^{(n)}) \), obtaining

\[
\tilde{\pi}(w | Y^{(n)}) \approx \int \pi_{G}(w | \theta, Y^{(n)}) \pi_{LA}^{AGHQ}(\theta | Y^{(n)}) d\theta. \tag{10}
\]

The integration in Equation (10) is approximated with the same AGHQ points and weights used to obtain \( \pi_{LA}^{AGHQ}(\theta | Y^{(n)}) \), so that \( \tilde{\pi}(w | Y^{(n)}) \) corresponds to a discrete mixture of Gaussian approximations with weights determined by AGHQ. Inferences for \( w \) are then made by sampling from this Gaussian mixture. The INLA method of Rue, Martino, and Chopin (2009) uses an alternative adaptive quadrature rule for the renormalization, and then another Laplace approximation to approximate the marginal distributions \( \pi(w | \theta, Y^{(n)}) \).

There is a growing body of evidence suggesting that approximations based on Equation (10) give results empirically similar to those returned by MCMC and other methods (Rue, Martino, and Chopin 2009; Brown 2011; Taylor and Diggle 2014; Stringer, Brown, and Stafford 2021; Wood 2020) in faster computational times. In Section 5.2 we show an example of a model for which a state-of-the-art MCMC algorithm runs for days and fails to converge (Section S.7.1 in the supplementary materials) to a suitable solution, while Equation (10) provides a potentially suitable (Section S.7.3 in the supplementary materials) solution in minutes. However, we stress that the convergence properties of Equation (10) are not known, and the apparent practical utility of this approximation makes establishing such properties an important area of research. Because AGHQ is used several times in computation of Equation (10), Theorem 1 is a first step toward this broader goal.

### 5.2. Example: Zero-Inflated Geostatistical Binomial Regression

Diggle and Giorgi (2016) introduce a zero-inflated geostatistical binomial regression model, where both the incidence rate and suitability of infection (zero-inflation probability) varies spatially. They argue that such models are of substantial importance in the mapping of tropical diseases, and make frequentist inferences for the parameters of interest. Here we make Bayesian inferences for the spatial patterns in incidence and suitability of infection of a tropical disease in Nigeria and Cameroon, based on a dataset of subjects who tested positive in \( n = 190 \) villages in this region (Giorgi, Schluter, and Diggle 2018). Data are obtained from the loaloa object in the geostatsp package (Brown 2011). A simpler model that does not allow for zero-inflation has been fit using INLA (Brown 2011) as well as MCMC and maximum likelihood (Giorgi and Diggle 2017). To our knowledge, no previous Bayesian implementation of this zero-inflated model exists.

We apply Equation (10) to fit this model. Let \( 0 \leq y_i \leq N_i, i \in [n] \) represent the counts of people infected and total number of people in the \( i \)th village out of the \( n = 190 \) included in the data, and let \( s_i \in \mathbb{R}^2 \) denote the geographical coordinates of this village. For every location \( s \in \mathbb{R}^2 \), let \( \rho(s) \) denote the probability that this location is capable of disease transmission (the suitability probability), and \( p(s) \) denote the probability that transmission occurs at this location, conditional on it being suitable (the incidence probability). Diggle and Giorgi (2016) stress the practical importance of allowing observed zero counts \( y_i = 0 \) to either be haphazard zeroes arising from sampling variability, or structural zeroes arising from a location being unsuitable for disease transmission. They also discuss how this makes joint inference of the underlying spatial fields governing suitability and incidence very challenging. The full model we consider is

\[
P(Y_i = y | \rho(s_i), \phi(s_i)) = [1 - \phi(s_i)]^I (y_i = 0) + \phi(s_i) \times \text{Binomial}(y_i; N_i, p(s_i)),
\]

\[
\log \left[ \frac{\phi(s)}{1 - \phi(s)} \right] = \beta_{\text{suit}} + u(s) \quad \text{and} \quad \log \left[ \frac{p(s)}{1 - p(s)} \right] = \beta_{\text{inc}} + v(s), s \in \mathbb{R}^2
\]

where the unknown functions \( u(\cdot), v(\cdot) \) are modeled as independent Gaussian Processes with the same Matérn covariance function, \( C_\theta \), with \( \theta = (\sigma, \rho) \) and the two intercepts are given independent Gaussian priors with variance 1000. We assign \( \sigma \) and \( \rho^{-2} \) independent exponential priors satisfying \( P(\rho < 200km) = P(\sigma < 4) = 97.5\% \), following Brown (2011) and Fuglstad et al. (2019).

