Efficient constrained Gaussian process approximation using elliptical slice sampling

Hassan Maatouk, Didier Rullière, Xavier Bay

Efficient constrained Gaussian process approximation using elliptical slice sampling

Hassan Maatouk[†,*1], Didier Rullière[‡2] and Xavier Bay[‡3]

(†) CY Cergy Paris Université, CY Tech Site du Parc, 95011 Cergy-Pontoise, France
(∗) EBInnov, School of Industrial Biology, 49 avenue des Genottes, CS 90009, 95895 Cergy Cedex, France
(‡) Mines Saint-Étienne, Univ Clermont Auvergne, CNRS, UMR 6158 LIMOS, Institut Henri Fayol, Saint-Étienne, F-42023, France

Abstract

In this paper, Bayesian shape-restricted function estimation using constrained Gaussian processes (GPs) is revisited. The finite-dimensional Gaussian process approximation proposed in [26] is considered. This approximation verifies a wide range of shape constraints such as monotonicity, convexity and boundedness constraints in the entire domain. Through this approach, shape constraints are reformulated as *equivalent* linear inequality constraints on the basis coefficients. To generate a sample from the resulting constrained posterior distribution, we employ a recently efficient circulant embedding technique. This technique involves absorbing a smooth relaxation of the constraint set into the likelihood, a prior distribution, and elliptical slice sampling (ESS). Our contribution in this article is threefold. First, we extend this technique to address sets of linear, quadratic and nonlinear inequalities, enabling the incorporation of more general and multiple shape constraints. These constraints can be applied individually, jointly, and sequentially. Furthermore, this generalization allows the proposed approach to be easily adapted to other basis functions and models. Second, we explore efficient samplers to approximate both the posterior and prior distributions, including Hamiltonian Monte Carlo and the Fast Fourier Transform. Furthermore, we employ a highly efficient, large-scale approach for sampling from the prior distribution, resulting in significant computational advantages. Third, we investigate the capability of this approach to handle higher-dimensional input spaces and manage a large number of observations. The proposed approach demonstrates flexibility, accuracy, and efficiency in both synthetic and real data studies.

Keywords Elliptical slice sampling · nonparametric regression · shape constraints · smooth relaxation · Toeplitz.

1 Introduction

Gaussian Processes (GPs) are employed as prior distributions over function spaces, used to solve tasks such as regression, classification, feature extraction or hyperparameter optimization. Their applicability extends across a multitude of disciplines, ranging from computer science to physics,

[1]hmk@cy-tech.fr
[2]drulliere@emse.fr
[3]bay@emse.fr

biology, engineering, finance, and machine learning [15, 20, 36, 45]. GP models are specified by their mean and covariance functions, which together define the behavior of the process. The flexibility of GPs primarily arises from the covariance function. This function enables the integration of prior knowledge, encompassing features such as smoothness, stationarity, and sparsity. It also allows for the imposition of constraints on derivatives, further enhancing the model's adaptability to diverse data characteristics [11].

Various real-world applications in fields such as nuclear physics [48] and econometrics [5, 9, 10, 12, 21] have demonstrated situations where the data indicate that the underlying function adheres to specific shape constraints. These constraints include properties like monotonicity, convexity, and boundedness. Incorporating such constraints remains an active area of research, aiming to enhance the applicability of GP models in diverse settings [6, 17, 23, 24, 25, 39, 43, 44]. Recently, the authors in [40] provided an overview and survey of various strategies for integrating shape constraints into GPs. In [42], constrained GPs have been utilized to estimate probability density functions (pdfs).

In the present paper, the finite-dimensional GP approximation originally proposed in [26] is considered. To the best of our knowledge, this is the only GP model in the literature that is capable of handling a wide range of shape constraints, such as monotonicity, boundedness, and convexity, whether applied alone, together and sequentially. The main idea is to approximate the parent GP using a finite-dimensional GP achieved through an appropriate basis expansion. These basis functions exhibit attractive properties not necessary shared by other basis such as Bernstein polynomials [13], regression splines [4, 30], and restricted splines [38]. Various constraints, including monotonicity, convexity, and boundedness, are *equivalently* translated into linear inequality constraints on the basis coefficients. This necessitates sampling from a high-dimensional truncated multivariate normal (tMVN) distribution, for which existing algorithms include Gibbs sampling [41], Metropolis-Hastings (MH) [8, 32], the minimax tilting method accept-reject sampler [2], and the highly efficient Hamiltonian Monte Carlo (HMC) sampler [34].

Recently, a new highly efficient approach for sampling from a tMVN distribution restricted to the positive orthant has been developed in [37]. Their method combines Elliptical Slice Sampling (ESS) developed in [31], a modified prior, and a *smooth relaxation* of hard constraints. Therefore, sampling from the full conditional tMVN distribution is accomplished by sampling from a prior MVN distribution. The authors in [37] employ the highly efficient sampler, based on the fast Fourier transform (FFT), for stationary GPs on a regular grid [47], denoted as samp.WC. Our contribution in this paper is threefold. First, we extend the approach in [37] to handle general linear inequalities. This extension enables the incorporation of various and multiple shape constraints, which can be applied individually, jointly, and sequentially. Moreover, this generalization allows the proposed approach to be easily adapted to other basis functions and models like those mentioned in the introduction. Second, we employ highly efficient large-scale methods for sampling from the prior distribution [28, 29], resulting in significant computational advantages. We further explore and compare this approach with other highly efficient samplers, such as HMC and FFT. Third, we explore the capability of this approach to handle higher-dimensional input spaces and manage a large number of observations.

This article is organized as follows: Section 2 introduces the nonparametric function estimation problem. In Section 3, we provide a brief review of the finite-dimensional GP approximation proposed in [26]. Section 4 outlines the development of the proposed approach for sampling from the posterior tMVN distribution. The performance of the proposed approach is demonstrated through numerical experiments in this section. Section 5 extends the proposed approach to multidimensional input spaces. Finally, Section 6 showcases the effectiveness of the proposed framework through applications with real-world data.

2 Shape-restricted function estimation

In this section, the nonparametric function estimation through GP regression is considered, where the unknown function satisfies shape constraints such as monotonicity, boundedness, or convexity. The methodology developed in the present paper is able to incorporate multiple shape constraints, either applied together and sequentially. We denote by $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ the set of n training data, where $\{\boldsymbol{x}_i\}$ and $\{y_i\}$ are the covariates and the data, respectively. The following regression problem is considered:

$$y_i = f(\boldsymbol{x}_i) + \epsilon_i, \quad \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2), \tag{1}$$

$i = 1, \ldots, n$, where f represents an unknown function that generates the data $\boldsymbol{y} = [y_1, \ldots, y_n]^\top$, $\boldsymbol{x}_i \in \mathbb{R}^d$ is a covariate of dimension d, and ϵ_i is an additive independent and identically distributed (i.i.d.) zero-mean Gaussian noise with a constant variance of σ^2. A GP prior distribution on the unknown function f is assumed [45]. Let $(Y(x))_{x\in\mathcal{D}}$ be a zero-mean GP with covariance function k, where $\mathcal{D}$ is a subset of $\mathbb{R}^d$, i.e., $Y \sim \mathcal{GP}(0, k)$. Throughout this paper, we assume, without loss of generality, that $\mathcal{D} = [0, 1]^d$. In GP modeling, the positive semi-definite covariance function k is crucial. It encodes assumptions about the underlying function, including differentiability and periodicity, through kernels. The Matérn family of kernels is widely used in computer experiments and, in the univariate case, is defined as

$$k(x, x') = \tau^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}}{\ell} |x - x'| \right)^\nu B_\nu \left(\frac{\sqrt{2\nu}}{\ell} |x - x'| \right), \tag{2}$$

for any $x, x' \in \mathcal{D}$, where $\Gamma(\cdot)$ is the Gamma function and $B_\nu(\cdot)$ denotes the modified Bessel function of the second kind of order ν. The parameter ν regulates the degree of smoothness of the GP sample paths. A process with the Matérn kernel of order ν is $\lceil \nu - 1 \rceil$ times differentiable [45]. The positive parameters τ^2 and ℓ are referred to as the signal variance and the correlation length-scale, respectively. These parameters are generally unknown and need to be estimated from data. We refer the reader to [16, 45] for a more detailed explanation on parameter estimation techniques.

In the following section, we briefly review the finite-dimensional approximation of GPs as proposed in [26]. This approach enables the satisfaction of various shape constraints, including monotonicity, convexity, and boundedness within both convex and non-convex sets. Furthermore, this method can accommodate a wide range of shape constraints, whether applied individually, collectively, and sequentially.

3 Finite-dimensional Gaussian approximation

In this section, the finite-dimensional approximation of GPs developed in [26] is briefly reviewed. Let $(Y(x))_{x\in\mathcal{D}}$ be a zero-mean GP with covariance function, where $\mathcal{D} \subset \mathbb{R}^d$. For the sake of simplicity, we begin with the one-dimensional case, i.e., $d = 1$ and $\mathcal{D} = [0, 1]$. We first introduce some notations. We denote by $\{u_j,\ j = 1, \ldots, N\}$ the set of equally spaced knots on $\mathcal{D}$ with spacing $\Delta_N = 1/(N - 1)$, where $u_j = (j - 1)\Delta_N$. Let us mention that the methodology developed in the present paper is also applicable to non-uniform discretization of $\mathcal{D}$. We recall the basis functions $\{\phi_j\}$ that will be used in the model developed in the present paper:

$$\phi_j(x) := \phi \left(\frac{x - u_j}{\Delta_N} \right), \tag{3}$$

for $j \in \{1, \ldots, N\}$, where $\phi(x) := (1 - |x|)\mathbf{1}_{[-1,1]}(x)$ is the *hat* function on $[-1, 1]$. It is worth noting that other basis functions, such as those suggested in [26, 27], can be easily adapted to the

methodology developed in the present paper. Let us recall that the *hat* functions $\{\phi_j\}$ in (3) have two nice properties. First, the value of any *hat* function at any knot is equal to Kronecker's delta function (i.e., $\phi_j(u_l) = \delta_{j,l}$), where $\delta_{j,l}$ is equal to one when $j = l$ and zero otherwise. Second, for any $x \in \mathcal{D}$, we have $\sum_{j=1}^{N} \phi_j(x) = 1$. Any continuous function $f : \mathcal{D} \to \mathbb{R}$, that is, $f \in C^0(\mathcal{D}, \mathbb{R})$ can be approximated by a piecewise linear interpolating between the function values at the knots $\{u_j\}$,

$$\tilde{f}_N(x) = \sum_{j=1}^{N} f(u_j)\phi_j(x), \quad \forall x \in \mathcal{D}, \tag{4}$$

where $\{\phi_j\}$ are the basis functions defined in (3).

Proposition 1 (Uniform convergence)**.** *For any $f \in C^0(\mathcal{D}, \mathbb{R})$, the piecewise linear function $\tilde{f}_N$ defined in* (4) *converges uniformly to f when N tends to infinity.*

Proof. From $\phi(x) \geq 0$ and $\sum_{j=1}^{N} \phi_j(x) = 1$, for any $x \in \mathcal{D}$, we have

$$\begin{aligned} |\tilde{f}_N(x) - f(x)| &= \left| \sum_{j=1}^{N} f(u_j)\phi_j(x) - f(x) \sum_{j=1}^{N} \phi_j(x) \right| = \left| \sum_{j=1}^{N} \left[f(u_j) - f(x) \right] \phi_j(x) \right| \\ &\leq \sum_{j=1}^{N} \left[|f(u_j) - f(x)| \phi_j(x) \right] \leq \sup_{|x-x'| \leq \Delta_N} |f(x) - f(x')| \,. \end{aligned}$$

The proof of the proposition holds due to the fact that f is uniformly continuous on the compact interval $\mathcal{D}$. □

Now, the finite-dimensional GP approximation is defined as follows:

$$Y^N(x) = \sum_{j=1}^{N} Y(u_j)\phi_j(x) = \sum_{j=1}^{N} \xi_j \phi_j(x), \quad x \in \mathcal{D}, \tag{5}$$

where we denote $Y(u_j) = \xi_j$ for any $j \in \{1, \ldots, N\}$. Given that $(Y(x)_{x \in \mathcal{D}} \sim \mathcal{GP}(0, k)$, the coefficient vector $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_N]^\top$ is a zero-mean Gaussian vector with a covariance matrix $\tau^2 \boldsymbol{K}$, where

$$\tau^2 \boldsymbol{K}_{j,l} = k(u_j - u_l), \quad j, l = 1, \ldots, N,$$

with τ^2 the signal variance. Due to the uniform discretization of the input domain $\mathcal{D}$, the covariance matrix $\tau^2 \boldsymbol{K}$ retains its stationary property and displays a Toeplitz structure. This property provides a significant computational advantage in the sampling procedure (Section 4).

