Two-step Bayesian methods for generalized regression driven by partial differential equations

PRITHWISH BHAUMIK1 E-mail: pbhaumi@ncsu.edu
WENLI SHI1 E-mail: wenli_shi@ncsu.edu
and SUBHASHIS GHOSAL1 E-mail: sghosal@ncsu.edu
1Department of Statistics, North Carolina State University, 2311 Stinson Drive, Raleigh, North Carolina 27695-8203, USA

In certain non-linear regression models, the functional form of the regression function is not explicitly available, but is only described by a set of differential equations. For regression models described by a set of ordinary differential equations (ODEs), both Bayesian and non-Bayesian methods for inference were developed in the literature. In this paper, we consider a Bayesian approach to non-linear regression with respect to a multidimensional predictor variable given by a set of partial differential equations (PDEs). We consider a computationally convenient two-step approach by first representing the functions nonparametrically, constructing a finite random series prior using tensor products of B-splines and directly inducing a posterior distribution on parameter space through an appropriate projection map. By considering three different choices of the projection map, we propose three different approaches with their merits. We allow generalized non-linear regression with the response variable following an exponential family of distributions, extending the method beyond regression with additive normal errors. We establish Bernstein-von Mises type theorems which show $\sqrt{n}$-consistency and asymptotically correct frequentist coverage of Bayesian credible regions. We also conduct a simulation study to evaluate the finite sample performances of the proposed methods.

Keywords: Partial differential equation, Generalized regression, Two-step method, Projection posterior, Bernstein-von Mises theorem, Contiguity, B-splines, Tensor products.

1. Introduction

We consider a non-linear generalized regression for a response variable $Y$ on an $s$-dimensional predictor $t = (t_1, \ldots, t_s)$. Let $Y$ follow a natural exponential family density given by $p(y|\xi) = \exp[y\xi - \psi(\xi)], \xi \in \Xi$, an open interval in $\mathbb{R}$, with respect to a $\sigma$-finite measure $\nu$ and parameter $\xi = f_\theta(t)$ depending on $t$ and some parameter $\theta \in \Theta \subset \mathbb{R}^p$ indexing the model. We consider the situation where the regression functions are specified only indirectly by a set of partial differential equations (PDEs). Let $D^r$ stand for the mixed partial differential operator $\partial^{r_1} / \partial t_1^{r_1} \cdots \partial t_s^{r_s}$ of order $r = (r_1, \ldots, r_s), |r| = \sum_{j=1}^s r_j$. Let the regression function $f_\theta(t)$ be a solution of

$$F(t, (D^r f_\theta : |r| \leq \alpha), \theta) = 0,$$

where $\alpha$ is a positive integer and $F$ is a smooth binding function. The formulation in terms of the natural parameter is for the convenience of the presentation only; see Remark 2.1 for an extension
to any arbitrary parameterization. It is also possible to consider the situation where $Y$ (and hence $f_\theta$ and $F$) are vector-valued by slightly generalizing our proposed approach, but for simplicity of the notations, we shall restrict to the case where the response variable is a scalar. Our setup includes the situation $s = 1$, which leads to ordinary differential equation (ODE) models. Even for ODE models, our results are new as we cover generalized regression and also offer another method based on finite-difference quotients (see Subsection 2.2).

There is a rich literature on parameter estimation for ODE models. If the system of ODEs allow explicit solutions, the nonlinear least-squares (NLS) method (Levenberg [16], Marquardt [17]) may be used. However, in most situations, the solutions are not explicitly known, and hence the NLS method must be aided by numerical solutions such as the 4-stage Runge-Kutta algorithm (RK4) (Hairer et al. [13, page 134] and Mattheij and Molenaar [18, page 53]). Xue et al. [28] established the $\sqrt{n}$-consistency, asymptotic normality and asymptotic efficiency of the estimator, but this approach is computationally intensive. Bayesian computational methods using Markov chain Monte Carlo (MCMC) sampling with likelihood evaluated by numerical solutions of the ODEs were developed by Gelman et al. [9], Rogers et al. [21] and Girolami [11]. Baake et al. [1] proposed an estimator by viewing the ODE as a multi-point boundary value problem. Other Bayesian approaches were considered by Campbell and Steele [6] and Chkrebtii et al. [7], but theoretical properties of the posterior distribution were not explored in the literature.

Varah [25] used a two-step method for estimating $\theta$. In the first step, each state variable is approximated by a cubic spline using the least-squares technique. In the second step, the corresponding derivatives are estimated by differentiating the nonparametrically fitted curve and the estimator is obtained by minimizing the sum of squares of the difference between the derivatives of the fitted spline and the derivatives suggested by the ODEs at the design points. This method does not depend on the initial or boundary conditions of the state variables and is computationally very efficient irrespective of the complexity of the model. An example given in Voit and Almeida [26] showed the computational superiority of the two-step approach over the usual least-squares technique. Brunel [5] replaced the sum of squares of the second step by a weighted integral of the squared deviation and proved $\sqrt{n}$-consistency as well as asymptotic normality of the resulting estimator. The order of the B-spline basis was determined by the smoothness of $F(\cdot, \cdot, \cdot)$ in its first two arguments. Gugushvili and Klaassen [12] used the same approach but used kernel smoothing instead of spline smoothing. They also established the $\sqrt{n}$-consistency of the estimator. Other variations of the two-step method were considered by Wu et al. [27] and Dattner and Gugushvili [8]. Bhaumik and Ghosal [2, 3, 4] considered a Bayesian analog of the Varah-Brunel two-step method. The idea is to use a “projection-posterior distribution”, which is a useful tool when the original parameter space has complicated restrictions so that putting a prior and computing the posterior is difficult. The approach is to embed the class of regression functions, given by the ODE, in an infinite-dimensional space of smooth functions. A prior distribution is given by the standard Bayesian nonparametric technique of a random series based on a B-spline basis with normal coefficients. Then a posterior distribution is obtained and that is directly induced on the original parameter space through some appropriate projection map. Bhaumik and Ghosal [2, 3, 4] established the $\sqrt{n}$-consistency and asymptotic normality of the posterior distribution of $\theta$. In a regular parametric family indexed by a finite-dimensional parameter space, often a Bayesian credible set is also an approximately frequentist confidence set as a consequence of the celebrated Bernstein-von Mises theorem. The result asserts that in
large samples, the posterior distribution is approximately normal with a center at a point that is itself approximately normally distributed in the frequentist sense around the true value of the parameter with the same spread as the posterior distribution. In particular, this implies that the Bayesian credible sets are also approximately frequentist confidence set, justifying the Bayesian approach to uncertainty quantification in the frequentist sense. In differential equation models, even though the underlying family is regular parametric, a Bernstein-von Mises type theorem is not immediate for the indirect Bayesian procedures using the two-step method. Establishing a Bernstein-von Mises theorem in this setup requires establishing a posterior asymptotic normality result in a parametric family of increasing dimension and linearization for a projection map that induces the two-step posterior on the original parameter space.

PDE models are often encountered in fields like epidemiology, biology, and finance. In most situations, these systems are not analytically solvable. Compared with ODE models, estimation techniques for PDE models have been much less explored. Müller and Timmer [20] considered two methods — a two-step approach similar to Varah [25] using splines to estimate the parameters, and the other by obtaining a solution of the PDE for the full trajectory of the experimental dataset through a dynamic minimization procedure called the multiple shooting approach. Xun et al. [29] used a two-step optimization process called the parameter cascading method. In the first step, a nonparametric B-spline model is fit using the penalized least squares approach for a given parameter vector. The parameter vector is estimated by the least-squares method in the second step. The \( \sqrt{n} \)-consistency and asymptotic normality of the estimator were established in their work. Xun et al. [29] also proposed a Bayesian P-splines approach.

We consider a generalized nonlinear regression model setting, where the regression function is described implicitly by a PDE in terms of the predictor variables involving some unknown model parameters. We pursue a Bayesian approach to inference on the model parameters using a posterior distribution. An advantage of using a Bayesian method is that it automatically gives uncertainty quantification in the form of Bayesian credible sets, which can be also justified in the frequentist sense through the Bernstein-von Mises theorem. One natural approach to the problem is to evaluate the likelihood function using a numerical solution of the PDEs and apply the Bayes theorem to obtain the posterior distribution. In principle, sampling from the posterior distribution can be drawn using MCMC methods, but the approach is very computationally expensive, and occasionally suffers from the problem of multimodality that may lead to local trapping of posterior sampling.

In this paper, we follow the projection-posterior approach by extending the ideas used in the Bayesian two-step method of Bhaumik and Ghosal [2, 3, 4] to nonlinear regression models with respect to a multidimensional predictor variable driven by PDEs. As in those papers, in the first step, a prior is put on the space of all sufficiently smooth functions ignoring the PDE through a standard nonparametric Bayesian technique such as a finite random series of tensor products of B-splines with normal coefficients. A typical sample from the resulting posterior distribution of the regression function is then projected on the space of solutions (or approximate numerical solutions) of the PDE. In general, the projection map is obtained by minimizing a discrepancy measure between a sufficiently smooth function and a solution (or approximate numerical solution) of the PDE. The discrepancy measure may be given by a discrepancy measure between the two regression functions, integrated with respect to a measure on the predictor space. However, as the functional form of the solution of the PDE is unknown, a direct distance between
those functions can be obtained only if an approximate numerical solution replaces the actual solution, provided a numerical solution scheme is available. This approach is a generalization of that of Bhaumik and Ghosal [4] from ODE to PDE models. In particular, the approach needs the evaluation of the numerical solution for various values of the parameter, and hence is computationally expensive. An alternative is to avoid computing a numerical solution by minimizing a discrepancy measure that measures how much the function fails to satisfy the PDE overall, for instance, the $L_2$-norm of $F(\cdot, (D^r f : |r| \leq \alpha), \theta)$ with respect to a measure $\mu$. The corresponding minimizer is the projection map inducing the projection-posterior. The approach is analogous to that in Bhaumik and Ghosal [2, 3] for ODE models. In general, the minimizer is obtained numerically, but when the equation is linear in the parameter, the minimizer may be found explicitly as a solution to a least-squares problem. In particular, this happens for a second-order PDE of three types — parabolic, elliptical and hyperbolic, in their canonical form.

It is natural to choose a measure $\mu$ with a density, and the corresponding procedure will be referred to as the Basic Two-Step Bayesian Method. Nevertheless, the associated integrals will be usually computed using numerical integration, by Riemann sums, for instance. This effectively reduces the underlying measure to a discrete one supported on a fine grid. Then it is also useful to replace the derivatives by the corresponding normalized finite differences. Depending on the type of the underlying measure $\mu$, we end up with two different versions of the procedure, with somewhat different limiting distributions. The latter is computationally more efficient, as only certain sums need to be calculated, and will be referred to as the Finite-Difference Two-Step Bayesian Method. Neither of these gives efficient Bayes estimators though, which can be only obtained by the more computationally intensive method of minimizing a particular discrepancy measure obtained from the Kullback-Leibler divergence between the unrestricted function and the numerical solution of the PDE (assuming such a one is available) integrated with respect to the density of the observation times. We call this the Efficient Two-Step Bayesian Method.

We present Bernstein-von Mises theorems that ensure coverage of credible regions for all three approaches in separate statements, where the result for the Finite-Difference Two-Step Bayesian Method is given only for a second-order parabolic PDE.