Inference for \( u(\cdot) \) and \( v(\cdot) \) is based on their values at the observed locations \( s_i \), and then posterior distributions for their values at any new location \( s \in \mathbb{R}^2 \) are obtained using standard methods for spatial interpolation. Define \( U = \{u(s_i) : i \in [n]\} \), \( V = \{v(s_i) : i \in [n]\} \), and let \( w = (U, \beta_{\text{suit}}, V, \beta_{\text{inc}}) \in \mathbb{R}^m, m = 2n + 2 \). The Gaussian process priors on \( u(\cdot) \) and \( v(\cdot) \) imply that \( U | \theta \sim N(0, \Sigma(\theta)) \) and \( V | \theta \sim N(0, \Sigma(\theta)) \) independently, where \( \Sigma(\theta)_{ij} = C_\theta(\|s_i - s_j\|), i, j \in [n] \). To infer \( U^* = \{u(s^*_i) : t \in [T]\} \) and \( V^* = \{v(s^*_i) : t \in [T]\} \) for any set of new locations \( \{s^*_i : t \in [T]\} \subseteq \mathbb{R}^2 \), we simulate from the predictive distribution \( (U^*, V^*) | Y \) by first drawing \( w \) from \( \tilde{\pi}(w | Y^{(n)}) \) using standard methods (Rue 2001), and then sampling from \( (U^*, V^*) | w \) using existing algorithms for conditional simulation of Gaussian fields, implemented in the geostatsp (Brown 2011) and RandomFields (Schlather et al. 2015) packages.

We fit the model using AGHQ with \( k = 7 \) and the approximations described in Section 5.1, and show the resulting spatial interpolations on a fine grid in Figure 4. Total computation time for parameter estimation was 225 sec. The predicted incidence probabilities appear visually similar to those reported by Brown (2011) and Giorgi and Diggle (2017) for the simpler model without zero-inflation, and the novel plot of predicted suitability probabilities identifies a cluster of villages that have a low posterior probability of being suitable for transmission. Owing to the
lack of available convergence theory in this problem, we include
a brief simulation study in Section S.7.3 in the supplementary
materials to assess the empirical accuracy of this procedure for
this model and these data.

To better illustrate the difficulty of fitting this model with
existing methods, we fit the model using MCMC by running
the “NUTS” sampler (Hoffman and Gelman 2014) through the
tmbstan package (Monnahan and Kristensen 2018) using the
default settings. Eight chains of 10,000 iterations each (includ-
ing a 1000 iteration warmup) were run in parallel on a remote
server at a total “wall” computation time of 66 hr. The resulting
chains exhibited divergent transitions according to STAN’s built
in diagnostics. We investigate this in Section S.7.1 in the sup-
plementary materials, finding that $\beta_{suit}$ is poorly identified
by the sampler. We ran both Equation (10) and MCMC with $\beta_{suit}$
and $\beta_{inc}$ fixed at their estimated posterior means obtained
from the initial fit of Equation (10). This sampler converged
without warnings in just over 19 hr for 10,000 iterations. The
Kolmogorov-Smirnov (KS) statistics for the difference between
approximate marginal CDFs from MCMC and Equation (10)
indicate that the two procedures provide mostly comparable
inferences, with disagreement in a small number of villages. See
Section S.7.2 in the supplementary materials for further details.
We reiterate that MCMC did not produce a complete answer for
$\beta_{suit}$ in this problem.

Inferences made using Equation (10) produce a complete
answer in around three and a half minutes on a modern server
for this problem of substantial practical importance (Diggle and
Giorgi 2016). In this same problem and on the same hardware,
MCMC either (a) runs for almost a day and produces an incom-
plete answer, or (b) runs for almost 3 days and fails. This example
illustrates why these types of approximations have such high
potential value in applied statistics, and why convergence theory
for Equation (10) is of such importance. Theorem 1 provides a
first step toward this goal.

6. Discussion

Using standard regularity assumptions, we have provided the
first stochastic convergence rate for adaptive quadrature in
Bayesian inference, and showed that this rate applies to the
approximate normalizing constant, posterior density, moments,
and marginal densities. Using our R package aghq, available
on CRAN, we demonstrated the use of AGHQ for Bayesian
inference in two challenging low-dimensional models and one
high-dimensional model. We now briefly discuss five open
problems that remain for the theory of adaptive quadrature
in Bayesian inference.