3.1 Conditioning on noisy observations with shape constraints

In this section, we introduce some notations to simplify the problem's presentation. Consider $\mathcal{C}$ as the convex set of functions that satisfy shape constraints, such as monotonicity, convexity, and boundedness. The non-convex scenario has been investigated in [27]. For example, we may have

$$\mathcal{C} = \begin{cases} \mathcal{C}_b := \{f \in C^0(\mathcal{D}, \mathbb{R}) \text{ s.t. } l_b \leq f(x) \leq u_b, \ \forall x \in \mathcal{D}, \text{ and } l_b < u_b \in \mathbb{R}\} \\ \mathcal{C}_m := \{f \in C^0(\mathcal{D}, \mathbb{R}) \text{ s.t. } f(x) \leq f(x'), \ \forall x < x' \in \mathcal{D}\} \\ \mathcal{C}_c := \left\{f \in C^0(\mathcal{D}, \mathbb{R}) \text{ s.t. } \frac{f(x')-f(x)}{x'-x} \leq \frac{f(x'')-f(x')}{x''-x'}, \ \forall x < x' < x'' \in \mathcal{D}\right\} \end{cases} \tag{6}$$

which corresponds to boundedness, monotonicity (nondecreasing), and convexity constraints, respectively. Here, $C^0(\mathcal{D}, \mathbb{R})$ denotes the set of continuous functions from $\mathcal{D}$ to $\mathbb{R}$. As outlined in

[26], the basis functions $\{\phi_j\}$ defined in (3) ensure an *equivalent* reformulation of the functional constraints $Y^N \in \mathcal{C}$ as a finite set of linear inequality constraints on the basis coefficients $\{\xi_j\}$ for various shape constraints

$$Y^N \in \mathcal{C} \quad \Leftrightarrow \quad \boldsymbol{\xi} \in \mathcal{E} \tag{7}$$

where $\mathcal{E}$ denotes a set of linear inequality constraints on the basis coefficients $\boldsymbol{\xi} \in \mathbb{R}^N$. According to (7), we obtain

$$\mathcal{E} = \left\{ \begin{array}{l} \mathcal{E}_b := \left\{ z \in \mathbb{R}^N : \ l_b \leq z_j \leq u_b, \ \forall j = 1, \ldots, N \right\} \\ \mathcal{E}_m := \left\{ z \in \mathbb{R}^N : \ z_{j-1} \leq z_j, \ \forall j = 2, \ldots, N \right\} \\ \mathcal{E}_c := \left\{ z \in \mathbb{R}^N : \ \dfrac{z_{j-1} - z_{j-2}}{t_{j-1} - t_{j-2}} \leq \dfrac{z_j - z_{j-1}}{t_j - t_{j-1}}, \ \forall j = 3, \ldots, N \right\} \end{array} \right. \tag{8}$$

which corresponds to boundedness, monotonicity, and convexity constraints, respectively. Let us mention that the above linear inequality constraints on the basis coefficients (8) can be written in matrix form as follows:

$$\left\{ \boldsymbol{\xi} \in \mathbb{R}^N \text{ s.t. } \boldsymbol{A}\boldsymbol{\xi} + \boldsymbol{b} \geq \boldsymbol{0}_m \right\}, \tag{9}$$

where $\boldsymbol{A} \in \mathbb{R}^{m \times N}$ represents the matrix of constraints, and $\boldsymbol{b}$ is a m-dimensional vector. For instance, when $\boldsymbol{\xi} \in \mathcal{E}_m$, we obtain

$$\boldsymbol{A}_{i,j} = \left\{ \begin{array}{lll} -1 & \text{if } j = i & \text{for any } i = 1, \ldots, N-1; \\ 1 & \text{if } j = i+1 & \text{for any } i = 1, \ldots, N-1; \\ 0 & \text{otherwise;} & \end{array} \right. \tag{10}$$

and $\boldsymbol{b} = [0, \ldots, 0]^\top \in \mathbb{R}^m$. This implies $m = N - 1$ linear inequality constraints on the coefficients vector $\boldsymbol{\xi}$. Let us recall that $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_N]^\top$ is a vector extracted from the zero-mean GP Y with covariance function k. Therefore, $\boldsymbol{\xi}$ is a zero-mean Gaussian vector with a covariance matrix $\tau^2 \boldsymbol{K}$. Consequently, the pdf of the constrained distribution $\{\boldsymbol{\xi} \text{ s.t. } \boldsymbol{\xi} \in \mathcal{E}\}$ is given by

$$p(\boldsymbol{\xi} | \boldsymbol{\xi} \in \mathcal{E}) \propto \underbrace{\exp\left(-\frac{1}{2\tau^2} \boldsymbol{\xi}^\top \boldsymbol{K}^{-1} \boldsymbol{\xi} \right)}_{\text{(untruncated) prior}} \underbrace{\mathbf{1}_{\mathcal{E}}(\boldsymbol{\xi})}_{\text{constrained set}}, \tag{11}$$

where $\tau > 0$. The above pdf in (11) is proportional to the product of an untruncated MVN prior and an indicator function representing the set of linear inequality constraints.

Now, Model (5) considering both noisy observations and shape constraints, is expressed as follows:

$$Y^N(x) = \sum_{j=1}^{N} \xi_j \phi_j(x) \quad \text{s.t.} \quad \left\{ \begin{array}{ll} Y^N(x_i) + \epsilon_i = y_i & \text{(noisy observations),} \\ Y^N \in \mathcal{C} & \text{(shape constraints),} \end{array} \right. \tag{12}$$

where $x_i \in \mathcal{D}$ is the covariate, $y_i \in \mathbb{R}$ is the response and $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$, with σ^2 the noise variance. Following the equivalent in (7), the conditional distribution (12) can be written in matrix form as follows:

$$\boldsymbol{X}\boldsymbol{\xi} + \boldsymbol{\epsilon} = \boldsymbol{y}, \quad \boldsymbol{\xi} \in \mathcal{E}, \tag{13}$$

where $\boldsymbol{y} = [y_1, \ldots, y_n]^\top$ is the vector of data, $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_n]^\top$ is the noise zero-mean Gaussian vector, and $\boldsymbol{X}$ is the $n \times N$ design matrix defined by $\boldsymbol{X}_{i,j} := \phi_j(x_i)$. With only noisy observations $\{\boldsymbol{X}\boldsymbol{\xi} + \boldsymbol{\epsilon} = \boldsymbol{y}\}$, we have $\{\boldsymbol{\xi} | \boldsymbol{y}\}$ follows a MVN distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ (see, for example, [45]), where

$$\left\{ \begin{array}{l} \boldsymbol{\mu} = (\boldsymbol{X}^\top \boldsymbol{X}/\sigma^2 + \boldsymbol{K}^{-1}/\tau^2)^{-1} \boldsymbol{X}^\top \boldsymbol{y}/\sigma^2; \\ \boldsymbol{\Sigma} = (\boldsymbol{X}^\top \boldsymbol{X}/\sigma^2 + \boldsymbol{K}^{-1}/\tau^2)^{-1}. \end{array} \right. \tag{14}$$

According to the previous notations, the constrained posterior distribution (13) is the following truncated MVN:

$$\{\boldsymbol{\xi}|\boldsymbol{y}, \boldsymbol{A}\boldsymbol{\xi} + \boldsymbol{b} \geq \mathbf{0}_m\} \sim \mathcal{N}_{\mathcal{T}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{A}, \boldsymbol{b}), \tag{15}$$

where $\mathcal{N}_{\mathcal{T}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{A}, \boldsymbol{b})$ is the truncated MVN distribution with mean vector $\boldsymbol{\mu}$, covariance matrix $\boldsymbol{\Sigma}$, constraints matrix $\boldsymbol{A}$, and constraints vector $\boldsymbol{b}$. In this context, the efficient HMC sampler developed in [34] and implemented in the `R package` *tmg* can be used. However, in situations when N is large, $(\boldsymbol{X}^\top\boldsymbol{X}/\sigma^2 + \boldsymbol{K}^{-1}/\tau^2)$ keeps changing over each McMC iteration due to updates in σ^2 and τ^2. This necessitates an $N \times N$ matrix inversion at every iteration. Moreover, within a large McMC algorithm, updating the unknown covariance function parameters involves computing the inversion of an $N \times N$ matrix at each step.

In the following (Section 4), we develop an efficient approach for sampling from the constrained posterior distribution given by Equation (15). This approach avoids matrix inversions (14) and the need to sample from the full conditional tMVN distribution, as described in (15). Furthermore, it can be trivially adapted with other basis functions and models. It is worth noting that in contrast to the HMC sampler, the initial step of the proposed McMC approach does not necessitate the validation of linear inequality constraints.

3.2 Mean and Maximum a posterior estimates (mAP and MAP)

Before presenting alternative approaches of sampling from the constrained posterior distribution (13), we first define the two posterior estimates, namely maximum a posterior (MAP) and mean a posterior (mAP). Let $\boldsymbol{\mu}^*$ be the posterior mode that maximises the pdf of the posterior distribution (15). Then, maximizing the posterior pdf is equivalent to maximize the quadratic problem

$$\boldsymbol{\mu}^* := \underset{\boldsymbol{\xi} \text{ s.t. } \boldsymbol{l} \leq \boldsymbol{A}\boldsymbol{\xi} \leq \boldsymbol{u}}{\arg\max} \left\{ -[\boldsymbol{\xi} - \boldsymbol{\mu}]^\top \boldsymbol{\Sigma}^{-1} [\boldsymbol{\xi} - \boldsymbol{\mu}] \right\}, \tag{16}$$

with condition mean $\boldsymbol{\mu}$ and conditional covariance $\boldsymbol{\Sigma}$ as in (14). This is a quadratic optimization problem subject to linear inequality constraints (see [3, 18]). The optimization problem in (16) is equivalent to

$$\boldsymbol{\mu}^* := \underset{\boldsymbol{\xi} \text{ s.t. } \boldsymbol{l} \leq \boldsymbol{A}\boldsymbol{\xi} \leq \boldsymbol{u}}{\arg\min} \left\{ \boldsymbol{\xi}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\xi} - 2\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\xi} \right\}. \tag{17}$$

Finally, the posterior mode $\boldsymbol{\mu}^*$ can then serve as a suitable starting point for the McMC sampler. It can play an important role in avoiding numerical instability in the sampling procedure for a large value of the approximate parameter η (see Section 4).

Definition 1 (MAP estimate)**.** *The Maximum a posteriori (MAP) estimate of Y^N conditionally on shape constraints and noisy observations is defined as*

$$M^N(x) := \sum_{j=1}^{N} \mu_j^* \phi_j(x) = \Phi(x)^\top \boldsymbol{\mu}^*, \quad x \in \mathcal{D},$$

where $\boldsymbol{\mu}^* = [\mu_1^*, \ldots, \mu_N^*]^\top$ is the posterior mode computed by (17) *and $\Phi(x) = [\phi_1(x), \ldots, \phi_N(x)]^\top$.*

Comments on the MAP estimate

The MAP estimate is independent of the sampling process and is determined only by solving a quadratic optimization problem subject to linear inequality constraints (17). However, it does depends on the noise variance parameter σ^2. In the sampling procedure developed in Section 4, the noise variance parameter σ^2 is estimated at each McMC iteration using the joint posterior, as

presented in [37]. For this reason, we propose to plug the average of the estimated values of σ^2 into the quadratic optimization problem to obtain the MAP estimate. The asymptotic behavior of this estimate has been discussed in both [1] and [19], where they established a generalization of the Kimeldorf-Wahba correspondence [22] for constrained noisy and noisy-free cases.

Definition 2 (mAP estimate)**.** *The mean a posteriori (mAP) estimate of Y^N conditionally on shape constraints and noisy observations is defined as*

$$m^N(x) := \mathrm{E}\left[Y^N(x)\middle|\boldsymbol{y}, Y^N \in \mathcal{C}\right] = \Phi(x)^\top \overline{\boldsymbol{\mu}},$$

where $\overline{\boldsymbol{\mu}} := \mathrm{E}\left[\boldsymbol{\xi}|\boldsymbol{y}, \boldsymbol{\xi} \in \mathcal{E}\right]$ is the posterior mean which is computed from simulations, with $\mathcal{E}$ the set of linear inequality constraints on the basis coefficients (8)*, and $\Phi(x) = [\phi_1(x), \ldots, \phi_N(x)]^\top \in \mathbb{R}^N$.*

The methodology developed in the following section for sampling from the constrained posterior distribution (15) can also be applied to various other basis functions and models.

4 Constrained posterior sampling: alternative approaches

According to (13), we are interested in sampling from a constrained posterior distribution with the following form:

$$p(\boldsymbol{\xi}|\boldsymbol{y}, \boldsymbol{\xi} \in \mathcal{E}) \propto \underbrace{\exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\xi}\|^2\right)}_{\text{likelihood function}} \underbrace{\exp\left(-\frac{1}{2\tau^2}\boldsymbol{\xi}^\top \boldsymbol{K}^{-1}\boldsymbol{\xi}\right)}_{\text{(untruncated) MVN prior}} \underbrace{\mathbf{1}_{\mathcal{E}}(\boldsymbol{\xi})}_{\text{constraint}}, \quad \boldsymbol{\xi} \in \mathbb{R}^N, \tag{18}$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is the vector of data, $\boldsymbol{X} \in \mathbb{R}^{n \times N}$ is the matrix of design, $\mathcal{E}$ is a convex set in $\mathbb{R}^N$ representing the set of linear inequality constraints on the basis coefficients (8), and $\boldsymbol{K}$ is a positive definite matrix. The methodology developed in this section generally applies to sets of linear inequalities, as in (9), and to any such $\boldsymbol{K}$. However, our specific interest lies in situations where $\boldsymbol{K}$ is derived from the evaluation of a stationary covariance function on a regular grid, i.e., $\boldsymbol{K}_{j,l} = k(u_j - u_j)$, for any $j, l = 1, \ldots, N$. This means that the covariance matrix $\boldsymbol{K}$ exhibits a Toeplitz structure. The constrained posterior pdf in (18) is proportional to the product of a likelihood function, an (untruncated) MVN prior, and an indicator function representing the set of constraints. A simple calculation yields that $p(\boldsymbol{\xi})$ in (18) is the density function of the following MVN distribution (see e.g., [45], Section 2.1.1):

$$\mathcal{N}\left(\left(\frac{\boldsymbol{X}^\top\boldsymbol{X}}{\sigma^2} + \frac{\boldsymbol{K}^{-1}}{\tau^2}\right)^{-1}\frac{\boldsymbol{X}^\top\boldsymbol{y}}{\sigma^2}, \left(\frac{\boldsymbol{X}^\top\boldsymbol{X}}{\sigma^2} + \frac{\boldsymbol{K}^{-1}}{\tau^2}\right)^{-1}\right) \tag{19}$$

truncated to the set of linear inequality constraints $\mathcal{E}$ in $\mathbb{R}^N$. The main results developed in this section draw inspiration from a recent innovative approach introduced in [37]. This approach is based on absorbing a *smooth relaxation* of the constraint into the likelihood function, along with a circulant embedding technique for sampling the constrained posterior distribution (13). This strategy eliminates the necessity of sampling from the full conditional tMVN distribution (15) and avoids the need to perform matrix inversions. Moreover, this approach allows us to perform sampling before conditioning, rather than after, which can offer significant advantages in preserving the stationary property during the sampling process. Our aim in this paper is to generalize this approach in order to deal with sets of linear and nonlinear inequalities. This allows the incorporation of multiple shape constraints, either together or sequentially. Furthermore, it enables the proposed approach to be trivially adapted to other basis functions. Through this generalization, we retain the significant

computational benefits without modifying the prior distribution of the basis coefficients.