The present paper makes several new contributions to the literature. It provides three Bayesian two-step methods for making inference in PDE models, and studies their properties theoretically. A Bernstein-von Mises theorem for PDE models is a completely new contribution to the literature. Moreover, as we consider a generalized regression setting that need not have additive errors, our results are new even when specialized to ODE models. A fallout of that extension is that at the first stage, the coefficients of the spline series do not have an exact normal posterior, and hence asymptotic posterior normality needs to be established. At the second step, for efficiency, the divergence measure has to be specified through the Kullback-Leibler divergence instead of the plain $L_2$-distance between the function and the numerical solution of the PDE.

Throughout the paper, let $\| \cdot \|_\infty$ stand for the supremum-norm for functions on $[0, 1]^s$, while $\| \cdot \|$ will stand for the Euclidean norm for a vector and the operator norm (maximum absolute eigenvalue) of a square matrix. For a function $f$, $\| f \|_{L_2, q_0}$ stands for the $L_2$-norm of $f$ under a density $q_0$ and $\| \cdot \|_{TV}$ stands for the total variation norm of a measure. The notation $N(\mu, \Sigma)$ will stand for the normal distribution with mean vector $\mu$ and dispersion matrix $\Sigma$. For a function of $t$, $\theta$ and other arguments, a dot on the top refers to differentiation with respect to $\theta$ (if $\theta$ is multidimensional, then this stands for the column vector of the partial derivatives, and two dots
Bayesian regression for PDE models

on the top refers to the Hessian matrix). In an integration over \( t \), we shall skip the dummy variable \( t \) and the domain \([0, 1]^s\). The space of continuous function and the space of \( m \)-times continuously differentiable functions on \([0, 1]^s\) are denoted respectively by \( C \) and \( C^m \). We shall write \( P \rightarrow \) for convergence in \( P \)-probability and \( \rightarrow \) for distributional convergence. We shall use \( O, o, O_P, o_P \) symbols for order at most, order smaller than and their stochastic counterparts respectively, and \( a_n \ll b_n \) and \( a_n \ll b_n \) respectively for \( a_n = O(b_n) \) and \( a_n = o(b_n) \). The symbol \( E \) will stand for the expectation operator and \( D \) for the dispersion matrix operator of a random vector under the true sampling distribution, unless explicitly mentioned otherwise.

The paper is organized as follows. In the next section, we describe the methodology and present the corresponding Bernstein-von Mises theorems. In Section 3, proofs of the main results are given. Certain auxiliary lemmas needed to prove the main results are stated and proved in Section 4. A simulation study on the proposed methods is presented in Section 5. The paper concludes with a discussion of possible extensions in Section 6.

2. Method Descriptions and Main Results

Consider samples \((t_1, Y_1), \ldots, (t_n, Y_n)\), where \( t \) is itself randomly obtained by sampling from a fixed positive and continuous density \( q_0 \), and for a given value of \( t = (t_1, \ldots, t_n) \), the conditional density of \( Y \) given \( t \) with respect to a \( \sigma \)-finite measure \( \nu \) is given by an exponential family

\[
p_\theta(y|t) = p(y|t; f_\theta) = \exp\{y f_\theta(t) - \psi(f_\theta(t))\},
\]

where \( f_\theta(\cdot) \) satisfies the PDE (1.1).

Remark 2.1. If \( \eta = g(\xi) \) is another smooth reparameterization of the natural parameter \( \xi \) of the exponential family, then the corresponding regression function \( h_\theta = g(f_\theta) \) satisfies \( \frac{\partial}{\partial \xi} h_\theta(t) = g'(f_\theta(t)) \frac{\partial}{\partial \xi} f_\theta(t) \) by the chain-rule of differentiation, and similarly for higher order derivatives. This means that the equation \( F(t, (D^r f_\theta : |r| \leq \alpha), \theta) = 0 \) can be written equivalently as \( H(t, (D^r h_\theta : |r| \leq \alpha), \theta) = 0 \), where \( H \) is completely specified in terms of the known functions \( F \) and \( g \). In other words, a PDE system for a parameter \( \eta \) not in the natural form can be equivalently described by a system of the form (1.1) for the natural parameter, and vice versa. Thus it suffices to assume that the PDEs describe the natural parameter in terms of the predictor variable \( t \) in our formulation.

We denote the data \((t_1, Y_1), \ldots, (t_n, Y_n)\) at stage \( n \) by \( \mathbb{D}_n \). We write \( P_n \) for the associated empirical distribution and \( \mathbb{G}_n \) the corresponding normalized empirical process, and also let \( Y \) stand for the column vector \((Y_1, \ldots, Y_n)^T\). Let \( \theta_0 \) denote the true value of \( \theta \). For brevity, we write the true regression function \( f_{\theta_0} \) as \( f_0 \), and hence the density of the true data-generating process is \( p(\cdot|t, f_0) \). The distribution corresponding to the density \( p_0 = p(\cdot; f_0) \) is denoted by \( P_0 \). All of the proposed methods use tensor products of B-spline functions to construct prior distributions, for which it will be convenient to have the domain of \( t \) to be \([0, 1]^s\). We shall assume that the domain is rectangular and possibly after some affine transformations, has been reduced to \([0, 1]^s\). For every method, the projection-posterior distribution given the data \( \mathbb{D}_n \) will be generically denoted by \( \Pi(\cdot|\mathbb{D}_n) \). Even though these are different measures, from the context it will be clear which projection-posterior the symbol \( \Pi(\cdot|\mathbb{D}_n) \) refers to.
Being solutions of differential equations, the functions \( t \mapsto f_\theta(t) \) are smooth, and hence continuous, for all \( \theta \). Therefore on the compact domain \([0, 1]^s\), all these functions are bounded. We further assume that the dependence of \( f_\theta(t) \) on \( \theta \) is Lipschitz continuous. This will be typically a consequence of Lipschitz continuous dependence of the binding function \( F \) on \( \theta \), which is immediate from continuous differentiability of \( F \) in the argument \( \theta \). In particular, \( f_\theta(t) \) are uniformly bounded for \( \theta \) in compact subsets of the natural parameter space \( \Theta \).

The natural Bayesian method based on the likelihood computed by a numerical solution of the PDEs leads to a Bernstein-von Mises theorem under some natural smoothness and boundedness conditions on the family of regression functions, provided that a numerical solution with accuracy better than \( n^{-1} \) can be obtained. These conditions ensure that the \( n \)-fold product measures based on the true regression function and its numerical approximations are mutually contiguous, allowing translating convergence statements in terms of the second measure to those in terms of the first. In particular, the \( n^{-1/2} \)-contraction rate of the posterior may be established, which is then used to derive the Bernstein-von Mises theorem using Theorem 2.1 of Kleijn and van der Vaart [15]. In order to avoid a digression, we do not present the statement and the proof of this result.

In this paper, we follow a Bayesian two-step approach to construct projection-posterior distributions and study their convergence properties. We first disregard the PDE and only regard the regression function as an element of a space of smooth functions. We put a prior on it through a representation for any \( \beta \)

\[
\mathbf{X}_{\alpha} = \left( \prod_{j=1}^{s} N_{jk}(t_{ik}) : 1 \leq i \leq n, 1 \leq j \leq j_{k_{n} + m - 1} \right), \quad \{ N_{j}(\cdot) : j = 1, \ldots, k_{n} + m - 1 \}
\]

are B-spline basis functions of fixed order \( m \geq \alpha \) with \( k_{n} - 1 \) equispaced interior knots, and \( \beta \in \mathbb{R}^{(m+k_{n}-1)s} \) is the coefficient vector. The solution of the PDE is assumed to be smooth to at least the \( m \)th order. We put a fixed prior with Lebesgue density \( \pi \) supported on a compact interval of \( \mathbb{R} \), independently on each coordinate of \( \beta \). Let \( \mathbf{X}_{\alpha} = \prod_{j=1}^{s} N_{jk}(t_{ik}) : 1 \leq i \leq n, 1 \leq j \leq j_{k_{n} + m - 1} \), a matrix with \( n \) rows and \( (k_{n} + m - 1)^s \) columns. The corresponding posterior is not restricted to the solutions of the PDE, and hence does not immediately give a posterior distribution on the parameter \( \theta \). Thus we use a projection map \( \chi \) from the space of smooth functions to the space of the solutions of the PDE to induce the corresponding projection-posterior distribution. Depending on the choice of \( \chi \), we end up with different two-step Bayesian methods. When \( \chi \) stands for the Kullback-Leibler divergence between a general smooth \( f \) and a solution of the PDE proxied by a numerical solution, the corresponding Bayes estimator will be shown to be asymptotically efficient. This approach, to be called the Efficient Two-Step Bayesian Method, will be discussed in Subsection 2.3.

The existence and the accuracy of the natural Bayesian method as well as the Efficient Two-Step Bayesian Method are both dependent on the existence of a numerical solution of the PDEs. Unlike ODEs, there is no universal numerical solution for PDEs, and for some types of PDEs, no numerical method may be known yet. Available numerical methods include the finite difference method (Morton and Mayers [19]) and finite element method (Johnson [14]). For linear PDEs, numerical solutions with good convergence properties are available (Tadmor [23]). Even if a numerical solution exists, its repeated computation may be excessively time consuming. An alternative is to construct a projection map using the PDE itself rather than a numerical solution. This can be obtained by minimizing a discrepancy measure \( R_{f}(\theta) \) given by the \( L_2 \)-norm of the binding function \( F \) of the PDE with respect a measure \( \mu \). Then the projection map is given by
\[ \chi(f) = \arg \min \{ R_f(\theta) : \theta \in \Theta \}, \]

where

\[ [R_f(\theta)]^2 := \int \| F(t, (D^r f(t)) : |r| \leq \alpha) \|^2 \mu(dt). \] (2.2)

Since \( f_\theta \) satisfies the PDE with parameter \( \theta \), it is easy to check that \( \chi(f_\theta) = \theta \) for all \( \theta \in \Theta \). Thus the map \( \chi \) extends the definition of the parameter \( \theta \) beyond the model. Clearly, the true parameter \( \theta_0 = \arg \min \{ R_{f_\theta}(\eta) : \eta \in \Theta \} \), and hence the notion extends the definition of \( \theta \) beyond the given PDE model to all sufficiently smooth regression functions. From now on, we shall write \( \theta \) for \( \chi(f) \) and treat it as the parameter of interest. A posterior is induced on \( \Theta \) through the mapping \( \chi \) acting on \( f(\cdot, \beta) = \beta^T N(\cdot) \) and the posterior of \( \beta \).

In many examples, the binding function \( F \) is linear in the parameters, that is,

\[ F(t, (D^r f(t)) : |r| \leq \alpha), \theta \) = \( G(t, (D^r f(t)) : |r| \leq \alpha)) - \theta H(t, (D^r f(t)) : |r| \leq \alpha) \] (2.3)

for some (possibly vector or matrix valued) functions \( G \) and \( H \). Then the minimizer \( \chi(f) \) of (2.2) is obtained explicitly as

\[ \chi(f) = \left( \int H(t, (D^r f(t)) : |r| \leq \alpha)) G(t, (D^r f(t)) : |r| \leq \alpha) \right)^{-1} \] (2.4)

\[ \times \int H(t, (D^r f(t)) : |r| \leq \alpha)) G(t, (D^r f(t)) : |r| \leq \alpha) \mu(dt). \]

2.1. Basic Two-step Bayesian Method

It is natural to choose the measure \( \mu \) in (2.2) absolutely continuous on \([0, 1]^s\) with Radon-Nikodym derivative \( w \), to be referred to as the weight function below. Let the mixed partial derivatives of \( w \) up to the \((s-1)\)th order vanish on the boundary. The weight function \( w \) will be required to have the property that the density decays to 0 at the boundaries along with certain derivatives, so that certain terms arising from integration-by-parts will vanish.