First, computing approximate quantiles and credible sets
requires further integration of the approximate posterior over
a subset of the parameter space, and hence a quadrature rule
is needed that satisfies a truncated version of $P(p, k)$. Providing
a robust method for this computation with correspond-
ing theoretical guarantees (analogous to Corollaries 2 and 3)
will complete the justification of using AGHQ for all facets
of Bayesian inference in low-dimensional models. The current
implementation uses an interpolation-based method with no
theoretical guarantees, but appears to provide reasonable output
in challenging examples. Second, for high-dimensional models,
the current implementation uses a Gaussian approximation and
an adaptive quadrature approximation with reused points and
weights. Providing full theoretical guarantees for the output of
this entire procedure remains an open problem, and will not
only validate the use of the aghq package for such models
but also provide the first theoretical guarantees for INLA-like
methods; we believe that Theorem 1 is an important first step
toward this goal. Third, our theoretical guarantees are all asym-
ptotic and worst-case subject to the regularity assumptions.
A challenging open problem is to provide theoretical guarantees
that hold for finite samples and adapt to properties such as
smoothness and sparsity, leading to improved performance for
“benign” data and models. Fourth, a principled choice of $k$ in
any given practical application, for any given dataset, remains
an open problem. The recommendation from Section 4.1 is feasible
due to the fast run time of AGHQ, and a more formally moti-
vated approach based on this could lead to a clearer and more
useful practical recommendation. Lastly, developing methods
with comparable accuracy to AGHQ that are computationally
feasible in very high dimensions remains a challenging open
problem.

Supplementary Materials

The supplementary materials are available at https://www.tandfonline.com/
doi/full/10.1080/01621459.2022.2141635, where the proofs of the main
theorems and the code necessary to produce the figures in the article can
be found.
Acknowledgments

We thank Jeffrey Negrea, Nancy Reid, Daniel Roy, and Jamie Stafford for helpful comments and suggestions.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

BB acknowledges support from an NSERC Canada Graduate Scholarship and the Vector Institute. AS acknowledges support from an NSERC Postgraduate Scholarship and the Centre for Global Health Research at St. Michael’s Hospital, Toronto, Canada. YT acknowledges support from an NSERC Postgraduate Scholarship and the Vector Institute.

ORCID

Blair Bilodeau https://orcid.org/0000-0002-3933-1427
Alex Stringer https://orcid.org/0000-0002-4133-6884
Yanbo Tang https://orcid.org/0000-0002-6228-7049