Our main idea is to approximate the indicator function in (18) by a smooth (infinitely differentiable) function and incorporate it into the likelihood function. Let us provide the following general result:

Proposition 2 (Approximation of indicator function of inequality constraints)**.** *Let g be any piecewise continuous function on $\mathbb{R}$. Then, the indicator function $\mathbf{1}_{\{g(\cdot)\geq 0\}}$ can be approximated by a scaled logistic sigmoid function as follows:*

$$\mathbf{1}_{\{g(\cdot)\geq 0\}}(x) \approx \frac{1}{1+\exp[-\eta g(x)]}, \quad x \in \mathbb{R}, \tag{20}$$

where the parameter $\eta > 0$ controls the quality of the approximation. As η increases, the quality of the approximation improves. The approximate function in (20) *possesses the same degree of differentiability as g and we have*

$$\left| \mathbf{1}_{\{g(\cdot)\geq 0\}}(x) - \frac{1}{1+e^{-\eta g(x)}} \right| \leq \frac{1}{1+e^{\eta |g(x)|}},$$

for any $x \in \mathbb{R}$.

Proof. Let us consider two cases: when $g(x) < 0$, then, $\mathbb{1}_{\{g(x)\geq 0\}} = 0$. Thus,

$$\left| \mathbf{1}_{\{g(\cdot)\geq 0\}}(x) - \frac{1}{1+e^{-\eta g(x)}} \right| \leq \frac{1}{1+e^{-\eta |g(x)|}} \quad \Leftrightarrow \quad \frac{1}{1+e^{-\eta g(x)}} \leq \frac{1}{1+e^{\eta |g(x)|}}.$$

The second case when $g(x) \geq 0$ holds since $1 - 1/[1+\exp\{-\eta g(x)\}] = 1/[1+\exp\{\eta g(x)\}]$, which completes the proof of the proposition. □

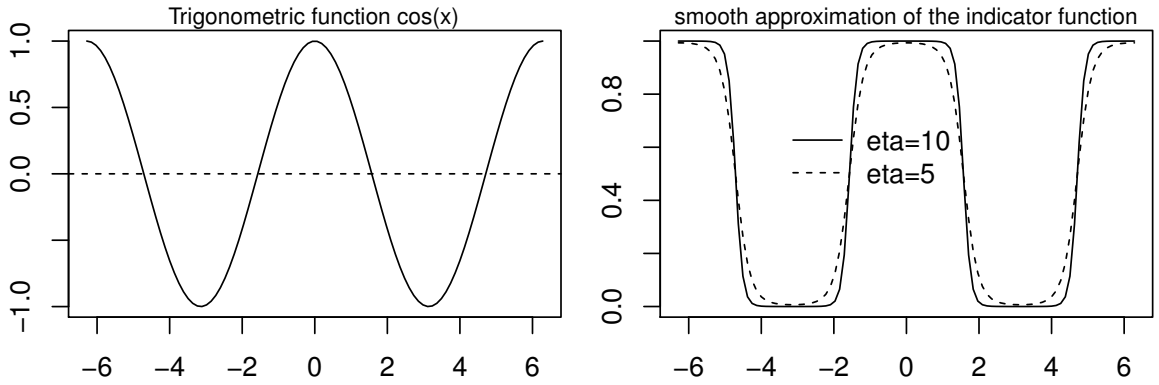

Figure 1: The nonlinear function $g(x) = \cos(x)$ on $[-2\pi, 2\pi]$ (left) and the smooth approximation of the corresponding indicator function $\mathbf{1}_{g(\cdot)\geq 0}(\boldsymbol{\xi})$ for two different values of η (right).

Figure 1 shows the the proposed approximation J_η of the indicator function in (20) for the trigonometric function $g(x) = \cos(x)$. The results in Proposition 2 can be applied to m nonlinear inequality constraints $g_i(\boldsymbol{\xi}) \geq 0$, for any $i = 1, \ldots, m$. Furthermore, these results can be used to handle m quadratic inequality constraints $\boldsymbol{\xi}^\top \boldsymbol{C}_i \boldsymbol{\xi} + \boldsymbol{d}_i^\top \boldsymbol{\xi} + e_i \geq 0$, for any $i = 1, \ldots, m$, where $\boldsymbol{C}_i$ is a $N \times N$ matrix of dimension, $\boldsymbol{d}_i$ is an N-dimensional vector, and e_i is a real number.

In the following section, we first focus on monotonicity constraints, where the set of linear inequality constraints on the basis coefficients $\mathcal{E}$ in (18) is replaced by $\mathcal{E}_m$, see Equation (8). The results of Proposition 2 well be employed in the algorithm developed in the next sections.

4.1 Monotonicity constraints

In this section, the monotonicity constraint is considered. The authors in [26] proved that from Model (5), Y^N is monotone (non-decreasing) *if and only if* $\boldsymbol{\xi} \in \mathcal{E}_m$, where

$$\mathcal{E}_m = \{z \in \mathbb{R}^N \ : \ z_{j-1} \leq z_j, \ j = 2, \ldots, N\}. \tag{21}$$

As a result, the constrained posterior pdf in (18) has the following form:

$$p(\boldsymbol{\xi}|\boldsymbol{y}, \boldsymbol{\xi} \in \mathcal{E}_m) \propto \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\xi}\|^2\right) \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\xi}^\top \boldsymbol{K}^{-1}\boldsymbol{\xi}\right) \mathbf{1}_{\mathcal{E}_m}(\boldsymbol{\xi}), \quad \boldsymbol{\xi} \in \mathbb{R}^N. \tag{22}$$

Sampling from the above constrained posterior pdf in (22) using an efficient circulant embedding technique can be summarized in three steps that avoid matrix inversions and direct sampling from the full conditional posterior tMVN distribution in (15).

First step: smooth approximation of the indicator function In the first step, we approximate the indicator function $\mathbf{1}_{\mathcal{E}_m}(\boldsymbol{\xi})$ in (22) through sigmoid functions as in Proposition 2. Specifically, we use the logistic sigmoid function $1/[1 + \exp(-x)]$, which corresponds to the cumulative distribution function (cdf) of the logistic distribution. To be more precise, we use the scaled logistic sigmoid approximation

$$\mathbf{1}_{\xi_1 \leq \xi_2} \approx \frac{e^{\eta(\xi_2 - \xi_1)}}{1 + e^{\eta(\xi_2 - \xi_1)}} = \frac{1}{1 + e^{-\eta(\xi_2 - \xi_1)}},$$

for large $\eta > 0$ to derive a smooth approximation called $\mathbb{J}_\eta(\cdot)$ for $\mathbf{1}_{\mathcal{E}_m}(\boldsymbol{\xi})$,

$$\mathbf{1}_{\mathcal{E}_m}(\boldsymbol{\xi}) \approx \mathbb{J}_\eta(\boldsymbol{\xi}) = \prod_{j=2}^{N} \frac{e^{\eta(\xi_j - \xi_{j-1})}}{1 + e^{\eta(\xi_j - \xi_{j-1})}} = \frac{e^{\eta(\xi_N - \xi_1)}}{\prod_{j=2}^{N}[1 + e^{\eta(\xi_j - \xi_{j-1})}]},$$

where η plays the role of controlling the quality of the approximation.

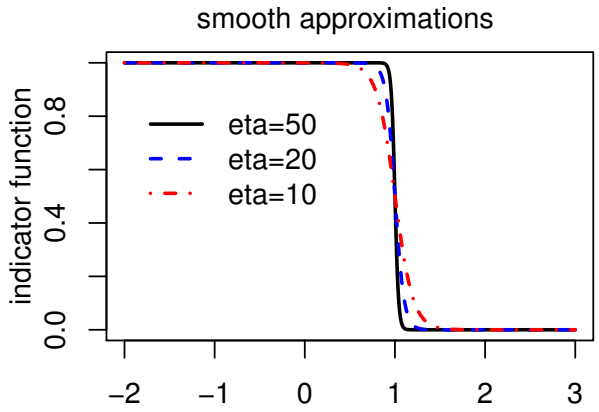

Figure 2: Smooth approximations of the indicator function $\mathbf{1}_{\xi_1 \leq \xi_2 = 1}$ for different values of η.

Figure 2 displays smooth approximations of the indicator function $\mathbf{1}_{\xi_1 \leq \xi_2 = 1} \approx \frac{1}{1+e^{-\eta(1-\xi_1)}}$ for different values of η on the interval $[-2, 3]$.

By substituting the smooth approximate $\mathbb{J}_\eta(\cdot)$ of the indicator function $\mathbf{1}_{\mathcal{E}_m}(\boldsymbol{\xi})$ in (22), we obtain the following density approximation $\tilde{p}$ of p, where the *smooth relaxation* of the constraints $\mathbb{J}_\eta$ has

been absorbed into the likelihood:

$$\begin{aligned}\tilde{p}(\boldsymbol{\xi}|\boldsymbol{y}, \boldsymbol{\xi} \in \mathcal{E}_m) &\propto \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\xi}\|^2\right)\exp\left(-\frac{1}{2\tau^2}\boldsymbol{\xi}^\top \boldsymbol{K}^{-1}\boldsymbol{\xi}\right)\mathbb{J}_\eta(\boldsymbol{\xi}) \\ &= \left[\exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\xi}\|^2\right)\mathbb{J}_\eta(\boldsymbol{\xi})\right]\exp\left(-\frac{1}{2\tau^2}\boldsymbol{\xi}^\top \boldsymbol{K}^{-1}\boldsymbol{\xi}\right) \\ &= \underbrace{\left[\exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\xi}\|^2\right)\prod_{j=2}^{N}\frac{e^{\eta(\xi_j-\xi_{j-1})}}{1+e^{\eta(\xi_j-\xi_{j-1})}}\right]}_{\textit{redefined likelihood}}\underbrace{\exp\left(-\frac{1}{2\tau^2}\boldsymbol{\xi}^\top \boldsymbol{K}^{-1}\boldsymbol{\xi}\right)}_{\text{(untruncated) MVN prior}}. \quad (23)\end{aligned}$$

The parameter η plays the role of controlling the quality of the approximation; a higher value of η yields a better approximation.

Second step: elliptical slice sampling In the second step, we use the efficient ESS method developed in [31] to sample from the approximate posterior distribution in (23). Let us briefly review the ESS. The ESS is a general technique employed for sampling from distributions with the following form:

$$p(\boldsymbol{\xi}) \propto L(\boldsymbol{\xi})\mathcal{N}(\boldsymbol{\xi}; \boldsymbol{0}, \boldsymbol{K}). \quad (24)$$

The above density (24) is proportional to a product of a general likelihood function $L(\cdot)$ and a zero-mean MVN prior, where L represents the *redefined likelihood* function in (23). In this context, sampling from (24) can be performed using Metropolis-Hastings (MH) proposals [8, 32]:

$$\boldsymbol{\xi}' = \rho\boldsymbol{\nu} + \sqrt{1-\rho^2}\boldsymbol{\xi}, \quad \boldsymbol{\nu} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{K}), \quad (25)$$

where $\rho \in [-1, 1]$ is a step-size parameter, $\boldsymbol{\xi}$ is the current state, and $\boldsymbol{\xi}'$ is the proposal state. Let us recall that the MH acceptance ratio $\alpha = \min\{1, L(\boldsymbol{\xi}')/L(\boldsymbol{\xi})\}$ is only relies on the likelihood ratio and is independent of ρ. Moreover, this method is simple to implement and can be immediately applied to a much wider variety of models with Gaussian priors.

The ESS is founded on the parametrization $\rho = \sin(\theta)$ in (25), which offers an adaptive and automated approach for tuning the step-size parameter ρ, ensuring acceptance at each step. Thus, the MH proposal in (25) is reformulated as follows:

$$\boldsymbol{\xi}' = \sin(\theta)\boldsymbol{\nu} + \cos(\theta)\boldsymbol{\xi}, \quad \boldsymbol{\nu} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{K}), \quad (26)$$

where the angle θ is uniformly generated from a $[\theta_{\min}, \theta_{\max}]$ interval which is shrunk exponentially fast until an acceptable state is reached. For each specific value of θ, a uniform random number is generated and compared to the likelihood ratio $L(\boldsymbol{\xi}')/L(\boldsymbol{\xi})$. If the proposal $\boldsymbol{\xi}'$ is not acceptable, one shrinks the bracket of θ, and continues this process until acceptance. Detailed instructions on how to shrink the bracket can be found in [31].