We establish the asymptotic normality of the posterior distribution \( \Pi(\cdot | \eta_0) \) through that of the posterior distribution of \( \beta \) and by showing that \( \theta \) can be represented as an approximately linear function of \( \beta \). The former is obtained through a Bernstein-von Mises theorem in a generalized linear model with dimension increasing to infinity. Then a contiguity result allows reinterpreting the convergence to be under the true data generating process. This scheme is common for all three projection-based methods. The contiguity result is formally stated and proved in Section 4.

We make the following assumption about well-separation of a unique minimum \( \theta_0 \) of \( R_{f_\theta} \).

(A1) For all \( \epsilon > 0 \), \( \inf \{ R_{f_\theta}(\eta) : \| \eta - \theta_0 \| \leq \epsilon \} > 0 \).

We denote \( g = ((D^r f(\cdot, \beta))^T : |r| \leq \alpha) \) and \( g_\alpha = ((D^r f_\theta)^T : |r| \leq \alpha) \). For a function \( \phi(t, g(t), \theta) \), denote \( D_g \phi(t, g(t), \theta) := \frac{\partial \phi(t, g(t), \theta)}{\partial g} \) and \( D_{g_\alpha} \phi(t, g(t), \theta) := \frac{\partial \phi(t, g(t), \theta)}{\partial g} \) respectively; here the derivative with respect to a vector argument refers to a column vector of derivatives with respect to each component of the vector argument.
Let \( G(t, g(t), \theta) = (D_\theta F(t, g(t), \theta))^T F(t, g(t), \theta), M(g_0, \theta_0) = \int D_\theta (G(\cdot, g_0, \theta_0)) \, w, \)

\[
\begin{align*}
\Gamma(f(\cdot, \beta)) &= - \sum_{|r| \leq \alpha} \int (-1)^{\sum_j r_j} D^r \{ D_{g,r} \, (G(\cdot, g_0(\cdot), \theta_0)) \} \, f(\cdot, \beta), \\
A(t) &= -(M(g_0, \theta_0))^{-1} \sum_{|r| \leq \alpha} (-1)^{\sum_j r_j} D^r \{ D_{g,r} \, (G(t, g_0(t), \theta_0)) \} \, w(t),
\end{align*}
\]

where \( D_{g,r} \) stands for the mixed partial derivative operator differentiating with respect to the \((r_1, \ldots, r_s)\)th component of the second argument. Then it follows that

\[
(M(g_0, \theta_0))^{-1} \Gamma(f(\cdot, \beta)) = \int A \beta^T N = H^{\text{TM}}_{n1} \beta, \tag{2.5}
\]

where \( H^{\text{TM}}_{n1} = \int A N^T \) is a \( p \times (k_n + m - 1)^s \)-matrix.

By Corollary 6.21 of Schumaker [22], there exists a \( \beta_0 = ((\beta_{0,j_1, \ldots, j_s})) \in \mathbb{R}^{(k_n + m - 1)^s}, |\beta_{0,j_1, \ldots, j_s}| \leq \|f\|_{\infty} \) for all \( j_1, \ldots, j_s \), such that

\[
\|D^r f_0 - D^r f(\cdot, \beta_0)\|_{\infty} = O(k_n^{-m+|r|}), \tag{2.6}
\]

where \( f(t, \beta_0) = \sum_{j_1=1}^{k_1+m-1} \cdots \sum_{j_s=1}^{k_s+m-1} \beta_{0,j_1, \ldots, j_s} N_{j_1}(t_1) \cdots N_{j_s}(t_s), \ t = (t_1, \ldots, t_s). \)

Let \( \mu_{n1} = \sqrt{n} H^{\text{TM}}_{n1} (X_n^T \Delta X_n)^{-1} X_n^T (Y - E(Y)) \) and \( \Sigma_{n1} = n H^{\text{TM}}_{n1} (X_n^T \Delta X_n)^{-1} H^{\text{TM}}_{n1}, \)

where \( \Delta = \text{diag}(\psi''(f(t_1, \beta_0)), \ldots, \psi''(f(t_n, \beta_0))) \).

We also make the following assumption on the prior and the true regression function.

(A2) The support of the prior \( \pi \) is large enough to include all values of \( f_0(t) \) in its interior.

**Theorem 2.1.** Assume Conditions (A1), (A2) and let \( M(g_0, \theta_0) \) be nonsingular. Let \( f_0 \) have at least all \( m > \max(\alpha + 2s, 2\alpha + s/2) \) mixed order continuous partial derivatives and \( n^{1/2} \lesssim k_n \lesssim n^{1/(2m - \alpha)} \) \(< \delta_1 \leq \delta_2 < \min(1/(4\alpha), 1/(2\alpha + s)) \). Then there exists \( E_n \subseteq (C)^{\# \{ |r| \leq \alpha \} \times \Theta} \) with \( \Pi(E_n^c | \mathbb{D}_n) = o_P(1) \), such that

\[
\sup_{(g, \theta) \in E_n} \left\| \sqrt{n}(\theta - \theta_0) - (M(g_0, \theta_0))^{-1} \sqrt{n}(\Gamma(f(\cdot, \beta)) - \Gamma(f_0)) \right\| \to 0, \tag{2.7}
\]

the projection-posterior of \( \theta \) contracts at the rate \( n^{-1/2} \) in \( P^n_0 \)-probability around \( \theta_0 \) and that

\[
\Pi \left( \sqrt{n}(\theta - \theta_0) \in \cdot | \mathbb{D}_n \right) - N \left( \mu_{n1}, \Sigma_{n1} \right) ||_{\text{TV}} \overset{P^n_0}{\to} 0.
\]

### 2.2. Finite-Difference Two-Step Bayesian Method

While it is natural to choose the measure \( \mu \) to be absolutely continuous with respect to the Lebesgue measure in (2.2), having density \( w \), in practice, the minimizer of (2.2) will have to be computed numerically through a Riemann sum on a fine grid (if (2.3) holds, by a Riemann sum approximation of (2.4)). In this situation, the partial derivatives may also be replaced by the corresponding finite differential coefficients, leading to a variation of the Basic Bayesian Two-Step Method. We present this method only for canonical second-order PDEs with respect to two variables, where the binding function is separable.
Consider a PDE given by \( L_\theta f = 0 \) with \( \theta \in \Theta \), where \( L_\theta \) is a linear differential operator guided by a scalar parameter \( \theta \). Instead of a continuous density, the discrete uniform measure on a lattice is considered with a small mesh-width \( h = h_n \) and the integrations are replaced by summations over the grid in the discrepancy measure. Further, we replace the derivatives by finite difference quotients and consider the projection map

\[
\hat{\beta}(h \eta, \theta) = \arg \min_{\hat{\beta}(h \eta, \theta)} \mathbb{E}_{V_h} f(h \eta, \theta),
\]

where \( L_{\theta,h} \) is the corresponding finite-difference operator and \( \mathcal{G}_h \) is a lattice grid for the domain.

We consider the important case \( s = 2 \) so that \( t = (t_1, t_2) \). A second-order PDE of the parabolic, elliptical or hyperbolic type in their canonical form, the projection map may be explicitly computed by solving a simple least-squares problem. The primary advantage of using this projection map is that no numerical integration is needed, making the approach computationally the fastest. The canonical forms of the elliptical and hyperbolic PDEs are identical (except for the difference in the sign of the parameter) while the parabolic form is different. Thus there are two projection maps, one for the parabolic form and the other for the elliptical (or hyperbolic) form. For each method, we establish a Bernstein-von Mises theorem for the corresponding projection-posterior distribution.

The canonical parabolic PDE is given by

\[
L_{\theta} f(t_1, t_2) = \frac{\partial}{\partial t_2} f(t_1, t_2) - \theta \frac{\partial^2}{\partial t_1^2} f(t_1, t_2) = 0,
\]

and the corresponding finite-difference operator is

\[
L_{\theta,h} f(ih, jh) = f^{(1)}_{i,j} = \frac{f(ih, (j+1)h) - f(ih, jh)}{h},
\]

\[
L_{\theta,h} f(ih, jh) = f^{(2)}_{i,j} = \frac{f((i+1)h, jh) - 2f(ih, jh) + f((i-1)h, jh)}{h^2}.
\]

Therefore, for the parabolic equation, minimization over \( \Theta \) leads to

\[
\theta = \arg \min_{\eta \in \Theta} \mathbb{E}_{V_h} f(\cdot, \beta, \eta, \theta) = \arg \min_{\eta \in \Theta} \sum_i \sum_j \left( \frac{1}{h^2} f^{(1)}_{i,j} - \frac{\eta}{h^2} f^{(2)}_{i,j} \right)^2 = \frac{1}{H_{n^2}} \sum_i \sum_j \left( f^{(1)}_{i,j} - \frac{\eta}{h^2} f^{(2)}_{i,j} \right)^2.
\]

Let \( \beta_0 \) be as defined in (2.6), and

\[
H_{n^2} = 2 \int \int \left( \frac{\partial^2 f_0}{\partial t_1^2} \frac{\partial N}{\partial t_2} + \theta_0 \frac{\partial f_0}{\partial t_2} \frac{\partial^2 N}{\partial t_1^2} \right) \left( \frac{\partial^2 f_0}{\partial t_1^2} \right)^2.
\]

The canonical hyperbolic PDE is

\[
L_{\theta} f(t_1, t_2) = \frac{\partial^2}{\partial t_1^2} f(t_1, t_2) - \theta \frac{\partial^2}{\partial t_1^2} f(t_1, t_2) = 0,
\]

and the corresponding finite-difference operator is

\[
f^{(3)}_{i,j} = \frac{f(ih, (j+1)h, \beta) - 2f(ih, jh, \beta) + f(ih, (j-1)h, \beta)}{h^2}.
\]

Therefore, for the hyperbolic PDE, minimization over \( \Theta \) leads to

\[
\theta = \arg \min_{\eta \in \Theta} \mathbb{E}_{V_h} f(\cdot, \beta, \eta) = \arg \min_{\eta \in \Theta} \sum_i \sum_j \left( \frac{1}{h^2} f^{(3)}_{i,j} - \frac{\eta}{h^2} f^{(2)}_{i,j} \right)^2 = \frac{1}{H_{n^2}} \sum_i \sum_j \left( f^{(3)}_{i,j} - \frac{\eta}{h^2} f^{(2)}_{i,j} \right)^2.
\]
The canonical elliptic PDE is
\[
\frac{\partial^2}{\partial t^2} f(t_1, t_2) + \theta \frac{\partial^2}{\partial t_1^2} f(t_1, t_2) = 0
\]
and hence can be handled in the same way a hyperbolic PDE is treated by flipping the sign of the parameter.

For the finite-difference two-step method, we have two similar posterior asymptotic normality results for the canonical parabolic PDE or hyperbolic (elliptical) PDEs and we only formally state the one for the canonical parabolic PDE.