References

Almquist, W., Warriyar, K. V., and Deardon, R. (2020), “Continuous Time Individual-Level Models of Infectious Disease: EpiliMCT,” arXiv:2006.00135v1. [691,695]
Bojanov, B., and Petrov, P. P. (2001), “Gaussian Interval Quadrature Formula,” Numerische Mathematik, 87, 625–643. [692]
Brown, P. (2011), “Model-based Geostatistics the Easy Way,” Journal of Statistical Software, 73, 423–498. [698]
Cagnone, S., and Monari, P. (2013), “Latent Variable Models for Ordinal Data by using the Adaptive Quadrature Approximation,” Computational Statistics, 28, 597–619. [690]
Davis, P. J., and Rabinowitz, P. (1975), Methods of Numerical Integration, New York: Academic Press. [690,692]
Dick, J., Gantner, R. N., Le Gia, Q. T., and Schwab, C. (2019), “Higher Order Quasi-Monte Carlo Integration for Bayesian PDE Inversion,” Computers and Mathematics with Applications, 77, 144–172. [691]
Diggle, P. J., and Giorgi, E. (2016), “Model-based Geostatistics for Prevalence Mapping in Low-Resource Settings,” Journal of the American Statistical Association, 111, 1096–1120. [698,699]
Eadie, G. M., and Harris, W. E. (2016), “Bayesian Mass Estimates of the Milky Way: The Dark and Light Sides of Parameter Assumptions,” The Astrophysical Journal, 829. [696]
Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H. (2019), “Constructing Priors that Penalize the Complexity of Gaussian Random Fields,” Journal of the American Statistical Association, 114, 445–452. [698]
Geirsson, O. P., Hrafnkelsson, B., Simpson, D., and Sigurdarson, H. (2020), “LGM Split Sampler: An Efficient MCMC Sampling Scheme for Latent Gaussian Models,” Statistical Science, 35, 218–233. [698]
Genz, A., and Keister, B. D. (1996), “Fully Symmetric Interpolatory Rules for Multiple Integrals over Infinite Regions with Gaussian Weight,” Journal of Computational and Applied Mathematics, 71, 299–309. [693]
Giorgi, E., and Diggle, P. J. (2017), “PrevMap: An R Package for Prevalence Mapping,” Journal of Statistical Software, 78, 1–29. [698]
Giorgi, E., Schluter, D. K., and Diggle, P. J. (2018), “Bivariate Geostatistical Modelling of the Relationship between Loa loa Prevalence and Intensity of Infection,” Environmetrics, 29, e2447. [698]
Heiss, F., and Winschel, V. (2008), “Likelihood Approximation by Numerical Integration on Sparse Grids,” Journal of Econometrics, 144, 62–80. [692,693]
Hoffman, M. D., and Gelman, A. (2014), “The no-U-turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo,” Journal of Machine Learning Research, 15, 1593–1623. [699]
Jin, S., and Anderson, B. (2020), “A Note on the Accuracy of Adaptive Gauss-Hermite Quadrature,” Biometrika, 107, 737–744. [690,694]
Kass, R., and Steffey, D. (1989), “Approximate Bayesian Inference in Conditionally Independent Hierarchical Models (Parametric Empirical Bayes Models),” Journal of the American Statistical Association, 84, 717–726. [698]
Kass, R. E., Tierney, L., and Kadane, J. B. (1990), “The Validity of Posterior Expansions based on Laplace's Method,” in Bayesian and Likelihood Methods in Statistics and Econometrics, eds. S. Geisser, J.S. Hodges, S.J. Press, and A. Zellner, pp. 473–488, Amsterdam: North Holland. [690,691,693,694]
Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., and Bell, B. M. (2016), “TMB: Automatic Differentiation and Laplace Approximation,” Journal of Statistical Software, 70, 1–21. [697]
Liu, Q., and Pierce, D. A. (1994), “A Note on Gauss-Hermite Quadrature,” Biometrika, 81, 624–629. [690,694]
Mannanahan, C., and Kristensen, K. (2018), “No-U-turn Sampling for Fast Bayesian Inference in ADMB and TMB: Introducing the adnuts and mbstan R Packages,” Plos One, 13, e0197954. [699]
Naylor, J., and Smith, A. F. M. (1982), “Applications of a Method for the Efficient Computation of Posterior Distributions,” Journal of the Royal Statistical Society, Series C, 31, 214–225. [690,692]
Pinheiro, J. C., and Bates, D. M. (1995), “Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model,” Journal of computational and Graphical Statistics, 4, 12–35. [690]
Rue, H. (2001), “Fast Sampling of Gaussian Markov Random Fields,” Journal of the Royal Statistical Society, Series B, 63, 325–338. [698]
Rue, H., Martino, S., and Chopin, N. (2009), “Approximate Bayesian Inference for Latent Gaussian Models by using Integrated Nested Laplace Approximations,” Journal of the Royal Statistical Society, Series B, 71, 319–392. [690,697,698]
Schillings, C., and Schwab, C. (2016), “Scaling Limits in Computational Bayesian Inversion,” ESAIM: Mathematical Modelling and Numerical Analysis, 50, 1825–1856. [690,691]
Schlather, M., Malinowski, A., Menck, P. J., Oesting, M., and Strokorb, K. (2015), “Analysis, Simulation and Prediction of Multivariate Random Fields with Package Randomfields,” Journal of Statistical Software, 63, 1–25. [698]
Stringer, A., Brown, P., and Stafford, J. (2021), “Approximate Bayesian Inference for Case Crossover Models,” Biometrics, 77, 785–795. [698]
Taylor, B. M., and Diggle, P. J. (2014), “INLA or MCMC? A Tutorial and Comparative Evaluation for Spatial Prediction in Log-Gaussian Cox Processes,” Journal of Statistical Computation and Simulation, 84, 2266–2284. [698]
Tierney, L., and Kadane, J. B. (1986), “Accurate Approximations for Posterior Moments and Marginal Densities,” Journal of the American Statistical Association, 81, 82–86. [690,692,693,698]
van der Vaart, A. W. (1998), Asymptotic Statistics, Cambridge: Cambridge University Press. [692]
Wachter, A., and Biegler, L. T. (2006), “On the Implementation of a Primal-Dual Interior Point Filter Line Search Algorithm for Large-Scale Nonlinear Programming,” Mathematical Programming, 106, 25–57. [696]
Wood, S. (2020), “Simplified Integrated Nested Laplace Approximation,” Biometrika, 107, 223–230. [698]
Wood, S., Pya, N., and Säfken, B. (2016), “Smoothing Parameter and Model Selection for General Smooth Models,” Journal of the American Statistical Association, 111, 1548–1575. [698]