Third step: sampling from the (untruncated) MVN prior In the final step, sampling the (untruncated) MVN prior, denoted as $\boldsymbol{\nu}$, from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{K})$ can be accomplished using highly efficient samplers. Since the covariance matrix, $\boldsymbol{K}$, is derived from a stationary covariance function defined on a regular grid, the process of sampling $\boldsymbol{\nu}$ from the prior MVN distribution $\mathcal{N}(\boldsymbol{0}, \boldsymbol{K})$ is equivalent to generating a stationary GP on a regular grid. In this scenario, different highly efficient approaches can be employed. For example, the efficient Fast Fourier Transform (FFT) algorithm, as introduced by [47], can be utilized. Furthermore, the authors of this paper have recently developed two approaches for sampling from very large Gaussian vectors derived from stationary GPs. The first approach is based on a large-scale Karhunen-Love Expansion (KLE), abbreviated as LS.KLE,

and is detailed in [29]. The second approach, known as the fast large-scale method, abbreviated as Fast.LS, is presented in [28].

From Equation (23), the logarithm of the *redefined likelihood* is given by

$$\log[L(\boldsymbol{\xi})] = -\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\xi}\|^2 + \eta(\xi_N - \xi_1) - \sum_{j=2}^{N}\left[\log\left(1 + e^{\eta(\xi_j - \xi_{j-1})}\right)\right]. \tag{27}$$

Hence, the computational complexity for evaluating the log-likelihood in (27) is $\mathcal{O}(nN)$, where n represents the number of samples and N denotes the dimension of the MVN prior $\boldsymbol{\xi}$. The logarithm of the *redefined likelihood* (27) can be expressed in matrix form as follows:

$$\log[L(\boldsymbol{\xi})] = -\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\xi}\|^2 - \sum_{i=1}^{m}\left[\log\left(1 + e^{-\eta \boldsymbol{a}_i^\top \boldsymbol{\xi}}\right)\right], \tag{28}$$

where $\boldsymbol{a}_i \in \mathbb{R}^N$ is the i^{th} row of the constraints matrix $\boldsymbol{A} \in \mathbb{R}^{m\times N}$ defined in (10), where $m = N-1$. In the numerical examples presented in the paper and based on Taylor's approximation, we absorb the sum in (28) into the logarithm to avoid computing the logarithm of $(1 + e^{-\eta \boldsymbol{a}_i^\top \boldsymbol{\xi}})$ for $m = N-1$ times. While there is little operational difference between the two approaches, our method offers significant computational benefits in terms of running time, prediction accuracy, and coverage, as we shall demonstrate below (Section 4.3). Furthermore, employing this technique makes the proposed approach less sensitive to the parameter η and, thus, to the mass-shifting phenomenon described in [49]. Similar to the methods introduced in the MUR technique [7, 29, 46], the proposed approach involves conducting sampling before conditioning rather than after. It offers potential advantages by maintaining the stationary property throughout the sampling process. Furthermore, it allows simulating the posterior distribution in cases of extreme constraints, a task that becomes impossible for the HMC sampler. Finally, through the proposed approach, there is no need to compute the inverse of an $n \times n$ matrix at each McMC iteration. This can be of a significant advantage, especially when the size of the data set n is high. Additionally, sampling the prior is independent of the noise variance parameter σ^2. This can be advantageous in some situations, which is not the case when sampling directly from the full conditional posterior distribution (15).

4.2 Boundedness constraints

In this section, the boundedness constraints is considered. From [26] and Model (5), Y^N verifies boundedness constraints (i.e., $Y^N(x) \in [l_b, u_b]$, for any $x \in \mathcal{D}$) *if and only if* $\boldsymbol{\xi} \in \mathcal{E}_b$, where

$$\mathcal{E}_b = \{z \in \mathbb{R}^N \ : \ l_b \le z_j \le u_b, \ j = 1, \ldots, N\}. \tag{29}$$

As a result, the constrained posterior pdf in (18) has the following form:

$$p(\boldsymbol{\xi}|\boldsymbol{y}, \boldsymbol{\xi} \in \mathcal{E}_b) \propto \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\xi}\|^2\right)\exp\left(-\frac{1}{2\tau^2}\boldsymbol{\xi}^\top \boldsymbol{K}^{-1}\boldsymbol{\xi}\right)\mathbf{1}_{\mathcal{E}_b}(\boldsymbol{\xi}), \quad \boldsymbol{\xi} \in \mathbb{R}^N. \tag{30}$$

The only difference from the monotonicity (nondecreasing) constraints developed in Section 4.1 lies in the indicator function in (22), where $\mathcal{E}_m$ is replaced by $\mathcal{E}_b$. Consequently, only the first step of the circulant embedding technique will be modified. However, we still use the scaled logistic sigmoid approximation

$$\mathbf{1}_{l_b \le \xi \le u_b} \approx \frac{1}{1 + e^{-\eta(\xi - l_b)} + e^{-\eta(u_b - \xi)} + e^{-\eta(u_b - l_b)}}$$

for large $\eta > 0$ to derive a smooth approximation called $\mathbb{J}_\eta(\cdot)$ for $\mathbf{1}_{\mathcal{E}_b}(\boldsymbol{\xi})$,

$$\mathbf{1}_{\mathcal{E}_b}(\boldsymbol{\xi}) \approx \mathbb{J}_\eta(\boldsymbol{\xi}) \quad = \quad \prod_{j=1}^{N} \frac{1}{1 + e^{-\eta(\xi_j - l_b)} + e^{-\eta(u_b - \xi_j)} + e^{-\eta(u_b - l_b)}} \tag{31}$$

where η plays the role of controlling the quality of the approximation.

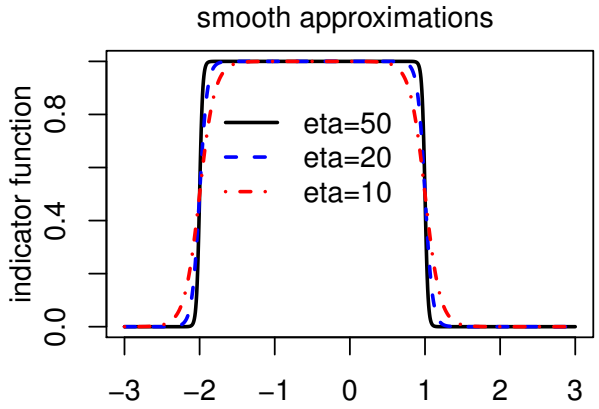

Figure 3: Smooth approximations of the indicator function $\mathbf{1}_{[-2,1]}(\cdot)$ for different values of η.

Figure 3 displays smooth approximations on the interval $[-3, 3]$ of the indicator function $\mathbf{1}_{[-2,1]}(\cdot)$ using Equation (31), with $l_b = -2$ and $u_b = 1$, for various values of η.

As for monotonicity constraints, by substituting the smooth approximate $\mathbb{J}_\eta(\cdot)$ of the indicator function $\mathbf{1}_{\mathcal{E}_b}(\boldsymbol{\xi})$ in (30), we obtain the following density approximation $\tilde{p}$ of p, where the *smooth relaxation* of the constraints $\mathbb{J}_\eta$ has been absorbed into the likelihood:

$$\begin{aligned} \tilde{p}(\boldsymbol{\xi}|\boldsymbol{y}, \boldsymbol{\xi} \in \mathcal{E}_b) \quad &\propto \quad \exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\xi}\|^2\right) \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\xi}^\top \boldsymbol{K}^{-1}\boldsymbol{\xi}\right) \mathbb{J}_\eta(\boldsymbol{\xi}) \\ &= \quad \underbrace{\left[\exp\left(-\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\xi}\|^2\right) \mathbb{J}_\eta(\boldsymbol{\xi})\right]}_{\textit{redefined likelihood}} \underbrace{\exp\left(-\frac{1}{2\tau^2}\boldsymbol{\xi}^\top \boldsymbol{K}^{-1}\boldsymbol{\xi}\right)}_{\text{(untruncated) MVN prior}}. \end{aligned} \tag{32}$$

According to (31), the logarithm of the *redefined likelihood* in (32) can be expressed as follows:

$$-\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\xi}\|^2 - \sum_{j=1}^{N} \log\left[1 + e^{-\eta(\xi_j - l_b)} + e^{-\eta(u_b - \xi_j)} + e^{-\eta(u_b - l_b)}\right]. \tag{33}$$

It is worth noting that the logarithm of the likelihood function (33) above can be expressed in matrix form as follows:

$$-\frac{1}{2\sigma^2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\xi}\|^2 - \sum_{i=1}^{m} \log\left[1 + e^{-\eta(\boldsymbol{a}_i^\top \boldsymbol{\xi} + b_i)}\right], \tag{34}$$

where $\boldsymbol{a}_i \in \mathbb{R}^N$ represents the i^{th} row of the constraints matrix $\boldsymbol{A} \in \mathbb{R}^{m \times N}$ and $b_i \in \mathbb{R}^m$ is the i^{th} component of the vector $\boldsymbol{b} \in \mathbb{R}^m$, which contains the lower and upper bounds constraints. In this boundedness case, we have $m = 2 \times N$, $\boldsymbol{A} = \begin{bmatrix} \text{diag}(N) \\ -\text{diag}(N) \end{bmatrix}$ and $\boldsymbol{b} = \begin{bmatrix} -\boldsymbol{l} \\ \boldsymbol{u} \end{bmatrix}$, with $\boldsymbol{l}$ and $\boldsymbol{u}$ the vectors representing the lower and upper bounds, respectively, and $\text{diag}(N)$ the N-dimensional identity matrix. This matrix form expression allows the proposed approach to handle boundedness constraints, where the lower and upper bounds can be any functions forming a convex set. For a further discussion on this case, we refer the reader to [27].

Remark 1. *For computational numerical advantages, the sum in* (34) *can be expressed as follows:*

$$\sum_{i=1}^{m} \log\left[1 + e^{-\eta(\boldsymbol{a}_i^\top \boldsymbol{\xi} + b_i)}\right] = \sum_{i=1}^{m} \log\left[1 + e^{-\eta(\boldsymbol{\xi}-\boldsymbol{l})_i} + e^{-\eta(\boldsymbol{u}-\boldsymbol{\xi})_i} + e^{-\eta(\boldsymbol{u}-\boldsymbol{l})_i}\right],$$

where, for example, $(\boldsymbol{\xi} - \boldsymbol{l})_i$ *represents the* i^{th} *element of the vector* $(\boldsymbol{\xi} - \boldsymbol{l})$*. This expression is applicable to boundedness constraints, where there is no need to compute the matrix product* $\boldsymbol{A\xi}$ *at each McMC iteration. This advantage will be highlighted in the numerical examples presented in this paper. Finally, if the constraints matrix* $\boldsymbol{A} \in \mathbb{R}^{m\times N}$ *is such that the diagonal elements are equal to* $\lambda_i \in \mathbb{R}$ *for* $i = 1, \ldots, m$ *and zero otherwise, then*

$$\mathbf{1}_{\{\boldsymbol{A\xi}+\boldsymbol{b}\geq \mathbf{0}_m\}} \approx \prod_{i=1}^{m} \frac{1}{1 + e^{-\eta(\boldsymbol{a}_i^\top \boldsymbol{\xi} + b_i)}} = \prod_{i=1}^{m} \frac{1}{1 + e^{-\eta(\lambda_i \xi_i + b_i)}}.$$

This is the case when we have either an upper or a lower bound constraint.

The following section presents empirical demonstrations of the computational complexity and predictive accuracy of the proposed approach in the context of shape-restricted function estimation.

4.3 Numerical performance

The aim of this section is to illustrate the performance of the proposed approach developed in Section 4 in terms of computational running time, prediction accuracy, and uncertainty quantification. Additionally, we provide a comparative analysis against the HMC sampler developed in [34] and the embedding technique introduced in [37, 49].

Example 1 (Monotonicity constraints)**.** *In this numerical example, we consider the monotone (nonincreasing) function*

$$f_{m_1}(x) = -2/[1 + \exp(-12x + 3)], \tag{35}$$

with $x \in [0, 1]$*. This is a challenging situation because* f_{m_1} *is monotone (nonincreasing) and remains nearly flat between 0.6 and 1.*

Figure 4 illustrates the monotone (nonincreasing) function estimation using the finite-dimensional GP approximation (5)*. We randomly generate* $n = 100$ *samples from* (1) *using the true nonincreasing function* f_{m_1} *(black solid curve) and a noise standard deviation of* $\sigma = 0.5$*. The covariates* $\{x_i\}$*,* $i = 1, \ldots, n$ *are generated using the* `lhsDesign` *function implemented in the* `R package` *DiceDesign [14]. The Matérn covariance function* (2) *with a smoothness parameter* $\nu = 1.5$ *has been employed. The length-scale parameter* ℓ *has been chosen such that the correlation at the maximum possible separation between the covariates equals 0.05. We adopt a joint posterior distribution based on the inverse Gamma distribution to estimate the variance noise parameter* σ^2 *and the signal variance parameter* τ^2 *at each McMC iteration. The blue dashed-dotted curve represents the mAP estimate (Definition 2), while the red dashed curve represents the MAP estimate (Definition 1). The black solid curve represents the true nondecreasing function, and the gray shaded area represents the posterior 95% confidence interval. The proposed* `LS-ESS` *approach developed in 4.1 has been employed in the top two panels with* $N = 27$ *(left) and* $N = 100$ *(right). To avoid the mass-shifting phenomenon [49] on the flat region* $[0.6, 1]$*, we fix the parameter* η *at* 100*, thereby ensuring a "good" approximation of* (23)*. In the Middle two panels, the FFT approach [47] is employed to general the prior MVN distribution in the last step of the algorithm developed in Section 4.1 with* $N = 27$ *(left) and* $N = 100$ *(right). This is the* `FFT-WC ESS` *approach developed in [37]. In the bottom two panels, the HMC sampler has been used to generate samples from the full conditional tMVN distribution* (15) *with* $N = 27$ *(left) and* $N = 100$ *(right). The computational time (in second) of*