**Theorem 2.2.** Let \( h \ll n^{-1/2} \), \( n^{\delta_1} \lesssim k_n \lesssim n^{\delta_2} \) where \( 1/(2(m - 2)) < \delta_1 \leq \delta_2 < 1/12 \) for some \( m \geq 9 \), and Condition (A2) hold. Then \( \| \Pi \left( \sqrt{n}(\theta - \theta_0) \right) \|_{TV} \overset{P_n}{\rightarrow} 0 \), where the asymptotic mean is \( \mu_{n2} = H_{n2}^T (\int \psi(f_0)NN^Tq_0) - 1 X_n^T (Y - E(Y)) \) and the asymptotic variance is \( \Sigma_{n2} = H_{n2}^T (\int \psi(f_0)NN^Tq_0) - 1 H_{n2} \).

### 2.3. Efficient Two-Step Bayesian Method

We assume that a numerical solution \( f_{\theta_0} \) to the PDE exists, even though a particular form is not needed. As in the Basic Two-Step Bayesian Method, we first obtain a posterior distribution in the nonparametric model using a finite random series prior based on a tensor product of B-splines basis through the representation \( f(t, \beta) = NN^T(t) \beta \), but we define \( \theta \) by the projection
\[
\theta = \arg \min \{ -E_f \log p_{\eta,n} : \eta \in \Theta \},
\]
where \( E_f \) denotes expectation with respect to \( f(\cdot, \beta), p_{\eta,n}(y|t) = \exp \{ y f_{\eta,n}(t) - \psi(f_{\eta,n}(t)) \} \), and \( f_{\eta,n} \) is a numerical solution of (1.4). Then the posterior of \( \beta \) induces a posterior on \( \theta \). For ordinary linear regression, the objective function in (2.10) reduces to \( \int (f(t, \beta) - f_{\eta,n}(t))^2q_0(t)dt \). Let
\[
V_{\theta_0} = \int \psi''(f_0)f_0f_0^Tq_0, \quad C(t) = V_{\theta_0}^{-1}f_0(t), \quad H_{n3}^T = \int \psi''(f_0)CN^Tq_0 \quad \text{and} \quad \Delta = \text{diag}(\psi''(f(t_1, \beta_0)), \ldots, \psi''(f(t_n, \beta_0))),
\]
where \( f_0 \) stands for \( f_0 \) at \( \theta = \theta_0 \). Let
\[
\mu_{n3} = \sqrt{n}H_{n3}^T(X_n^T\Delta X_n)^{-1}X_n^T(Y - E(Y)), \quad \Sigma_{n3} = nH_{n3}^T(X_n^T\Delta X_n)^{-1}H_{n3}.
\]

**Theorem 2.3.** Suppose that a numerical solution \( f_{\theta_0} \) for \( \theta_0 \) exists such that \( \| f_{\theta_0} - f_0 \|_\infty = o(n^{-1/2}) \) uniformly in \( \theta \in \Theta \). Assume that the Fisher information matrix \( V_{\theta_0} \) for \( \theta \) is nonsingular, and \( n \rightarrow -E(\log p_\eta) \) has a well-separated minimum at \( \theta_0 \). Let Condition (A2) hold, \( f_{\theta_0} \) have at least all \( m \geq \alpha + 2s \) mixed order continuous partial derivatives and \( n^{\delta_1} \lesssim k_n \lesssim n^{\delta_2} \) for some \( 1/(2(m - \alpha)) < \delta_1 \leq \delta_2 < 1/(4s) \). Then \( \| \Pi \left( \sqrt{n}(\theta - \theta_0) \right) \|_{TV} \overset{P_n}{\rightarrow} 0 \), the posterior of \( \theta \) contracts at the rate \( n^{-1/2} \) in \( P_0^{n/2} \)-probability around \( \theta_0 \) and the Bayes estimator for \( \theta \) is asymptotically efficient.

The important point to note here is that in the generalized regression setting, in order to achieve asymptotic efficiency through a projection posterior, the projection map needs to minimize the Kullback-Leibler divergence between \( p(\cdot|\cdot, f(\cdot, \beta)) \) and \( p(\cdot|\cdot, f_{\eta,n}(\cdot)) \), rather than their weighted \( L_2 \)-distance, as in Bhaumik and Ghosal [4] for normal regression driven by ODEs. This is because, in an exponential family, the Fisher information is obtained as the curvature of the Kullback-Leibler divergence map rather than that of the \( L_2 \)-map. The discrepancy disappears for the normal model considered by Bhaumik and Ghosal [4].
3. Proofs of the Main Results

Proof of Theorem 2.1. By the definitions of \( \theta \) and \( \theta_0 \), \( \int G(\cdot, \theta)w = 0 \) and \( \int G(\cdot, \theta_0)w = 0 \). Subtracting the second equation from the first and applying the mean-value Theorem, we obtain that \( \int \left[ D_\theta(G(\cdot, \theta_0))(\theta - \theta_0) + D_{\theta_0}(G(\cdot, \theta_0))(\theta - \theta_0) \right] w = O(||g - g_0||^2_{\infty}) + O(||\theta - \theta_0||^2) \). We shall show that \( \int D_{\theta_0}(G(\cdot, \theta_0))(\theta - \theta_0)w \) is a linear functional of \( f(\cdot, \beta) - f_0 \). 

Now \( \int D_{\theta_0}(G(\cdot, \theta_0))(g - g_0)w = \sum_{|r| \leq \alpha} \int D_{\theta_0}^r(G(\cdot, \theta_0))(D^r f(\cdot, \beta) - D^r f_0)w. \) Rewrite each term as \((-1)^r \sum_{i=1}^{r} \int D_{\theta_0}^r(G(\cdot, \theta_0))w(f(\cdot, \beta) - f_0) \) using integration by parts, which is a linear functional of \( f(\cdot, \beta) - f_0(\cdot) \). This gives the relation

\[
M(g_0, \theta_0)(\theta - \theta_0) = \Gamma(f(\cdot, \beta) - f_0(\cdot)) + O(||g - g_0||^2_{\infty}) + O(||\theta - \theta_0||^2).
\]

From Lemma 4.4, we have \( \Pi(||g - g_0||_{\infty} \leq \delta_n|D_n) \rightarrow 1 \) in \( P^0_0 \)-probability for some \( \delta_n = o(n^{-1/4}) \). Then using the steps of the proof of Lemma 3 of Bhamik and Ghosal [2], we can find \( \epsilon_n \rightarrow 0 \) such that for \( E_n = \{ (g, \theta) : ||g - g_0||_{\infty} \leq \delta_n, ||\theta - \theta_0|| \leq \epsilon_n \} \), we have \( \Pi(E_n|D_n) \rightarrow 1 \) for each \( n \). 

By Lemma 4.5, the first term takes values in a large compact set with posterior probability arbitrarily close to one, in \( P^0_0 \)-probability. Thus on a set of high posterior probability, the posterior distribution of \( \sqrt{n}(\theta - \theta_0) \) is approximated by that of \( \sqrt{n}(M(g_0, \theta_0)^{-1} \Gamma(f(\cdot, \beta) - f_0)) \), which is asymptotically equal to the stated normal distribution with \( P^0_0 \)-probability tending to one, in view of Lemma 4.3. \( \square \)

Proof of Theorem 2.2. We prove the result for only the parabolic case; other cases can be handled similarly. For a given function \( f \) on \([0, 1]^2\), we have

\[
G_h(f, \theta) = V_h(f) = -2h^2 \sum_i \sum_j \left( \frac{f^{(1)}_{ij}}{h} - \theta \frac{f^{(2)}_{ij}}{h^2} \right) f^{(2)}_{ij}.
\]

For a given \( h \) and \( \beta \), let \( \theta \) and \( \theta_{0,h} \) respectively denote the solutions of \( G_h(f(\cdot, \beta), \eta) = 0 \) and \( G_h(f_0, \eta) = 0 \). By (2.8), \( \theta_{0,h} = h \sum_i \sum_j f^{(1)}_{ii,j} f^{(2)}_{ii,j} / \sum_i \sum_j (f^{(2)}_{ii,j})^2 \), where

\[
f^{(1)}_{0i,j} = f_0(ih, (j+1)h) - f_0(ih, jh), f^{(2)}_{0i,j} = f_0((i+1)h, jh) - 2f_0(ih, jh) + f_0((i-1)h, jh).
\]

By the convergence of Riemann sums to integrals, it is immediate that \( \theta_{0,h} = \theta_0 + O(h) \rightarrow \theta_0 \) as \( h \rightarrow 0 \). By Lemma 4.3, on a set of \( P^0_0 \)-probability tending to one, with arbitrarily high posterior probability, \( ||\beta - \beta_0||_{\infty} \leq ||\beta - \beta_0|| \leq k_n/\sqrt{n} \). A Taylor series expansion of \( G_h(f(\cdot, \beta), \theta) \) in \( (\beta, \theta) \) at \((\beta_0, \theta_{0,h})\) gives

\[
G_h(f(\cdot, \beta_0), \theta_{0,h}) + \frac{\partial G_h(f(\cdot, \beta_0), \theta_{0,h})}{\partial \theta}(\theta - \theta_{0,h}) + \frac{\partial G_h(f(\cdot, \beta_0), \theta_{0,h})}{\partial \beta}(\beta - \beta_0).
\]
up to $O(||\beta - \beta_0||^2) + O(||\theta - \theta_{0,h}||^2)$. As $G_h(f(\cdot, \beta), \theta) = 0 = G_h(f_0(\cdot), \theta_{0,h})$, this leads to the linearization
\[
\sqrt{n}(\theta - \theta_{0,h}) = -\sqrt{n} \left( \frac{\partial G_h(f(\cdot, \beta_0), \theta_{0,h})}{\partial \theta} \right)^{-1} \left( (G_h(f(\cdot, \beta_0), \theta_{0,h}) - G_h(f_0(\cdot), \theta_{0,h})) \right) \]
\[+ \frac{1}{\sqrt{n}} \frac{\partial G_h(f(\cdot, \beta_0), \theta_{0,h})}{\partial \beta} (\beta - \beta_0) \] (3.1)

We shall show that on a set of probability tending to one, the first, third and fourth terms are arbitrarily small with high posterior probability, and the second term has a normal distributional limit in probability.

The middle factor of the first term in (3.1) is
\[
\frac{\partial G_h(f(\cdot, \beta_0), \theta_{0,h})}{\partial \theta} = 2h^2 \sum_i \sum_j \left( f^{(2)}_{ij} / h^2 \right)^2 \to \int \int \left( \frac{\partial^2 f_0(t_1, t_2)}{\partial t_1^2} \right)^2 dt_1 dt_2
\]
by the convergence of the Riemann sum and the convergence of $f(\cdot, \beta_0)$ to $f_0$. Thus the first term is of the order
\[
\sqrt{n}(\theta - \theta_{0,h}) = -\frac{1}{\sqrt{n}} \frac{\partial G_h(f(\cdot, \beta_0), \theta_{0,h})}{\partial \theta} \left. \right|_{\beta = \beta_0} \leq \sqrt{n} ||f_0 - f(\cdot, \beta_0)||_\infty \leq \sqrt{n} k_n^{-m} \to 0
\]
as $n \to \infty$ by the choice of $k_n$ and since $f_0$ is smooth to the $m$th (actually infinite) order.

From Lemma 4.3 and the choice of $k_n$, it follows that $\sqrt{n}||\beta - \beta_0||$ in (3.1) is $o(1)$ with probability tending to 1.

We next bound the fourth term in (3.1). First observe that, the dimension of $\beta$ is of the order $k_n^2$. Let $E = \{ \beta : ||\beta - \beta_0|| \leq C k_n / \sqrt{n} \}$, which has high posterior probability in $P^n_0$-probability in view of Lemma 4.3. Then
\[
||f(\cdot, \beta) - f_0||_\infty \lesssim ||\beta - \beta_0||_\infty + ||f(\cdot, \beta_0) - f_0||_\infty \leq \sqrt{k_n^2/n + k_n^{-m}} \ll n^{-1/2}
\] (3.2)
for our choice of $k_n$. Using properties of the tensor product of B-splines basis, $|f^{(1)}_{ij} - f^{(1)}_{0ij}| = O(hk_n ||\beta - \beta_0||) + O(hk_n^{-5}) = O(hk_n^2 / \sqrt{n} + h k_n^{-5})$, $|f^{(2)}_{ij} - f^{(2)}_{0ij}| = O(h^2 k_n^4 ||\beta - \beta_0||) + O(h^2 k_n^{-4}) = O(h^2 k_n^2 / \sqrt{n} + h k_n^{-4})$, whenever $\beta \in E$. Now by the convergence of Riemann sums
\[
\begin{align*}
&\cdot \ h^2 \sum_i \sum_j f^{(1)}_{0ij} / h \to \int \frac{\partial f_0}{\partial t_2} dt_1 dt_2, \\
&\cdot \ h^2 \sum_i \sum_j f^{(2)}_{0ij} / h^2 \to \int \int \frac{\partial^2 f_0}{\partial t_1^2} dt_1 dt_2, \\
&\cdot \ h^2 \sum_i \sum_j \left( f^{(1)}_{0ij} / h \right) \left( f^{(2)}_{0ij} / h^2 \right) \to \int \int \frac{\partial f_0}{\partial t_1} \frac{\partial f_0}{\partial t_2} dt_1 dt_2 \\
&\cdot \ h^2 \sum_i \sum_j \left( f^{(2)}_{0ij} / h^3 \right) \to \int \left( \frac{\partial^2 f_0}{\partial t_1^2} \right)^2 dt_1 dt_2
\end{align*}
\]
Thus it follows that for $\beta \in E$,
\[
||\theta - \theta_{0,h}|| = \left| \frac{h^2 \sum_i \sum_j \left( f^{(1)}_{0ij} / h \right) \left( f^{(2)}_{0ij} / h^2 \right)^2}{h^2 \sum_i \sum_j \left( f^{(2)}_{0ij} / h^2 \right)^2} - \frac{h^2 \sum_i \sum_j \left( f^{(1)}_{ij} / h \right) \left( f^{(2)}_{ij} / h^2 \right)^2}{h^2 \sum_i \sum_j \left( f^{(2)}_{ij} / h^2 \right)^2} \right| = O\left( \frac{k_n^3}{\sqrt{n}} + k_n^{-m+2} \right).
\]
Note that \( \sqrt{n} h^3 / \sqrt{n} \leq n^6 \delta_2^{-1/2} \rightarrow 0 \) and \( \sqrt{n} h^{2(m-2)} \leq n^{-2} \delta_1 (m-2) + 1/2 \rightarrow 0 \) as \( 1/(2(m-2)) < \delta_1 \leq \delta_2 < 1/12 \). Hence \( \sqrt{n} \| \theta - \theta_{0,h} \| \) in (3.1) is small with high posterior probability in true probability.

Therefore, (3.1) reduces to the study of the second term. We now show that its posterior distribution is asymptotically normal with mean \( \mu_{n2} \) and variance \( \Sigma_{n2} \). Let \( \mathcal{N}_{i,j} = \mathcal{N}(i h_j, j h_j) \), where \( \mathcal{N} \) is the vector of tensor products B-spline functions and let \( \Delta_1 \) and \( \Delta_2 \) respectively stand for the difference operators in the first coordinate and the second coordinate respectively.

By straightforward calculations, all components of the vector

\[
\frac{\partial G(f(\cdot, \beta_0), \theta_{0,h})}{\partial \beta} = -2h^2 \left[ \sum_i \sum_j \frac{\Delta_2 N_{i,j}}{i} \frac{f_{ij}^{(2)}(\cdot, \beta_0)}{h^2} + \theta_{0,h} \sum_i \sum_j \frac{f_{ij}^{(1)}(\cdot, \beta_0)}{h^2} \right]
\]

are uniformly close to those of the corresponding integrals \( -2 \int \int (\frac{\partial^2 f_0}{\partial t^2} + \theta_0 \frac{\partial f_0}{\partial t} \frac{\partial^2 N}{\partial t^2}) \), by the convergence of Riemann sums, uniform convergence of \( f(\cdot, \beta_0) \) to \( f_0 \) as \( h \rightarrow 0 \) as \( h \rightarrow 0 \).

By Lemma 4.3, with probability tending to 1, the posterior of \( \sqrt{n} (\beta - \beta_0) \) is approximately \( \mathcal{N}(\sqrt{n} (X_n^T \Delta X_n)^{-1} X_n^T (Y - EY), n(X_n^T \Delta X_n)^{-1}) \). By the law of large numbers and uniform spline approximation, \( (X_n^T \Delta X_n)/n \) is uniformly close to \( \int \psi(f_0) NN^T q_0 \). Hence the posterior distribution of the second term is approximately normal with mean

\[
\frac{2}{\sqrt{n}} \int (\frac{\partial^2 f_0}{\partial t^2})^2 \int (\frac{\partial^2 f_0}{\partial t^2} \frac{\partial N}{\partial t_2} + \theta_0 \frac{\partial f_0}{\partial t_2} \frac{\partial^2 N}{\partial t^2}) \left( \int \psi(f_0) NN^T q_0 \right)^{-1} X_n^T (Y - EY)
\]

and variance

\[
\frac{4 \int (\frac{\partial^2 f_0}{\partial t^2} \frac{\partial N}{\partial t_2} + \theta_0 \frac{\partial f_0}{\partial t_2} \frac{\partial^2 N}{\partial t^2})^T \left( \int \psi(f_0) NN^T q_0 \right)^{-1} \int (\frac{\partial^2 f_0}{\partial t^2} \frac{\partial N}{\partial t_2} + \theta_0 \frac{\partial f_0}{\partial t_2} \frac{\partial^2 N}{\partial t^2})}{\left( \int (\frac{\partial^2 f_0}{\partial t^2})^2 \right)^2}.
\]

Finally, recall that \( \theta_{0,h} - \theta_0 = O(h) = o(n^{-1/2}) \) by the choice \( n \ll n^{-1/2} \). This shows that \( \sqrt{n} (\theta - \theta_{0,h}) \) and hence \( \sqrt{n} (\theta - \theta_0) \) also, have posterior distribution approximately \( \mathcal{N}(\mu_{n2}, \Sigma_{n2}) \) with high probability.

**Proof of Theorem 2.3.** By the definition of the projection, and a Taylor series expansion, it is easy to see that \( \sqrt{n} (\theta - \theta_0) \) admits the linearization \( \sqrt{n} H_{n3}^T (\beta - \beta_0) \). The arguments use the closeness of the numerical solution to the actual one for the PDE in the same way Bhaumik and Ghosal [4] did in the ODE context. The details of the arguments are omitted. This leads to the asymptotic mean \( \mu_{n3} \) and asymptotic covariance matrix \( \Sigma_{n3} \) in view of Lemma 4.3.

As for the efficiency, the log-density is \( Y f_0(t) - \psi(f_0(t)) + \log q_0(t) \) and the Fisher information is given by \( \int \psi''(f_0) f_0 f_0^T q_0 = V_{\theta_0} \). By Lemma 4.6, \( \Sigma_{n3} \rightarrow V_{\theta_0}^{-1} \) and \( \mu_{n3} \rightarrow N(0, V_{\theta_0}^{-1}) \).

Thus the asymptotic dispersion matrix can be replaced by \( V_{\theta_0}^{-1} \) and the asymptotic mean is asymptotically equivalent with an efficient estimator, establishing that the Bayes estimator is asymptotically efficient. 

\( \square \)
4. Auxiliary Lemmas

Let \( Q_n^0 \) be the probability measure with density function (with respect to \( \nu^n \))

\[
\frac{dQ_n^0(y)}{d\nu^n} = h(y, \beta_0) = \exp\{\sum_{i=1}^n y_i f(t_i, \beta_0) - \psi(f(t_i, \beta_0))\}. \tag{4.1}
\]

The measure \( Q_n^0 \) stands for the joint distribution of the observations in a generalized linear model setting with predictors given by tensor product B-spline functions of the original predictor \( t \) under the parameter value \( \beta_0 \). Even though this distribution is not the true data generating process, it closely resembles that, in that all asymptotic properties derived under one distribution translated to the same properties under the other. This follows from the contiguity result below, which plays a key role in deriving asymptotic properties of all three two-step methods.

**Lemma 4.1.** Let \( f_0 \in C^m \) for some \( m \geq 1 \) and \( k_n \gg n^{1/(2m)} \). Then the measures \( P_n^0 \) and \( Q_n^0 \) are mutually contiguous, i.e., for any sequence of events \( A_n \), \( Q_n^0(A_n) \to 0 \) if and only if \( P_n^0(A_n) \to 0 \).

**Proof.** Note that the log-likelihood ratio of \( Q_n^0 \) to \( P_n^0 \) is given by

\[
\sum_{i=1}^n (Y_i - \psi'(f_0(t_i))) (f(t_i, \beta_0) - f_0(t_i)) - \sum_{i=1}^n (\psi(f(t_i, \beta_0)) - \psi(f_0(t_i)))
\]

\[
+ \sum_{i=1}^n \psi'(f_0(t_i)) (f(t_i, \beta_0) - f_0(t_i))
\]

\[
= \sum_{i=1}^n (Y_i - \psi'(f_0(t_i))) (f(t_i, \beta_0) - f_0(t_i)) + O(n\|f(\cdot, \beta_0) - f_0\|_{2, \infty}).
\]

Under \( P_n^0 \)-probability, the first term has a mean 0. Its variance is bounded by a constant multiple of \( n\|f(\cdot, \beta_0) - f_0\|_{2, \infty} = o(nk_n^{-2m}) = o(1) \), and the second term converges to 0 as well. Hence the likelihood ratio converges in probability to 1, thus remains bounded away from zero and infinity. Lemma 6.4 of van der Vaart [24] now implies mutual contiguity. \( \square \)

The next result gives the order of the eigenvalues of the scale matrix appearing in posterior asymptotic normality the coefficients of B-spline functions.

**Lemma 4.2.** All eigenvalues of \( (X_n^T \Delta X_n)^{-1} \) are of the order \( k_n^{2m}/n \) in \( P_n^0 \)-probability.

**Proof.** One can proceed like the proof of Lemma A.9 of Yoo and Ghosal [30], where it was shown that all eigenvalues of \( X_n^T X_n \) are of the order \( n/k_n^{2m} \) in \( P_n^0 \)-probability. Since \( \Delta \) is a diagonal matrix with entries \( \psi''(f(t_1, \beta_0)), \ldots, \psi''(f(t_n, \beta_0)) \), all of which remain confined between two fixed positive numbers, a minor modification of their argument incorporating these weights in forming linear combinations will establish the estimate. \( \square \)
The following result is the main ingredient for all two-step methods, establishing that in the tensor product B-splines-based nonparametric model, the posterior distribution of the coefficient vector $\beta$ is approximately normal on a set of samples with high probability.