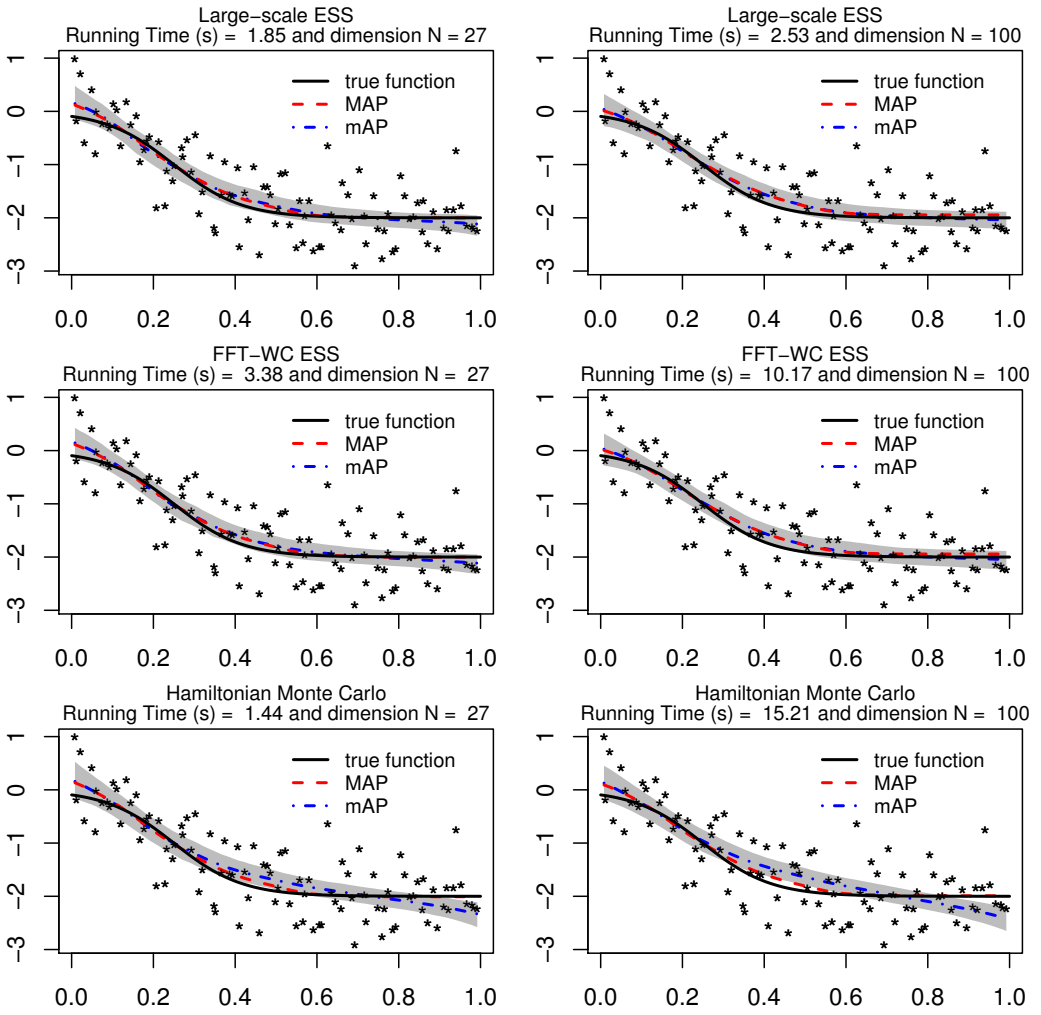

Figure 4: Monotone function estimation using model (5) with three different approaches. Top: the proposed LS-ESS approach with $N = 27$ (left) and $N = 100$ (right). Middle: the FFT-WC ESS approach with $N = 27$ and $N = 100$ (right). Bottom: the HMC sampler with $N = 27$ (left) and $N = 100$ (right). The computation time (in second) of running the McMC sampler for 6,000 iterations is displayed at the main part of each panel, where the first 1,000 iterations are discarded as burn-in.

running the McMC sampler for 6,000 iterations is displayed in the main part of each panel, where the first 1,000 iterations are discarded as burn-in. Under the same setting, when the dimension N is fixed at 100*, the proposed LS-ESS approach is six times (resp. four times) faster than the HMC sampler (resp. FFT-WC ESS approach). It is worth noting that when all the parameters ℓ, τ^2, and σ are fixed, the proposed LS-ESS approach can generate 6,000 monotonic trajectories in 11.05 second when the dimension N is fixed at 500, while the HMC approach takes 35.72 seconds. Unlike the circulant embedding technique approach we propose, when using the HMC sampler, the constrained posterior distribution fails to capture the flat region (when $x \in [0.6, 1]$). This is attributed to the mass-shifting phenomenon when using the full conditional tMVN distribution, as described in [49], which becomes more pronounced as the dimension N increases and as the posterior mode gets closer to boundary constraints. The methodology developed in the present paper tackles this problem through the parameter of approximation η. It is clear that the proposed LS-ESS approach visually outperforms the HMC sampler in terms of prediction accuracy and coverage. Unlike the HMC sampler, when employing the proposed approach, the mean of the constrained posterior distribution (mAP estimate) follows the MAP estimate. It is important to note that the performance*

of the HMC sampler, as well as that of the FFT-WC ESS approach, depends on the smoothness parameter ν of the covariance function. In this numerical example, the results were obtained using a smoothness parameter of $\nu = 1.5$. However, when $\nu < 1.5$, the computational running time of the HMC sampler is longest. Furthermore, the FFT performs poorly when ν is high, as demonstrated in [29]. The proposed LS-ESS approach does not suffer from this issue; it performs well in both cases, i.e., with high and low values of the smoothness parameter ν.

Example 2 (Boundedness constraints)**.** *In this example, we examine the behavior of the posterior distribution of the proposed LS-ESS approach as well as the HMC sampler using the nonnegative function introduced in [35]. The function is defined as follows:*

$$f(x) = \frac{1}{[1 + (10x)^4]} + \frac{1}{2} \exp\left[-100(x - 0.5)^2\right], \quad x \in [0, 1]. \tag{36}$$

This is a challenging situation because the underlying function f is bounded between 0 and 1, flat, and approximately zero on the interval $[0.7, 1]$.

Figure 5 depicts the posterior distribution behavior of the three approaches: LS-ESS (top panels), FFT-WC ESS (middle panels), and HMC sampler (bottom panels), using the model defined in equation (5). We randomly generate $n = 100$ samples from (1) using the true bounded function f defined in (36) and a noise standard deviation of $\sigma = 0.1$. The covariates $\{x_i\}$, $i = 1, \ldots, n$ are generated using the lhsDesign function implemented in the `R package` DiceDesign [14]. The Matérn covariance function (2) with a smoothness parameter $\nu = 1.5$ has been employed. The length-scale parameter ℓ has been chosen such that the correlation at the maximum possible separation between the covariates equals 0.05. The noise and signal variance parameters, σ^2 and τ^2 respectively, have been estimated at each McMC iteration. The blue dashed-dotted curve represents the mean a posterior (mAP) estimate (i.e., the mean of the posterior samples), while the red dashed curve represents the MAP estimate. The black solid curve represents the true bounded function, and the gray shaded area represents the posterior 95% confidence interval. The proposed LS-ESS approach has been employed in the top two panels with $N = 27$ (left) and $N = 100$ (right). The parameter η is fixed at 100, providing a "good" approximation of (32). The FFT approach to generate the prior MVN distribution in the last step of the algorithm developed in Section 4.1 is employed in the middle panels with $N = 27$ (left) and $N = 100$ (right). In the bottom two panels, the HMC sampler has been used with $N = 27$ (left) and $N = 100$ (right). The computational time (in second) of running the McMC sampler for 6,000 iterations is displayed in the main part of each panel, where the first 1,000 iterations are discarded as burn-in. Under the same setting, when the dimension N is fixed at 100, the proposed LS-ESS approach is six times (resp. three times) faster than the HMC sampler (resp. FFT-WC ESS approach). Let us mention that when all the parameters ℓ, τ^2, and σ are fixed, the proposed LS-ESS approach can generate 6,000 bounded trajectories in 5.49 seconds when the dimension N is fixed at 500, while the HMC approach takes 19.69 seconds. This significant reduction in computation time is attributed to both sampling from an efficient prior, Fast.LS, through ESS, and the considerations outlined in Remark 1. As for the monotonicity constraint case, the computational running time of the HMC sampler (resp. FFT-WC ESS approach) becomes slower with a lower (resp. higher) value of the smoothness parameter ν. It is worth noting that the posterior credible intervals exhibit greater compression when employing the circulant embedding technique developed in the present paper compared to sampling from the full conditional tMVN (15).

Run-time comparison (LS-ESS versus HMC) In this paragraph, we provide empirically illustration of the computational complexity of the proposed LS-ESS approach. The monotone (nonincreasing) and bounded synthetic functions f_{m_1} and f, defined in (35) and (36) respectively, are employed. In both cases, the covariates are generated uniformly on $[0, 1]$. The parameter η is

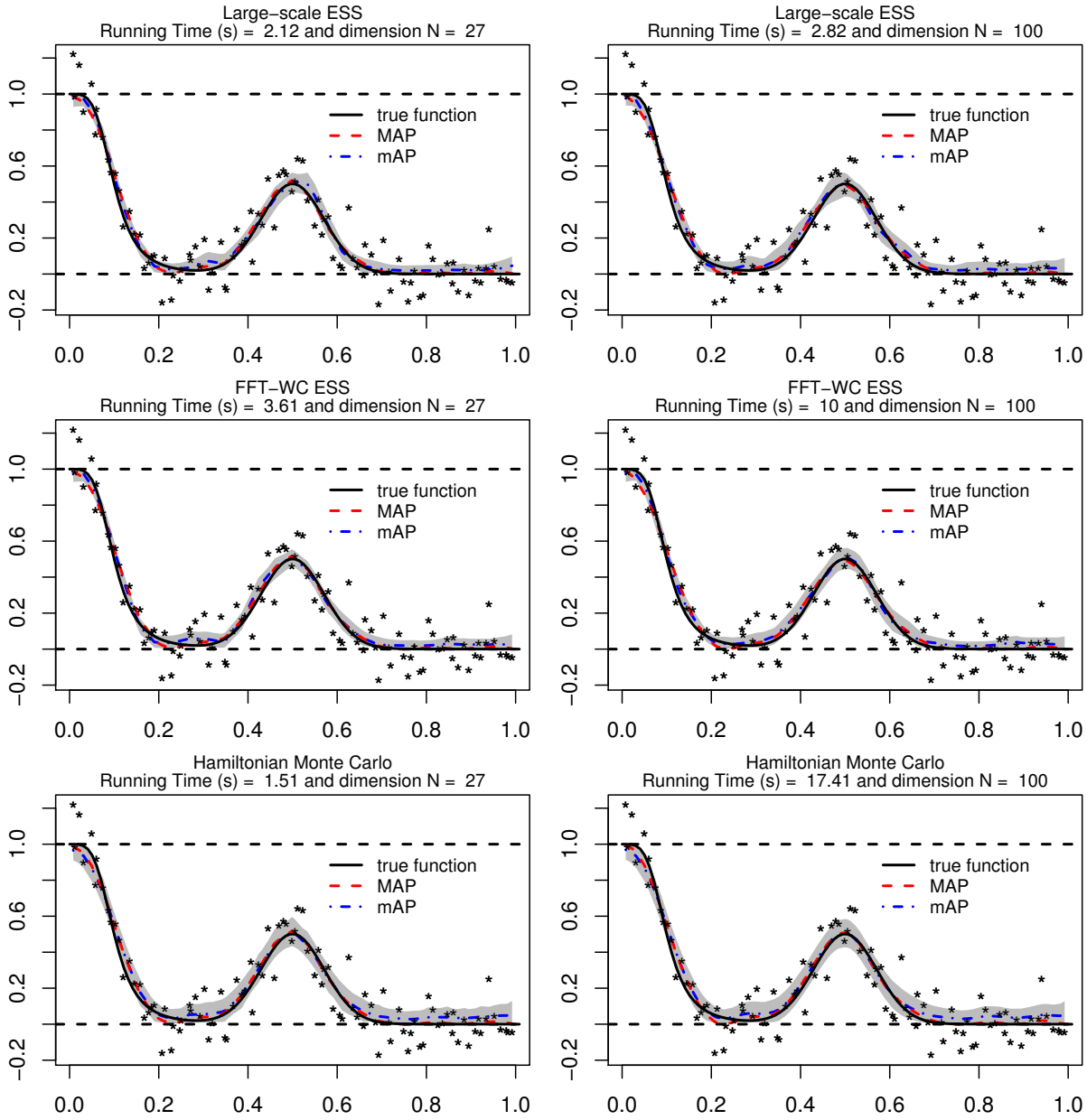

Figure 5: Same as Figure 4 for bounded function estimation.

fixed at 100 and the number of knots $N = \lceil \frac{n}{8} \rceil$, where n represents the number of samples. The Matérn covariance function (2) is employed, with the smoothness parameter ν fixed at 1.5, and the length-scale parameter ℓ chosen to achieve a correlation of 0.05 at the maximum possible separation between the covariates. The other parameters, σ^2 (noise variance) and τ^2 (signal variance), are updated using the joint posterior distribution, as described in [37]. The proposed LS-ESS approach developed in the present paper, is compared to the highly efficient HMC sampler developed in [34] and implemented in the `R package` *tmg*, in terms of computational running time per iteration.

Figure 6 displays the average running time per McMC iteration over ten replicated as a function of the sample size n for two McMC samplers: the LS-ESS approach proposed in this paper, represented by the black solid curve, and the HMC sampler developed in [34], represented by the black dashed curve. The left panel considers the monotonicity (nondecreasing) constraint, while the right panel employs the nonnegativity constraint. It is worth noting that the HMC sampler performs well in low dimensions, especially when $N \leq 125$ (i.e., when $n \leq 1,000$). However, as both the dimension N and the sample size n grow, the proposed LS-ESS demonstrates a clear advantage

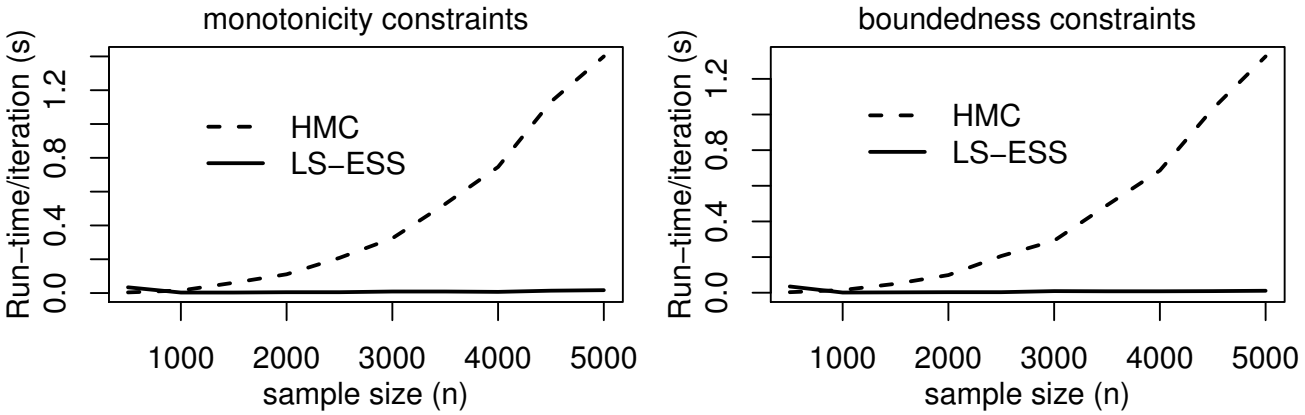

Figure 6: Average running time per McMC iteration (in seconds) over ten replicates plotted against the sample size n for two Gibbs samplers: the proposed approach denoted LS-ESS (black solid curve) and the HMC *tmg* (black dashed curve). The monotonicity (resp. boundedness) constraints is considered in the left (resp. right) panel. The dimension N is fixed at $\lceil n/8 \rceil$.

over the HMC sampler. Unlike the HMC sampler, the computational complexity of our approach increases linearly with both the dimension N and the sample size n.

Example 3 (Multiple constraints: monotonicity and boundedness)**.** *The extension of the proposed approach to sets of linear inequalities allows for the incorporation of multiple shape constraints through Model* (5)*. In this numerical example, monotonicity and boundedness constraints are jointly considered. To achieve this, we examine the logistic function* $f_{m_3}(x) = 1/[1 + \exp(-6x)]$*, where* $x \in \mathcal{D}$*. This function is monotone (increasing) and bounded from above by 1.*

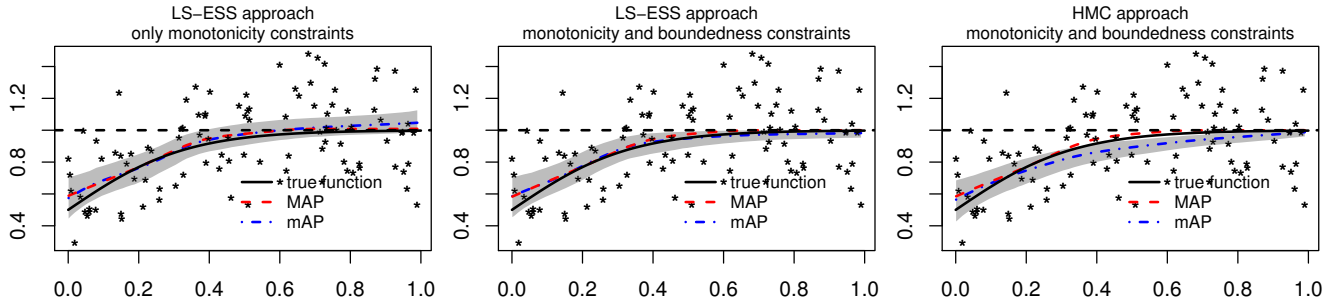

Figure 7: GP approximation from Model (5) using the proposed LS-ESS approach under only monotonicity constraints (left) and under both monotonicity and boundedness constraints (middle). The HMC approach is employed in the right panel under both constraints. The black dashed horizontal line represents the upper bound constraint.

Figure 7 presents a side-by-side comparison between the proposed LS-ESS approach and the HMC sampler using Model (5), with $N = 27$. The left panel illustrates the proposed approach under only monotonicity constraints, while the middle and right panels demonstrate the two competitor approaches under both monotonicity and boundedness constraints. The Matérn covariance function (2) is employed with a smoothness parameter $\nu = 2.5$. The length-scale parameter ℓ was chosen to achieve a correlation of 0.05 at the maximum possible separation between the covariates. The signal and noise variance parameters τ^2 and σ^2 are updated at each McMC iteration. The black stars represent the $n = 100$ synthetic training data generated from (1) using the function f_{m_3} and a noise standard deviation $\sigma = 0.2$. The black dashed horizontal line represents the upper bound constraint. In the left panel, the posterior estimates (MAP and mAP) and the posterior 95% credible interval

(gray shaded area) respect monotonicity constraints, but they violate the upper bound constraint. In the middle and right panels, the posterior estimates and the 95% credible interval adhere to both monotonicity and boundedness constraints. We visually observe that incorporating multiple shape constraints significantly enhances prediction accuracy and reduces credible intervals. Furthermore, the posterior distribution obtained through the proposed LS-ESS approach visually appears to be more flexible than that obtained through the HMC sampler. It is worth noting that when the dimension N is fixed at 100, the running time for generating 6,000 McMC iterations through the proposed approach under both monotonicity and boundedness constraints was 3.7 seconds, compared to 21.4 seconds for the HMC sampler. This significant reduction in computation time is attributed to both sampling from an efficient prior, Fast.LS, through ESS, and the considerations outlined in Remark 1.

Example 4 (Linear regression)**.** *The model employed in the present paper, referenced as Model* (5)*, generates piecewise linear trajectories, where the basis functions are the hat functions described in* (3)*. Consequently, when the number of knots N is fixed to two, $Y^N(\cdot)$ generates linear paths/functions. Hence, linear regression problems can be effectively addressed using the approach developed in this paper. Let us consider the following increasing linear function: $f(x) = 5x$, $x \in \mathcal{D}$.*

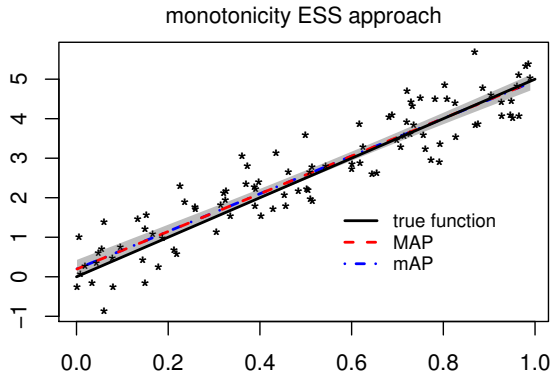

Figure 8: GP approximation from Model (5) using the proposed LS-ESS approach under monotonicity constraints. The black line represents the true monotone linear function $x \mapsto 5x$.

In Figure 8, the black line represents the target linear function $f(x) = 5x$, and the black stars represent the $n = 100$ observations obtained from (1) *using the function f and a noise standard deviation $\sigma = 0.5$. Model* (5) *is employed with $N = 2$. Additionally, the Matérn covariance function with a smoothness parameter $\nu = 2.5$ is used, with a length-scale parameter ℓ estimated through an adaptive cross-validation technique. The Gaussian vector $\boldsymbol{\xi} = [\xi_1, \xi_2]^\top$ is then generated using the LS-ESS approach developed in Section 4. The gray shaded area represents the 95% credible interval obtained from 6,000 McMC iterations, where the first 1,000 iterations are discarded as burn-in. The signal and noise variance parameters τ^2 and σ^2 are updated at each iteration. For example, the root-mean-squared-error (rmse) of fifty tested data is equal to 1.02×10^{-2} and 1.04×10^{-2} for MAP and mAP estimates, respectively.*

5 Higher dimensional input spaces

The finite-dimensional GP approximation defined in (5) can be extended to d-dimensional input spaces by tensorization (see [26] for a further discussion on imposing shape constraints for $d \geq 2$). Consider $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathcal{D}$ with input space $\mathcal{D} = [0,1]^d$, and a set of knots per dimension

$(u_1^1, \ldots, u_{N_1}^1), \ldots, (u_1^d, \ldots, u_{N_d}^d)$. Then, the proposed GP approximation can be expressed as

$$Y_{N_1,\ldots,N_d}(\boldsymbol{x}) = \sum_{j_1=1}^{N_1} \ldots \sum_{j_d=1}^{N_d} \xi_{j_1,\ldots,j_d} \left[\phi_{j_1}^1(x_1) \times \ldots \times \phi_{j_d}^d(x_d)\right], \tag{37}$$

where $\xi_{j_1,\ldots,j_d} = Y(u_{j_1}, \ldots, u_{j_d})$ and $\phi_{j_\zeta}^\zeta$ are basis functions as defined in (3). Conditionally on data and shape constraints, we get

$$Y_{N_1,\ldots,N_d}(\boldsymbol{x}_i) + \epsilon_i = y_i \quad \text{and} \quad \xi_{j_1,\ldots,j_d} \in \mathcal{E},$$

where $\boldsymbol{x}_i \in \mathcal{D}$, $y_i \in \mathbb{R}$ and $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$, with σ^2 the noise variance. As in (7), $\mathcal{E}$ represents a set of linear inequality constraints on the basis coefficient $\boldsymbol{\xi} = \xi_{j_1,\ldots,j_d}$. Then, the sampling approach developed in Section 4 can be employed to the zero-mean Gaussian vector $\boldsymbol{\xi}$ with a prior covariance matrix $\tau^2\boldsymbol{K}$. For instance, monotonicity constraints in two dimensions (i.e., $d = 2$) with respect to the two input variables x_1 and x_2 provide

$$\mathcal{E}_m = \{\boldsymbol{\xi} \in \mathbb{R}^{N_1 \times N_2} \ : \ \xi_{j-1,j'} \leq \xi_{j,j'} \text{ and } \xi_{j,j'-1} \leq \xi_{j,j'}\},$$

for $j \in \{2, \ldots, N_1\}$ for the first inequality and $j' \in \{2, \ldots, N_2\}$ for the second inequality.

As in one-dimensional cases, to avoid overfitting, it is more reasonable to choose the number of knots to be less than the number of training samples n. It is worth noting that having less knots than training samples can significantly impact the proposed McMC samplers, as they will then be performed in low-dimensional spaces, where $N = N_1 \times \ldots \times N_d \ll n$. Therefore, sampling from the posterior distribution can be achieved faster especially because of the incorporation of the data into the likelihood function (28), where there is not need to inverse a full-rank matrix of dimension $n \times n$. It is worth noting that when all elements of the vector $N = (N_1, \ldots, N_d)$ are equal to two (i.e., $N_\zeta = 2$, for any $\zeta \in 1, \ldots, d$), the proposed approach can be seen as a generalization of multidimensional linear regression problems.

5.1 Monotonicity in higher dimensions

The finite-dimensional GP approximation originally proposed in [26], was introduced with the aim of interpolating a given number of *noise-free* observations $\boldsymbol{y} = [y_1, \ldots, y_n]^\top$. It was strictly necessary to have more knots than training samples ($N > n$), a condition referred to as the *degree of freedom* in [26]. Moreover, it involves restricted domains where certain McMC samplers, performed poorly. Consequently, the extension to higher-dimensional input spaces was a challenge task. In this numerical example, we show that, by considering noisy observations, the proposed approach can be performed in higher dimensions. The relaxation of the set of constraints and the incorporation of the observations into the likelihood (28) allow this approach to deal with a large number of observations within a reasonable running time.

To do this, we consider the following 7D target function

$$f(\boldsymbol{x}) = 3/[1 + \exp(-10x_1 + 2.1)] + \arctan(5x_2) + \sum_{i=3}^{7}(i-1)x_i,$$

where $\boldsymbol{x} = (x_1, \ldots, x_7) \in [0, 1]^7$. This function is monotone non-decreasing with respect to all its input variables x_i, $i = 1, \ldots, 7$. While some computations could potentially be simplified, considering f as an additive function, we choose not to take advantage of this in order to test our approach on a fully 7D example. The target function f is evaluated on a Latin-hypercube design

at 5000 locations using lhsDesign function implemented in the `R package` *DiceDesign* [14]. The training data $\boldsymbol{y}$ are generated from (1) using the target function f and a noise standard deviation $\sigma = 0.1$. A tensor product of 1D Matérn covariance function (2) has been used with a smoothness parameter $\nu = 2.5$. We determine the number of knots as a balance between achieving high-quality resolution in the approximation and minimizing computational running time. Consequently, we set the number of knots per dimension to two, except for the first two dimensions (i.e., $N_\zeta = 2$, for $\zeta = 3, \ldots, 7$), owing to the linearity of the target function concerning the corresponding variables. For the first dimension, we choose $N_1 = 5$, and for the second dimension, we choose $N_2 = 4$, since the function f exhibits more variation along these dimensions compared to the others.