**Lemma 4.3.** Let B-spline functions in $N$ be of order $m > 2s$ with $k_n = 1$ equispaced interior knots, where $n^{3s} \leq k_n \leq n^6$ for $1/(2m) < \delta_1 < \delta_2 < 1/(4s)$. Consider the model $Y_i | (t_i, \eta_i) \sim f(\cdot, \eta_i)$ independently, where $f(y, \eta) = \exp[y\eta - \psi(\eta)]$ (density with respect to a $\sigma$-finite measure $\nu$), $\eta_i = N(t_i)\beta$, and each component of $\beta$ is independently given a prior $\pi$ satisfying Condition (A2). Let the data generating process follow (2.1) with true value $\theta_0$. Then with $P_0^\alpha$-probability tending one,

$$
||\Pi(\{X_n^T \Delta X_n\}^{1/2} (\beta - \beta_0) \in \cdot | D_n) - \mathcal{N}(\{X_n^T \Delta X_n\}^{1/2} X_n^T (Y - EY), I_{(k_n+m-1)^s})||_{TV} \to 0,
$$

and $||\beta - \beta_0|| \leq k_n^{s/\sqrt{n}} \leq n^{6s-1/2} \leq n^{-1/4}$ with high posterior probability, where $X_n$ and $\Delta$ are as defined in Section 2.1.

**Proof.** Let $\beta_0 = \langle (\beta_{0,i}, \ldots, \beta_{0,n}) \rangle \in R^{(k_n+m-1)^s}$ be such that $||f_0 - f(\cdot, \beta_0)||_{\infty} = O(k_n^{-m})$ and $||\beta_0||_{\infty} \leq ||f_0||_{\infty}$, in view of Corollary 6.21 of Schumaker [22]. We shall study the behavior of the posterior conditioned on the values of $t_1, \ldots, t_n$. First, we study convergence under the measure $Q_0^\alpha$ defined in Lemma 4.1, that is, pretending that the true value of $\eta_i$ is given by $f(t_i, \beta_0) := N(t_i)\beta_0$. We note that the dimension of $\beta$ is $p = (k_n + m - 1)^s$, which satisfies $p^2 \log p/4 \to 0$ because $k_n^{s/\sqrt{n}} \leq n^{6s-1/2}$ and $4\delta_2s < 1$. On a set of values for $(t_1, \ldots, t_n)$ with probability tending to 1, $||X_n^T \Delta X_n|| = O(k_n^{s/\sqrt{n}})$ by Lemma 4.2. In view of Condition (A2), the prior has compact support and contains the true parameter in its interior. Thus all conditions of Theorem 2.1 of Ghosal [10] hold, and hence the posterior of $(X_n^T \Delta X_n)^{1/2}(\beta - \beta_0)$ is approximately $\mathcal{N}(\{X_n^T \Delta X_n\}^{1/2} X_n^T (Y - EQ_0^\alpha Y), I_{(k_n+m-1)^s})$ in terms of the total variation distance under the measure $Q_0^\alpha$. The conclusion translates to the same assertion under the true measure $P_0^\alpha$ in view of the contiguity result Lemma 4.1, since

$$
||f_0 - f(\cdot, \beta_0)||_{\infty} = O(k_n^{-m}) = O(n^{-m/(2(m-\alpha))}) = o(n^{-1/2}).
$$

In the expression, we can replace $E_{Q_0^\alpha}(Y)$ by $E_0(Y)$, since with probability tending to 1,

$$
||\{X_n^T \Delta X_n\}^{1/2} X_n^T (E_0(Y) - E_{Q_0^\alpha}(Y))|| \leq k_n^{s/\sqrt{n}} ||X_n||_{\infty}||E_0(Y) - E_{Q_0^\alpha}(Y)||,
$$

the entries of $X_n$ are uniformly bounded by 1 and $Y$ is $n$-dimensional, implying that the above bound reduces to $k_n^{s/\sqrt{n}} ||f_0(t) - f(t, \beta_0)||_{\infty} \leq k_n^{-m-s} \to 0$ by the assumption $m > s$ and $k_n \to \infty$. This proves the first part of the assertion.

Further, we note that $D(||\{X_n^T \Delta X_n\}^{1/2} X_n^T (Y - EY)||) = I_p$, so that

$$
||\{X_n^T \Delta X_n\}^{1/2} X_n^T (Y - EY)||^2 = O_{P_0^\alpha}(\text{tr}(I_{(k_n+m-1)^s})) = O_{P_0^\alpha}(k_n^s).
$$

Also, in the approximate posterior, the dispersion matrix of $\beta$ is $(X_n^T \Delta X_n)^{-1}$, which has eigenvalues of the order $k_n^s/n$ by Lemma 4.2. This implies that the approximately normal posterior for $\beta$ concentrates in $O(k_n^s/\sqrt{n}) = O(n^{6s-1/2}) = o(n^{-1/4})$ neighborhoods in $P_0^\alpha$-probability. Because of the convergence in total variation, the actual posterior for $\beta$ also concentrates in $o(n^{-1/4})$-neighborhoods in $P_0^\alpha$-probability. \qed
The next result is to indicate the posterior consistency of the derivatives of the nonparametric estimator, which is important to prove Theorem 2.1.

**Lemma 4.4.** Under the setup and conditions of Theorem 2.1, for any $r$ with $|r| \leq \alpha$, we have $\Pi(\sqrt{n}\|D^r f(\cdot, \beta) - D^r f_0\|_\infty > \epsilon |D_n|) \overset{P_n}{\to} 0$ for any $\epsilon > 0$.

**Proof.** Since $\sqrt{n}\|D^r f(\cdot, \beta_0) - D^r f_0\|_\infty \leq \sqrt{n}k_n^{2m+2\alpha} \lesssim \sqrt{n}k_n^{-m-2s} \to 0$, it suffices to show that for any $\epsilon > 0$, $\Pi(\sqrt{n}\|D^r f(\cdot, \beta) - D^r f(\cdot, \beta_0)\|_\infty > \epsilon |D_n|) \overset{P_n}{\to} 0$. By the property of tensor product B-splines, $\|D^r f(\cdot, \beta) - D^r f(\cdot, \beta_0)\|_\infty \leq k_n^{r_1^\alpha}k_n^{r_2^\alpha}/\sqrt{n} = o((n^{-1/2} + \delta_2(2\alpha + s))$. Hence by Lemma 4.3, and because $\sqrt{n}(n^{-1/2} + \delta_2(2\alpha + s))^2 = o(1)$, the desired conclusion follows. \qed

Consider the setting of Theorem 2.1 and then the next lemma is true and crucial.

**Lemma 4.5.** Eigenvalues of $H_{n1}^T(X_n^T \Delta X_n)^{-1} X_n^T \Delta_0 X_n (X_n^T \Delta X_n)^{-1} H_{n1}$, the conditional dispersion matrix of $H_{n1}^T(X_n^T \Delta X_n)^{-1} X_n^T (Y - EY)$ given $t_1, \ldots, t_r$, are bounded by a multiple of $n^{-1}$ in $P_n^0$-probability, where $\Delta_0 = \text{diag}(\psi''(f_0(t_1)), \ldots, \psi''(f_0(t_r)))$.

**Proof.** Since by Lemma 4.2, both $X_n^T \Delta X_n$ and $X_n^T \Delta_0 X_n$ have all eigenvalues of the order $n/k_n^{\alpha}$, it suffices to show that the eigenvalues of $H_{n1}^T(X_n^T \Delta X_n)^{-1} H_{n1}$ are bounded by a multiple of $n^{-1}$.

We observe that all components of $A$, namely $A_k$, $k = 1, \ldots, p$, are $(m - \alpha)$-smooth and hence are uniformly bounded. Hence by the approximation property of B-splines basis (see Theorem 12.7 of Schumaker [22]), we have $\|A_k - \tilde{A}_k\| = O(k_n^{(m-\alpha)}) = O(n^{-\delta_1(m-\alpha)}) = o(n^{-1/2})$ as $\delta_1 > 1/(2(m-\alpha))$ by our choice, where $\tilde{A}_k = \alpha_k^T N$, for some $\alpha_k$, with entries bounded. Write

$$
H_{n1}^T(X_n^T \Delta X_n)^{-1} H_{n1} = (H_{n1} - \tilde{H}_{n1})^T(X_n^T \Delta X_n)^{-1}(H_{n1} - \tilde{H}_{n1})
+ \tilde{H}_{n1}^T(X_n^T \Delta X_n)^{-1} \tilde{H}_{n1} + (H_{n1} - \tilde{H}_{n1})^T(X_n^T \Delta X_n)^{-1} \tilde{H}_{n1}
+ \tilde{H}_{n1}^T(X_n^T \Delta X_n)^{-1}(H_{n1} - \tilde{H}_{n1}),
$$

where $[\tilde{H}_{n1}]_{k_1} = \int \tilde{A}_k N^T$ for $k = 1, \ldots, p$.

First, we handle the second term as

$$
\|\tilde{H}_{n1}^T(X_n^T \Delta X_n)^{-1} \tilde{H}_{n1}\| \lesssim \frac{k_n^s}{n} \|\tilde{H}_{n1}^T \tilde{H}_{n1}\| = \frac{k_n^s}{n} \|\tilde{A}N^T \int N \Lambda^T\|.
$$

This can be written as $\|\Lambda(\int NN^T)\Lambda^T\|$, where $\Lambda = (\alpha_1, \ldots, \alpha_p)$. It is contained in the proof of Lemma A.9 of Yoo and Ghosal [30] is that the eigenvalues of $\int NN^T$ are of order $k_n^s$. Therefore, the expression is bounded by a constant multiple of $(nk_n^s)^{-1} \text{tr}(\Lambda \Lambda^T)$. Since
the entries of $\Lambda$ are bounded, $\Lambda$ is a $p \times (k_n + m - 1)^s$-matrix with fixed $p$, it follows that the trace is $O(k_n^s)$. Thus $\|H_{n1}^T(X_n^T \Delta X_n)^{-1}H_{n1}\| \lesssim n^{-1}$.

The $k$th diagonal entry $(H_{n1} - \tilde{H}_{n1})^T(X_n^T \Delta X_n)^{-1}(H_{n1} - \tilde{H}_{n1})$ of the first term is

$$\int (A_k - \tilde{A}_k)N^T(X_n^T \Delta X_n)^{-1}N \lesssim \frac{k_n}{n} \|A_k - \tilde{A}_k\| \leq \frac{1}{n} \int N^T N.$$

Because $\|A_k - \tilde{A}_k\| \lesssim k_n^{-(m - s)}$, $\|(X_n^T \Delta X_n)^{-1}\| \lesssim k_n^{s}/n$, $\|N\|^2 \leq 1$ and $2m > 2\alpha + s$, the above expression bounded by a multiple of $n^k_{n}^{-(2m - 2\alpha - s)} = o(n^{-1})$. Since $p$ is fixed dimensional, this shows that $\|(H_{n1} - \tilde{H}_{n1})^T(X_n^T \Delta X_n)^{-1}(H_{n1} - \tilde{H}_{n1})\| = o(n^{-1}).$

To bound $\|(H_{n1} - \tilde{H}_{n1})^T(X_n^T \Delta X_n)^{-1}H_{n1}\|$, observe that its Frobenius norm, which dominates its operator norm, is bounded using the Cauchy-Schwarz inequality by

$$\sqrt{(H_{n1} - \tilde{H}_{n1})^T(X_n^T \Delta X_n)^{-1}(H_{n1} - \tilde{H}_{n1})} \lesssim n^{-1}.$$

The same argument applies to the fourth term. Hence $\|H_{n1}^T(X_n^T \Delta X_n)^{-1}H_{n1}\| \lesssim n^{-1}$.  