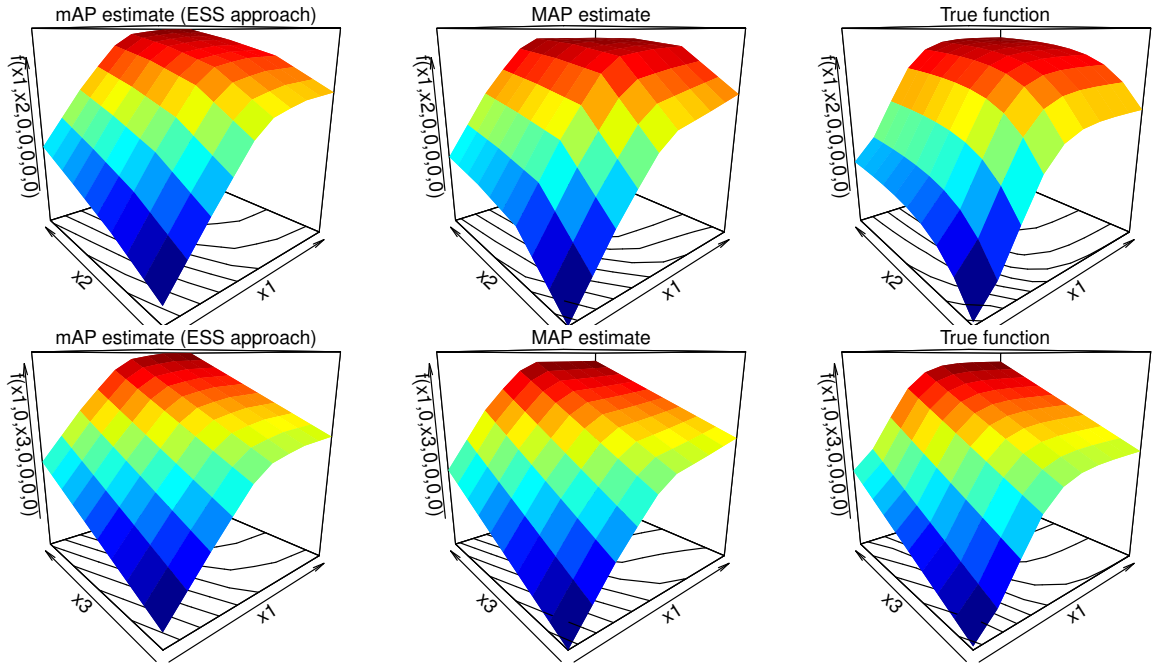

Figure 9: From left to right: the mAP estimate based on 6,000 McMC iterations, the MAP estimate based on the posterior mode (17), and the target function. All these functions are evaluated at $(x_1, x_2, 0, \ldots, 0)$ (top) and at $(x_1, 0, x_3, 0, \ldots, 0)$ (bottom), where $(x_1, x_2, x_3) \in [0, 1]^3$.

Figure 9 displays side by side the proposed MAP and mAP estimates together with the true function evaluated at $(x_1, x_2, 0, 0, 0, 0, 0)$. As in the one-dimensional cases, the mAP estimate is obtained through 6,000 McMC iterations of the algorithm developed in Section 4, with 1,000 iterations discarded as burn-in. While the MAP estimate is computed through the posterior mode (17). Firstly, it should be noted that the proposed approach (37) is capable of capturing the monotonicity dynamics of the target function across the entire domain. Secondly, generating 6,000 monotonic surfaces using the approach developed in Section 4 takes less than 10 minutes, which includes the estimation of signal and noise variance τ^2 and σ^2 respectively, at each iteration. However, computing the MAP estimate which is independent of the sampling procedure requires couple of seconds.

Now, in Figure 10 and Table 1, we investigate the quality of the prediction using the two proposed MAP and mAP estimates. Figure 10 shows side by side the predictive MAP and mAP estimates versus the true *noise-free* observations. Table 1 summarizes the regression error measurements for the two proposed MAP and mAP estimates. Based on the results from the one-dimensional cases and findings in [27], the MAP estimate consistently outperforms the mAP estimate across various criteria. For instance, the root-mean-squared-error (rmse) is equal to 7.56×10^{-2} when using the MAP estimate for prediction, while it is equal to 8.09×10^{-2} when using the mAP estimate. Additionally, the performance of the MAP estimate is further validated using the Q^2 criterion. This criterion is defined as $Q^2 = 1 - \text{SMSE}$, where SMSE represents the standardized mean squared error [45]. The Q^2 criterion equals one if the predictive mean matches the test data and is less

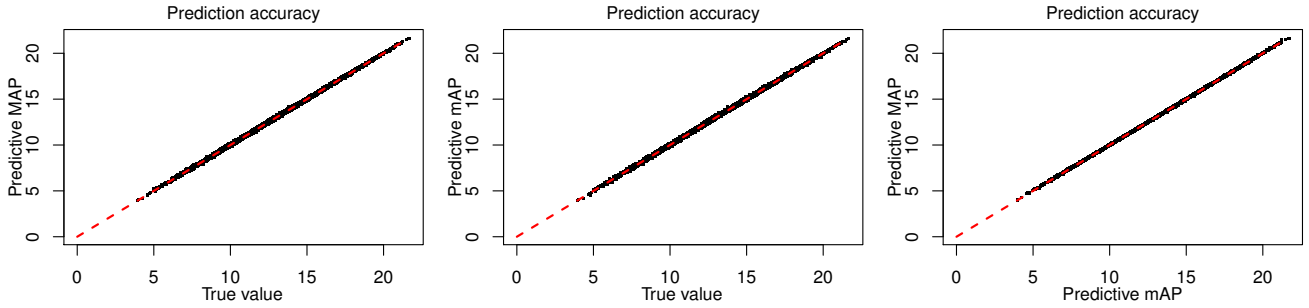

Figure 10: Side by side prediction accuracy: predictive MAP and mAP estimates versus true *noise-free* observations.

	mAP	MAP
MAE	6.04×10^{-2}	$\mathbf{5.83 \times 10^{-2}}$
MSE	0.65×10^{-2}	$\mathbf{0.57 \times 10^{-2}}$
RMSE	8.09×10^{-2}	$\mathbf{7.56 \times 10^{-2}}$
SMSE	7.91×10^{-4}	$\mathbf{6.90 \times 10^{-4}}$
Q^2	99.92×10^{-2}	$\mathbf{99.93 \times 10^{-2}}$
PVA	**0.90**	1.04
WAIC	-6.38×10^3	-6.38×10^3

Table 1: Comparison of regression error measures for the two proposed mAP and MAP estimates.

than one otherwise. For the MAP and mAP estimates, it is equal to 99.93 and 99.92, respectively. The computation time for the MAP estimate is significantly faster compared to the mAP estimate. Therefore, it may be considered an ideal and accurate estimate of the conditional posterior process. Finally, it is worth noting that the proposed sampling strategy developed in this paper improves the prediction of the mAP estimate and corrects the behavior of the posterior distribution.

5.2 Monotonicity and boundedness in 2D

The aim of this section is to investigate the performance of the proposed approach under multiple constraints. To achieve this, we consider the following target function

$$f(\boldsymbol{x}) = 1/[1 + \exp(-10x_1)] + x_2^3,$$

where $\boldsymbol{x} = (x_1, x_2) \in [0,1]^2$. This function is monotone non-decreasing with respect to x_1 and x_2 and bounded from above by the function $f_u(x_1, x_2) = 1/[1 + \exp(-10x_1)] + x_2^3 + 0.5$ (gray surface in Figure 11). The target function f is evaluated on a Latin-hypercube design at 1000 locations using lhsDesign function implemented in the `R package` *DiceDesign* [14]. The training data $\boldsymbol{y}$ are generated from (1) using the target function f and a noise standard deviation $\sigma = 0.1$. A tensor product of 1D Matérn covariance function (2) has been used with a smoothness parameter $\nu = 2.5$. We set the number of knots to $N_1 = 7$ for the first dimension and $N_2 = 5$ for the second dimension, resulting in 35 knots and basis functions. This is because the function f exhibits more variation along the first dimension compared to the second dimension.

In Figure 11, Model (37) is employed, where we display side by side the proposed MAP and mAP estimates along with the target function. The mAP estimate is obtained through 6,000 McMC iterations of the algorithm developed in Section 4, with 1,000 iterations discarded as burn-in. While the MAP estimate is computed through the posterior mode (17). With only 35 basis functions, the proposed approach (37) effectively captures the monotonicity dynamics of the target function across the entire domain while also adhering to boundedness constraints. Once again, the MAP

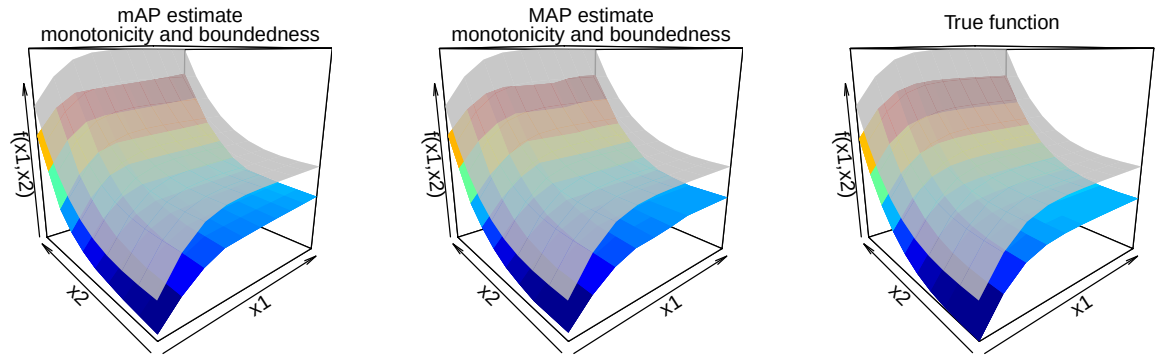

Figure 11: Prediction accuracy of model (37) under both monotonicity and boundedness constraints. From left to right: mAP estimate based on 6,000 McMC iterations, MAP estimate based on the mode (17), and the target function.

estimate outperforms the mAP estimate in terms of prediction accuracy, with an RMSE equal to 1.52×10^{-2} compared to 1.57×10^{-2}.

6 Real-world data application

In this section, we evaluate the performance of the approach developed in this paper through real-world data applications.

6.1 Age and income data

The dataset used comprises age (in years) and the logarithm of income (log.income) for 205 Canadian workers, sourced from a 1971 Canadian Census Public Use Tape. The objective is to estimate the logarithm of income as a function of age. The real-life data serves to highlight the superior prediction accuracy of the proposed LS-ESS approach developed in the present paper.

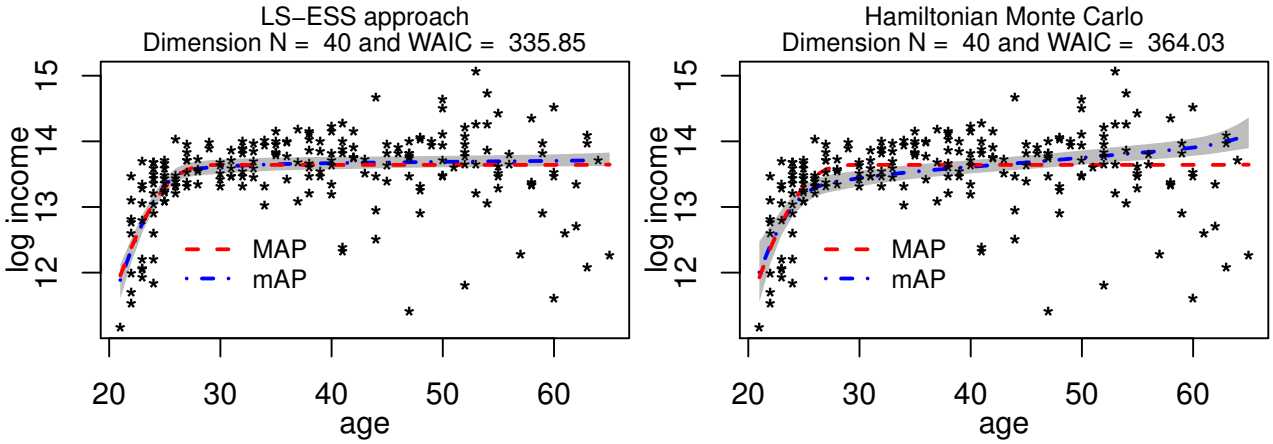

Figure 12: Estimation accuracy of the two competing methods applied on the age-income data. The black stars represent the training data. The blue dashed-dotted curve represents the mAP estimate, while the gray shaded area represents the 95% confidence interval. The red dashed curve represents the MAP estimate. The WAIC values corresponding to the two competing methods are displayed in the main of each panel based on 6,000 McMC posterior sample paths.

Figure 12 presents a computational comparison between the proposed large-scale approach (LS-ESS) and the HMC approach for age-income real-world data application. The black stars represent the $n = 205$ training samples. The GP approximation defined in (5) is employed, where N is fixed

at 40 to avoids overfitting [33]. The Matérn covariance function (2) is employed with a smoothness parameter ν fixed at 2.5. The length-scale parameter ℓ is estimated using an adaptive cross-validation technique based on the MAP as a predictive estimate. The noise variance parameter σ^2 and the signal variance parameter τ^2 are updated at each McMC iteration. In the left panel, the proposed LS-ESS approach is employed to generate 6,000 McMC sample paths from the constrained posterior distribution. In the right panel, the HMC sampler developed in [34] and implemented in the `R package` *tmg* is used. The blue dashed-dotted curve represents the mean of the posterior sample paths (mAP) estimate, while the red dashed curve represents the MAP estimate. The gray shaded area represent the 95% credible interval based on 6,000 McMC sample paths, where the first 1,000 iterations are discarded as burn-in. The two approaches fit the data well in general. However, unlike the HMC sampler, the proposed LS-ESS approach captures the flat region for age $\geq$ 26. For age $\leq$ 26, the HMC sampler fails to capture the trend, in contrast to the proposed LS-ESS approach. The Watanabe-Akaike information criterion (WAIC) value corresponding to the two methods is displayed in the main of each panel, and it is much smaller for the proposed approach. This indicates that the proposed LS-ESS approach fits the data better than the HMC sampler and it is more appropriate for fitting the age and income data. Furthermore, this confirms the results obtain in [49], where the authors demonstrated that the tMVN approach (i.e., sampling from the full conditional distribution (15)) is biased with a mass-shifting phenomenon appears in the marginal distribution. The WAIC value obtained using the proposed LS-ESS approach is close to that obtained using the efficient dependent global local prior DGL shrinkage approach developed in [49]. It is worth mentioning that the mAP estimate, when using the proposed LS-ESS approach, aligns with the MAP estimate, and the credible interval covers the MAP estimate. This means that the through the proposed LS-ESS approach, the behavior of the constrained posterior distribution is rectified.