The next lemma describes the asymptotic behavior of the mean and variance of the limiting normal distribution given by Theorem 2.3.

**Lemma 4.6.** *In Theorem 2.3, $\mu_{n3} \Rightarrow N(0, V_{\theta_0}^{-1})$ and $\Sigma_{n3} \rightarrow V_{\theta_0}^{-1}$.*

**Proof.** First we study the asymptotic dispersion matrix $\Sigma_{n3} = H_{n3}^T(X_n^T \Delta X_n)^{-1}H_{n3}$. Let $\hat{C}_k$ be a tensor product B-spline approximation of $C_k$ up to accuracy $k_n^{-(m + \alpha)}$, and let $\tilde{H}_{n3}$ stands for the expression for $H_{n3}$ with $C_k$ replaced by $\hat{C}_k$. Then following the arguments in the proof of Lemma 4.5, it can be established that

$$nH_{n3}^T(X_n^T \Delta X_n)^{-1}H_{n3} = n\tilde{H}_{n3}^T(X_n^T \Delta X_n)^{-1}\tilde{H}_{n3} + o(1).$$

Hence it suffices to show that $n\tilde{H}_{n3}^T(X_n^T \Delta X_n)^{-1}\tilde{H}_{n3} \rightarrow V_{\theta_0}^{-1}$.

By the law of large numbers and uniform spline approximation, $X_n^T \Delta X_n)$ is uniformly close to $\psi(f_0)N N^T q_0$. Thus the asymptotic dispersion matrix is approximately equal to

$$\int \psi''(f_0) N N^T q_0 \int \psi''(f_0) N N^T q_0^{-1} \int \psi''(f_0) N C^T q_0.$$

We shall show that this converges to $V_{\theta_0}^{-1}$.

Recall that $C$ can be uniformly approximated by $\hat{C} = \Lambda \tilde{N}$, for some matrix $\Lambda$ comprising of the coefficients of the tensor product B-spline approximation for $C$. Hence the above expression can be approximated by

$$\Lambda \int \psi''(f_0) NN^T q_0 \int \psi''(f_0) NN^T q_0^{-1} \int \psi''(f_0) NN^T q_0 \tilde{A}^T = \Lambda \int \psi''(f_0) NN^T q_0 \tilde{A}^T,$$
which, again by the uniform closeness of $C$ and $\tilde{C} = \Lambda N$, is
\[
\int \psi''(f_0)CC^T q_0 + o_{P^n}(1) = V_{\theta_0}^{-1} \int \psi''(f_0)\tilde{f}_0\tilde{f}_0^T q_0 V_{\theta_0}^{-1} + o_{P^n}(1) \xrightarrow{P^n} V_{\theta_0}^{-1},
\]
completing the proof that the asymptotic covariance matrix converges to $P^n$-probability to $V_{\theta_0}^{-1}$.

Finally, we show that $\mu_{n3} \xrightarrow{\mathcal{D}} N(0, V_{\theta_0}^{-1})$. Since conditionally on $t_1, \ldots, t_n$, $\mu_{n3}$ is the average of a set of independent random variables in a triangular array framework with uniformly bounded fourth order moments, it suffices to note that $D(\mu_{n3}) = \Sigma_{n3} \xrightarrow{P^n} V_{\theta_0}^{-1}$.

5. Simulation

In this section, we conduct two simulation studies to compare the proposed methods with a baseline method over the additive-error regression model and explore the finite-sample performance of the proposed methods in generalized regression models. The problem setup, methods and results of the additive-error regression and the generalized regression are introduced in Section 5.1 and Section 5.2, respectively.

5.1. Additive-error Regression

We first consider the situation that the targets are generated from the ordinary additive-error regression, where the error has a Gaussian distribution and the regression function follows the heat equation
\[
\frac{\partial f_\theta(t_1, t_2)}{\partial t_2} = \theta \frac{\partial^2 f_\theta(t_1, t_2)}{\partial t_1^2}, \quad (t_1, t_2) \in [0, 1] \times [0, 1],
\]
where the boundary conditions are given by
\[
f_\theta(t_1, 0) = \sin(\pi t_1); \quad f_\theta(0, t_2) = 0; \quad f_\theta(1, t_2) = 0.
\]

Given the heat equation, the data generating regimes are introduced for each model as follows:

- Step 1: Equispaced grid points for $t_1$ and $t_2$ are used for predictor values to avoid the sparsity of data in some regions and therefore, we pretend that $(t_1, t_2)$ jointly have a uniform distribution on $[0, 1] \times [0, 1]$.
- Step 2: The values of the function $f_\theta(t_1, t_2)$ are obtained through numeric solution on each data point $t_1$ and $t_2$ and it is denoted by $\tilde{f}_\theta(t_1, t_2)$.
- Step 3: Sample the target $Y$ from $\mathcal{N}(\tilde{f}_\theta(t_1, t_2), 1)$.

We consider the sample size $n$ equal to 400, 1600 and 6400, respectively and the true value of $\theta$ varies in $\theta_0 = 0.001, 0.005, 0.01$. Then, four methods are used to estimate $\theta$: the Maximum Likelihood Estimator (MLE), the Basic Two-step Bayesian Method (BTS), the Efficient Two-step Bayesian Method (ETS) and the Finite-Difference Two-Step Bayesian Method (FDT), where the MLE is treated as the baseline model. To compute the MLE, we first create a grid of $\theta$, which covers the true value. This grid consists of equispaced points in the log-scale by 0.05 from $e^{-10}$ to 1, which has 200 different values, because this grid is finer around the small values where the true values are located, and is coarser around the larger values. Then, using each $\theta$ in the grid, we calculate the numerical solutions $\tilde{f}_\theta(t_1, t_2)$ at each data point and minimize the
Bayesian regression for PDE models

Table 1. Point estimation biases and its standard errors of the four methods, where MLE denotes the Maximum Likelihood Estimator, BTS denotes the Basic Two-step Bayesian Method, FDTS denotes the Finite-Difference Two-Step Bayesian Method and ETS denotes the Efficient Two-step Bayesian Method.

<table>
<thead>
<tr>
<th>n</th>
<th>( \theta_0 )</th>
<th>MLE</th>
<th>BTS</th>
<th>FDTS</th>
<th>ETS</th>
</tr>
</thead>
<tbody>
<tr>
<td>400</td>
<td>0.004</td>
<td>0.0002 (0.0007)</td>
<td>0.0220 (0.0105)</td>
<td>0.0265 (0.0081)</td>
<td>0.0046 (0.0009)</td>
</tr>
<tr>
<td></td>
<td>0.005</td>
<td>0.0004 (0.0008)</td>
<td>0.0283 (0.0100)</td>
<td>0.0242 (0.0095)</td>
<td>0.0035 (0.0009)</td>
</tr>
<tr>
<td></td>
<td>0.010</td>
<td>0.0008 (0.0009)</td>
<td>0.0255 (0.0111)</td>
<td>0.0192 (0.0082)</td>
<td>0.0012 (0.0013)</td>
</tr>
<tr>
<td>1600</td>
<td>0.001</td>
<td>0.0001 (0.0004)</td>
<td>0.0138 (0.0045)</td>
<td>0.0146 (0.0038)</td>
<td>0.0016 (0.0038)</td>
</tr>
<tr>
<td></td>
<td>0.005</td>
<td>0.0002 (0.0004)</td>
<td>0.0107 (0.0049)</td>
<td>0.0108 (0.0044)</td>
<td>0.0002 (0.0051)</td>
</tr>
<tr>
<td></td>
<td>0.010</td>
<td>0.0004 (0.0004)</td>
<td>0.0067 (0.0057)</td>
<td>0.0063 (0.0041)</td>
<td>0.0002 (0.0063)</td>
</tr>
<tr>
<td>6400</td>
<td>0.001</td>
<td>0.0000 (0.0002)</td>
<td>0.0094 (0.0033)</td>
<td>0.0100 (0.0030)</td>
<td>0.0009 (0.0022)</td>
</tr>
<tr>
<td></td>
<td>0.005</td>
<td>0.0001 (0.0002)</td>
<td>0.0065 (0.0043)</td>
<td>0.0063 (0.0033)</td>
<td>0.0002 (0.0031)</td>
</tr>
<tr>
<td></td>
<td>0.010</td>
<td>0.0002 (0.0002)</td>
<td>0.0031 (0.0049)</td>
<td>0.0023 (0.0037)</td>
<td>-0.0002 (0.0033)</td>
</tr>
</tbody>
</table>

negative log-likelihood to find the MLE. The estimations of the rest three methods are calculated as stated in Section 2. The order of B-spline functions is chosen as 3 and the priors for the coefficients of the B-spline basis are \( N(0, 10) \) truncated between \(-50\) and \(50\). We choose \( k_n \) to be 2 for \( n = 400 \) and to be 3 for \( n = 1600 \) and \( n = 6400 \). The weight function in the basic two-step method is chosen as \( w(t_1, t_2) = t_1^3(1 - t_1)^3t_2^3(1 - t_2)^3 \) and we choose \( h = n^{-2/3} \) in the finite-difference two-step method. For the ETS, we deploy the same grid of \( \theta \) as the MLE and calculate the numerical solutions of the heat equation over an equispaced grid of \( t \) in \([0, 1] \times [0, 1] \) with segment 0.01 for computing the numerical integration.

For each method, \( n \) and true \( \theta_0 \), we carry out 500 replications. Under each replication, for the ETS, the samples of size 1500 are directly drawn from the posterior distribution of \( \theta \) and then we use the last 500 to construct the 95% credible interval for \( \theta \). For the MLE, the 95% confidence intervals are presented. However, each \( \theta \) has to be positive, which is not guaranteed by the BTS and the FDTS. Therefore, we generate sufficiently many samples and the first 500 positive samples of \( \theta \) are kept to calculate the 95% credible interval for \( \theta \). In fact, 1000 posterior samples are generated excluding the burn-in in each case.

We report the estimation biases of the posterior mean from the aforementioned methods and its standard error over these 500 replications shown in Table 1, where the standard errors are given inside the parentheses. From the table, both the estimation bias and its standard error of all the scenarios decrease as the sample size becomes larger. Comparing all four methods, we notice that the MLE has the best estimation accuracy and the smallest standard deviation, since this is the only method that does not rely on any approximation. Among all the three two-step methods, the ETS outperforms the other two with a smaller bias while the standard errors are comparable to these three methods. The results of the BTS and the FDTS are mostly comparable, which makes sense as the FDTS is almost the same as BTS except that it uses the finite difference operator.

We also calculate the coverage and the average length of the corresponding credible intervals over these 500 replications shown in Table 2, where the interval lengths are given inside the parentheses. From the table, lengths of the credible intervals shrink in all cases as \( n \) grows. This indicates that the posterior concentrates around its center. On the other hand, generally speaking, a smaller credible interval will provide a worse coverage result. Therefore, the coverage of the ETS-based credible interval is the worst, while its length is the smallest. Except for the case
Table 2. Coverages and average lengths of the credible/confidence intervals, where MLE denotes the Maximum Likelihood Estimator, BTS denotes the Basic Two-step Bayesian Method, FDTS denotes the Finite-Difference Two-Step Bayesian Method and ETS denotes the Efficient Two-Step Bayesian Method.