6.2 Light Detection and Ranging Data

The Light Detection and Ranging (LiDAR) data consist of 221 observations from a LiDAR experiment, containing information on range and logratio. The predictor range represents the distance traveled before the light is reflected back to its source, while the response variable logratio is the logarithm of the ratio of received light from two laser sources. The data suggests that the true underlying function exhibits a monotone (nonincreasing) constraints with a flat region.

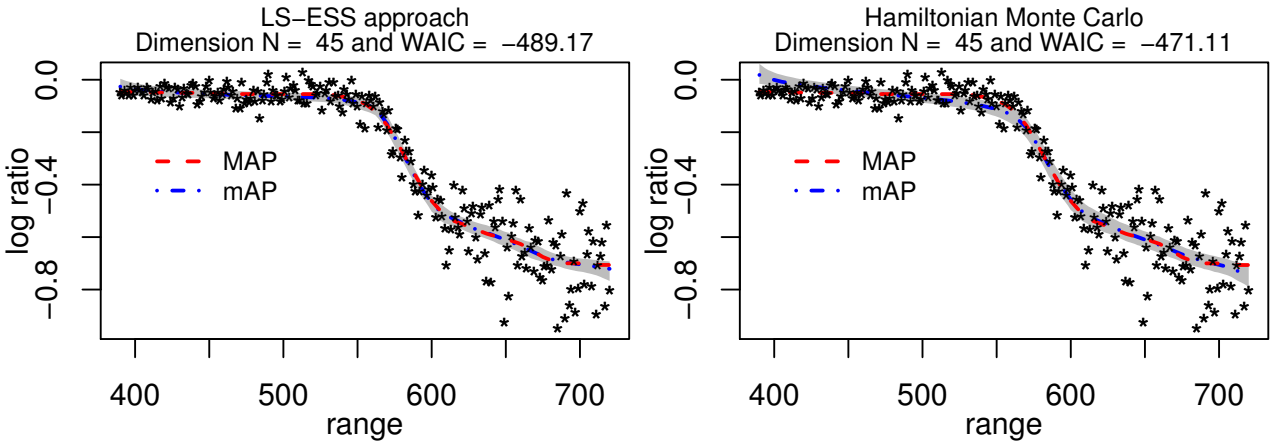

Figure 13: Same as Figure 12 for Light Detection and Ranging Data (LiDAR).

Figure 13 shows the estimation accuracy of the two competing methods applied on the LiDAR data. The GP approximation defined in (5) is employed, where N is fixed at 45 to avoids overfitting [33]. The Matérn covariance function (2) is employed with a smoothness parameter ν fixed at 2.5.

The length-scale parameter ℓ is estimated using an adaptive cross-validation technique. The noise variance parameter σ^2 and the signal variance parameter τ^2 are updated at each McMC iteration. The proposed LS-ESS approach is employed in the left panel, while the HMC sampler is applied in the right panel. The LS-ESS approach seems to fit the data better, especially over the region where the value of logratio starts to decrease. To confirm this result, we compute the overall WAIC value for the two competing approaches. The proposed LS-ESS approach obtains a smaller WAIC value of -489.17 again the value of -471.11 for the model with HMC sampler. This indicates that the proposed LS-ESS approach is more appropriate for fitting the LiDAR data. The WAIC value obtained using the proposed LS-ESS approach is approximately equal to the one obtained with the efficient shrinkage DGL approach developed in [49]. As for age and income data, the mAP estimate, when using the proposed LS-ESS approach, aligns with the MAP estimate, and the credible interval covers the MAP estimate.

7 Conclusion

In this paper, we revisit Bayesian shape-restricted function estimation. We consider the finite-dimensional Gaussian process approximation originally proposed in [26]. This method provides an equivalent formulation of shape constraints in terms of basis coefficients. Our contribution draws inspiration from the efficient approach developed in [37] for generating samples from the posterior distribution restricted to the positive orthant. First, we extend their method to manage sets of linear inequalities, enabling it to address more complex convex and non-convex constraints. Additionally, this adaptation simplifies its application to other models and basis functions. Second, we compare different highly efficient samplers for approximating both the posterior and prior distributions, including the Hamiltonian Monte Carlo and the Fast Fourier Transform methods. Third, we explore the extension of this approach to higher-dimensional input spaces. The effectiveness and accuracy of our proposed approach are demonstrated through studies using both synthetic and real data. The relaxation of the set of constraints allows the proposed approach to be flexible and capable of handling large datasets.

Acknowledgments

Part of this research was conducted with the support of the consortium in Applied Mathematics CIROQUO, gathering partners in technological and academia in the development of advanced methods for Computer Experiments. `https://doi.org/10.5281/zenodo.6581217`

References

[1] X. Bay, L. Grammont, and H. Maatouk. Generalization of the Kimeldorf-Wahba correspondence for constrained interpolation. *Electronic Journal of Statistics*, 10(1):1580–1595, 2016.

[2] Z. I. Botev. The normal law under linear restrictions: simulation and estimation via minimax tilting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):125–148, 2017.

[3] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

[4] B. Cai and D. B. Dunson. Bayesian multivariate isotonic regression splines: applications to carcinogenicity studies. *Journal of the American Statistical Association*, 102(480):1158–1171, 2007.

[5] M. Chataigner, A. Cousin, S. Crépey, M. Dixon, and D. Gueye. Beyond surrogate modeling: Learning the local volatility via shape constraints. *SIAM Journal on Financial Mathematics*, 12(3):SC58–SC69, 2021.

[6] H. A. Chipman, E. I. George, R. E. McCulloch, and T. S. Shively. mBART: Multidimensional Monotone BART. *Bayesian Analysis*, 17(2):515 – 544, 2022.

[7] Y. Cong, B. Chen, and M. Zhou. Fast simulation of hyperplane-truncated multivariate normal distributions. *Bayesian Analysis*, 12(4):1017 – 1037, 2017.

[8] S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White. MCMC methods for functions: modifying old algorithms to make them faster. *Statistical Science*, 28(3):424 – 446, 2013.

[9] A. Cousin, A. Deleplace, and A. Misko. Gaussian process regression for swaption cube construction under no-arbitrage constraints. *Risks*, 10(12):232, 2022.

[10] A. Cousin, H. Maatouk, and D. Rullière. Kriging of financial term-structures. *European Journal of Operational Research*, 255(2):631–648, 2016.

[11] H. Cramer and R. Leadbetter. *Stationary and related stochastic processes: sample function properties and their applications*. Wiley series in probability and mathematical statistics. Tracts on probability and statistics. Wiley, 1967.

[12] S. Crépey and M. F. Dixon. Gaussian process regression for derivative portfolio modeling and application to credit valuation adjustment computations. *Journal of Computational Finance*, 24(1), 2020.

[13] S. M. Curtis and S. K. Ghosh. A variable selection approach to monotonic regression with Bernstein polynomials. *Journal of Applied Statistics*, 38(5):961–976, 2011.

[14] D. Dupuy, C. Helbert, and J. Franco. DiceDesign and DiceEval: Two R Packages for Design and Analysis of Computer Experiments. *Journal of Statistical Software*, 65(11):138, 2015.

[15] G. Finocchio and J. Schmidt-Hieber. Posterior contraction for deep Gaussian process priors. *Journal of Machine Learning Research*, 24(66):1–49, 2023.

[16] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 3 edition, 2013.

[17] S. Golchi, D. R. Bingham, H. Chipman, and D. A. Campbell. Monotone emulation of computer experiments. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):370–392, 2015.

[18] D. Goldfarb and A. Idnani. A numerically stable dual method for solving strictly convex quadratic programs. *Mathematical Programming*, 27(1):1–33, 1983.

[19] L. Grammont, H. Maatouk, and X. Bay. Equivalent between constrained optimal smoothing and Bayesian estimation. working paper or preprint, March 2022.

[20] F. Huber, P.-C. Bürkner, D. Göddeke, and M. Schulte. Knowledge-based modeling of simulation behavior for Bayesian optimization. *Computational Mechanics*, 2024.

[21] A. L. Johnson and D. R. Jiang. Shape Constraints in Economics and Operations Research. *Statistical Science*, 33(4):527 – 546, 2018.

[22] G. Kimeldorf and G. Wahba. A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, pages 495–502, 1970.

[23] P. J. Lenk and T. Choi. Bayesian analysis of shape-restricted functions using Gaussian process priors. *Statistica Sinica*, pages 43–69, 2017.

[24] L. Lin and D. B. Dunson. Bayesian monotone regression using Gaussian process projection. *Biometrika*, 101(2):303–317, 2014.

[25] A. F. López-Lopera, F. Bachoc, N. Durrande, and O. Roustant. Finite-dimensional Gaussian approximation with linear inequality constraints. *SIAM/ASA Journal on Uncertainty Quantification*, 6(3):1224–1255, 2018.

[26] H. Maatouk and X. Bay. Gaussian process emulators for computer experiments with inequality constraints. *Mathematical Geosciences*, 49(5):557–582, 2017.

[27] H. Maatouk, D. Rullière, and X. Bay. Bayesian analysis of constrained Gaussian processes. working paper or preprint, April 2023.

[28] H. Maatouk, D. Rullière, and X. Bay. Large-scale constrained Gaussian processes for shape-restricted function estimation. working paper or preprint, December 2023.

[29] H. Maatouk, D. Rullière, and X. Bay. Sampling large hyperplane-truncated multivariate normal distributions. *Computational Statistics*, 2023.

[30] M. C. Meyer, A. J. Hackstadt, and J. A. Hoeting. Bayesian estimation and inference for generalised partial linear models using shape-restricted splines. *Journal of Nonparametric Statistics*, 23(4):867–884, 2011.

[31] I. Murray, R. Adams, and D. MacKay. Elliptical slice sampling. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 541–548. JMLR Workshop and Conference Proceedings, 2010.

[32] R. M. Neal. Regression and classification using Gaussian process priors. *Bayesian statistics 6*, pages 475–501, 1999.

[33] B. Neelon and D. B. Dunson. Bayesian isotonic regression and trend analysis. *Biometrics*, 60(2):398–406, 2004.

[34] A. Pakman and L. Paninski. Exact Hamiltonian Monte Carlo for truncated multivariate Gaussians. *Journal of Computational and Graphical Statistics*, 23(2):518–542, 2014.

[35] A. Pensoneault, X. Yang, and X. Zhu. Nonnegativity-enforced Gaussian process regression. *Theoretical and Applied Mechanics Letters*, 10(3):182–187, 2020.

[36] S. J Petit, J. Bect, and E. Vazquez. Relaxed Gaussian process interpolation: a goal-oriented approach to Bayesian optimization. *arXiv preprint arXiv:2206.03034*, 2022.

[37] P. Ray, D. Pati, and A. Bhattacharya. Efficient Bayesian shape-restricted function estimation with constrained Gaussian process priors. *Statistics and Computing*, 30(4):839–853, 2020.

[38] T. S. Shively, S. G. Walker, and P. Damien. Nonparametric function estimation subject to monotonicity, convexity and other shape constraints. *Journal of Econometrics*, 161(2):166–181, 2011.

[39] E. T. Spiller, R. L. Wolpert, P. Tierz, and T. G. Asher. The zero problem: Gaussian process emulators for range-constrained computer models. *SIAM/ASA Journal on Uncertainty Quantification*, 11(2):540–566, 2023.

[40] L. P. Swiler, M. Gulian, A. L. Frankel, C. Safta, and J. D. Jakeman. A survey of constrained Gaussian process regression: Approaches and implementation challenges. *Journal of Machine Learning for Modeling and Computing*, 1(2), 2020.

[41] J. Taylor and Y. Benjamini. RestrictedMVN: multivariate normal restricted by affine constraints. *R package version*, 1, 2016.

[42] T.-T. Tran, A. Fradi, and C. Samir. Learning, inference, and prediction on probability density functions with constrained Gaussian processes. *Information Sciences*, 642:119068, 2023.

[43] I. Ustyuzhaninov, I. Kazlauskaite, C. H. Ek, and N. Campbell. Monotonic Gaussian process flows. In *International Conference on Artificial Intelligence and Statistics*, pages 3057–3067. PMLR, 2020.

[44] X. Wang and J. O. Berger. Estimating shape constrained functions using Gaussian processes. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):1–25, 2016.

[45] C. K. Williams and C. E. Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

[46] J. T. Wilson, V. Borovitskiy, A. Terenin, P. Mostowsky, and M. P. Deisenroth. Pathwise conditioning of Gaussian processes. *Journal of Machine Learning Research*, 22(105):1–47, 2021.

[47] A. TA Wood and G. Chan. Simulation of stationary Gaussian processes in $[0,1]^d$. *Journal of Computational and Graphical Statistics*, 3(4):409–432, 1994.

[48] S. Zhou, P. Giulani, J. Piekarewicz, A. Bhattacharya, and D. Pati. Reexamining the proton-radius problem using constrained Gaussian processes. *Physical Review C*, 99:055202, May 2019.

[49] S. Zhou, P. Ray, D. Pati, and A. Bhattacharya. A mass-shifting phenomenon of truncated multivariate normal priors. *Journal of the American Statistical Association*, 0(ja):1–37, 2022.