<table>
<thead>
<tr>
<th>n</th>
<th>( \theta_0 )</th>
<th>MLE</th>
<th>BTS</th>
<th>FDTS</th>
<th>ETS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.79 (0.00255)</td>
<td>0.200 (0.085)</td>
<td>0.320 (0.074)</td>
<td>0.076 (0.0012)</td>
</tr>
<tr>
<td></td>
<td>0.005</td>
<td>0.798 (0.0037)</td>
<td>0.948 (0.086)</td>
<td>0.972 (0.077)</td>
<td>0.096 (0.0018)</td>
</tr>
<tr>
<td></td>
<td>0.010</td>
<td>0.652 (0.00395)</td>
<td>0.978 (0.089)</td>
<td>0.996 (0.078)</td>
<td>0.080 (0.0021)</td>
</tr>
<tr>
<td>1600</td>
<td>0.001</td>
<td>0.788 (0.0015)</td>
<td>0.712 (0.037)</td>
<td>0.746 (0.041)</td>
<td>0.08 (0.00057)</td>
</tr>
<tr>
<td></td>
<td>0.005</td>
<td>0.806 (0.00186)</td>
<td>0.988 (0.039)</td>
<td>0.994 (0.041)</td>
<td>0.054 (0.00096)</td>
</tr>
<tr>
<td></td>
<td>0.010</td>
<td>0.584 (0.00189)</td>
<td>0.994 (0.040)</td>
<td>1.000 (0.043)</td>
<td>0.052 (0.00093)</td>
</tr>
<tr>
<td>6400</td>
<td>0.001</td>
<td>0.782 (0.00088)</td>
<td>0.770 (0.024)</td>
<td>0.842 (0.027)</td>
<td>0.042 (0.00031)</td>
</tr>
<tr>
<td></td>
<td>0.005</td>
<td>0.786 (0.00092)</td>
<td>0.960 (0.025)</td>
<td>0.994 (0.027)</td>
<td>0.048 (0.00047)</td>
</tr>
<tr>
<td></td>
<td>0.010</td>
<td>0.536 (0.00094)</td>
<td>0.980 (0.028)</td>
<td>1.000 (0.029)</td>
<td>0.044 (0.00041)</td>
</tr>
</tbody>
</table>

where \( \theta = 0.001 \), the coverage of the BTS is closer to the target value compared with the FDTS and both methods outperform the results from the MLE, although both are on the conservative side. This may be due to the extra error term in the finite-difference approximation. Moreover, when \( \theta = 0.005 \) or 0.01, the coverage of the BTS decays towards the target value as the sample size \( n \) is changed from 1600 to 6400, which reflects the convergence of the credible interval coverage. When \( \theta = 0.001 \), the results indicate that a larger sample size might be needed for the small magnitude of the true value to gain an acceptable estimation. The lengths of the credible intervals are slightly increasing while the truth \( \theta_0 \) increases and those lengths from the FDTS always maintain a larger value than that from the BTS except when the sample size is small, i.e. \( n = 400 \). The finite-difference approximation might be one of the reasons for causing this difference.

Finally, we compare the computation time of the second step (which is the actual difference) among all the three two-step methods when the sample size is 6400. On average, the ETS takes 9.3 hours for each replicate, the BTS takes about 44.6 minutes and the FDTS takes about 7.2 minutes. Considering that the mesh-width of the FDTS increases as the sample size increases, we summarize that the fastest method is the FDTS even when the sample size is relatively large, say 6400. The ETS is the slowest method, which we would not recommend in practice due to the substantially higher computational cost. Therefore, the FDTS with a coarser grid is computationally much faster than the other methods. The trade-off between computational speed and accuracy will be an important factor for choosing between the BTS and the FDTS, on which we will mainly focus in the following discussions.

Next we explore how \( f_\theta(t_1, t_2) \) changes over different \( \theta \) at some \( t_1 \) and \( t_2 \). Therefore, we consider \( t_1 = (0.05, 0.25, 0.5, 0.75, 0.95) \) to plot the credible intervals of the numerical solution of \( f_\theta(t_1, t_2) \) over \( t_2 \) with the estimate of \( \theta \) from each replication. The same procedure is repeated to create the plots over \( t_2 = (0.05, 0.25, 0.5, 0.75, 0.95) \). Here we only display the results of the BTS and the FDTS since these two methods provide acceptable results at a reasonable computational cost. The credible intervals in Figure 1 are computed over the posterior sample means of \( \theta \) in each of the 500 replications. These plots demonstrate the convergence of the estimator as \( n \) increases by checking that the credible intervals gradually contain the truth from the blue lines to the red ones. Because the boundary conditions of \( t_1 \) are two-sided while that of \( t_2 \) is only at the origin, the plots of \( t_1 \) and \( 1 - t_1 \) are alike while the credible intervals become larger with \( t_2 \) in-
Figure 1: The numerically solved function values \( \tilde{f}_\theta(t_1,t_2) \) over different \( t_1 \) and \( t_2 \) using the posterior sample means of \( \theta \) in each replication. The black solid line is the one using the true value \( \theta = 0.01 \) and the rest form the 95\% credible intervals where the estimations of \( \theta \)'s of the solid line are calculated by the Basic Two-step Bayesian Method and the ones of the dash line are calculated by the Finite-Difference Two-Step Bayesian Method. The blue lines stand for the results of \( n = 400 \), green lines for \( n = 1600 \) and red lines for \( n = 6400 \).
Table 3. Coverages and average lengths of the Bayesian credible intervals, where BTS denotes the Basic Two-step Bayesian Method and FDTS denotes the Finite-Difference Two-Step Bayesian Method.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$\theta_0$</th>
<th>Poisson</th>
<th></th>
<th>Logistic</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>BTS</td>
<td>FDTS</td>
<td>BTS</td>
<td>FDTS</td>
</tr>
<tr>
<td>400</td>
<td>0.001</td>
<td>0.438 (0.055)</td>
<td>0.424 (0.057)</td>
<td>0.094 (0.119)</td>
<td>0.082 (0.117)</td>
</tr>
<tr>
<td></td>
<td>0.005</td>
<td>0.942 (0.057)</td>
<td>0.968 (0.057)</td>
<td>0.934 (0.122)</td>
<td>0.940 (0.123)</td>
</tr>
<tr>
<td></td>
<td>0.010</td>
<td>0.976 (0.060)</td>
<td>0.994 (0.060)</td>
<td>0.994 (0.119)</td>
<td>0.994 (0.120)</td>
</tr>
<tr>
<td>1600</td>
<td>0.001</td>
<td>0.810 (0.027)</td>
<td>0.726 (0.035)</td>
<td>0.644 (0.048)</td>
<td>0.616 (0.054)</td>
</tr>
<tr>
<td></td>
<td>0.005</td>
<td>0.980 (0.028)</td>
<td>1.000 (0.036)</td>
<td>0.998 (0.047)</td>
<td>0.996 (0.055)</td>
</tr>
<tr>
<td></td>
<td>0.010</td>
<td>0.998 (0.030)</td>
<td>1.000 (0.037)</td>
<td>1.000 (0.049)</td>
<td>1.000 (0.055)</td>
</tr>
<tr>
<td>6400</td>
<td>0.001</td>
<td>0.856 (0.016)</td>
<td>0.880 (0.021)</td>
<td>0.686 (0.038)</td>
<td>0.668 (0.044)</td>
</tr>
<tr>
<td></td>
<td>0.005</td>
<td>0.970 (0.018)</td>
<td>0.992 (0.022)</td>
<td>0.992 (0.039)</td>
<td>0.990 (0.044)</td>
</tr>
<tr>
<td></td>
<td>0.010</td>
<td>0.990 (0.020)</td>
<td>1.000 (0.023)</td>
<td>0.998 (0.041)</td>
<td>1.000 (0.046)</td>
</tr>
</tbody>
</table>

Checking that the BTS gives a larger credible interval than the FDTS in most cases, we conclude that the BTS could provide better coverage of the true function than the FDTS, which deploys the finite difference approximation, even though the BTS is computationally more expensive than the FDTS.

5.2. Generalized Regression

We then explore the situation that the targets are generated from the generalized regressions, for example, Poisson regression or logistic regression to evaluate the finite-sample performance of the proposed methods. Here, we only compare the BTS with the FDTS by a simulation study, since the ETS is more computationally intensive and harder to use in practice.

We again consider the heat equation alike Section 5.1 and the same way to generate the observations except that the Step 3 is replaced by

- Step 3: Sample the target $Y$ from the specific model
  - Poisson regression: $Y$ are sampled from $Poisson(\exp(\tilde{f}(t_1, t_2)))$;
  - Logistic regression: $Y$ are sampled from $Binomial(1, \frac{\exp(\tilde{f}(t_1, t_2))}{1+\exp(\tilde{f}(t_1, t_2))})$.

The rest of the setup, methods and manipulations are all the same as those in the ordinary additive-error regression except that the likelihood function and the objective function are changed towards the generalized regression model accordingly. Then we present the coverages and the average lengths of the 95% credible intervals over the 500 replications in Table 3, where the interval lengths are given inside the parentheses.

All the conclusions from the coverage and lengths of credible intervals are almost identical to the ones from the ordinary additive-error regression in Subsection 5.1, which demonstrates that the performance of these two proposed methods is acceptable in the generalized regression. On the other hand, the computational cost of the generalized regression models is comparable to the ones of ordinary additive-error regression. However, since the binary data will cause a loss of information compared to continuous or discrete data, larger sample size is needed for the logistic regression to achieve a similar estimation accuracy than in the Poisson regression or the ordinary additive-error regression.
6. Discussion

In this paper, we considered only the situation that the PDE model is well-specified, that is, the true data generating distribution belongs to the model. For ODE models with normal additive error, Bhaumik and Ghosal [2, 4, 3] allowed the true data generating distribution to lie outside the model, that is, the model can be misspecified. In the PDE model, this is also possible for normal additive errors in regression. However, in the generalized regression setting, for any of the three projection-posterior methods, a major step is to use the asymptotic normality of the posterior distribution of the B-spline coefficients before projection. Such a result is available only under the assumption that the true distribution is of that form for some value of the true parameter. Then we needed to convert the distributional properties under the true distribution through a contiguity result, which requires that the corresponding parameters are within $o(n^{-1/2})$. If the model is misspecified, there will be a fixed difference with the B-spline model, and hence the resulting distributions will not be mutually contiguous. The problem cannot be avoided even in ODE models, unless a misspecified Bernstein-von Mises theorem for increasing-dimensional exponential families is first obtained. For the normal regression model, the posterior for the B-spline coefficients is exactly normal and direct calculations allow derivations even under misspecification.

We considered only the setting of stochastic predictors. For all proposed methods, deterministic predictors can be addressed, provided that the estimates in Lemma 4.2 continue to hold, because the key result on posterior asymptotic normality holds for independent, non-identically distributed predictors. Indeed, we conditioned on the observed values of the predictors, restricted to an appropriate set of high probability, to apply the posterior asymptotic normality result for tensor product B-spline coefficients, from Theorem 2.1 of Ghosal [10]. The estimates for the eigenvalues can be assured if the predictor values satisfy (2.2) of Yoo and Ghosal [30].

References


