Bayesian analysis of constrained Gaussian processes

Hassan Maatouk, Didier Rullière, Xavier Bay

Bayesian analysis of constrained Gaussian processes

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

(†) CY Tech, CY Cergy Paris University, Laboratoire AGM, Site du Parc, 95011 Cergy-Pontoise, France
(‡) Mines Saint-Étienne, Univ Clermont Auvergne, CNRS, UMR 6158 LIMOS, Institut Henri Fayol, F-42023 Saint-Étienne France

Abstract Due to their flexibility Gaussian processes are a well-known Bayesian framework for nonparametric function estimation. Integrating inequality constraints, such as monotonicity, convexity, and boundedness, into Gaussian process models significantly improves prediction accuracy and yields more realistic credible intervals in various real-world data applications. The Gaussian process approximation, originally proposed in [22] is considered. It satisfies interpolation conditions and handles a wide range of inequality constraints everywhere. Our contribution in this paper is threefold. First, we extend this approach to handle noisy observations and multiple, more general convex and non-convex constraints. Second, we propose new basis functions in order to extend the smoothness of sample paths to differentiability of class C^p, for any $p \geq 1$. Third, we examine its behavior in specific scenarios such as monotonicity with flat regions and boundedness near lower and/or upper bounds. In that case, we show that, unlike the Maximum a posteriori (MAP) estimate, the mean a posteriori (mAP) estimate fails to capture flat regions. To address this issue, we propose incorporating multiple constraints, such as monotonicity with bounded slope constraints. According to the theoretical convergence and based on a variety of numerical experiments, the MAP estimate behaves well and outperforms the mAP estimate in terms of prediction accuracy. The performance of the proposed approach is confirmed through real-world data studies.

Keywords Gaussian processes; multiple constraints; convex and non-convex constraints; flat region; MAP estimate; HMC sampler.

1 Introduction

Gaussian processes (GPs) are a well-known nonparametric Bayesian framework for function estimation. They are widely used in many fields, such as computer science, physics, biology, engineering, and finance [33]. GP models are based on defining a prior distribution over function spaces. In general, a GP is characterized by its mean and covariance functions. The flexibility of GPs is attributed to their covariance function, which enables incorporating prior information, such as smoothness, stationarity, sparsity, and derivative constraints [10, 33].

Unconstrained GP models perform poorly in terms of prediction accuracy and yield unrealistic confidence intervals when applied to physical systems that satisfy inequality constraints such as monotonicity, boundedness, and convexity, see for example, [13, 18, 20, 34, 41]. Several real-world

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

cases where the data suggest that the underlying function satisfies specific inequality constraints are presented in physics [44] and econometrics [5, 8, 9, 11].

Including inequality constraints into a GP model improves its prediction accuracy and provides more realistic confidence intervals [6, 13, 20, 19, 35, 40, 45]. Recently, the authors in [38] provide an overview and survey of various strategies for incorporating shape constraints into a GP. In the present paper, the GP approximation proposed in [22] is considered, where various inequality constraints such as monotonicity, convexity, and boundedness are satisfied everywhere. To the best of our knowledge, it is the only model in the literature capable of dealing with a variety of shape constraints (either alone, together, or sequentially). The main idea is to approximate the samples of the parent GP by representing them in a finite-dimensional space of functions using an appropriate basis expansion. These basis functions possess attractive properties not necessary shared by other basis such as Bernstein polynomials [12], regression splines [4, 27], and restricted splines [37]. Various restrictions like monotonicity, convexity, and boundedness are *equivalently* translated into linear inequality constraints on the basis coefficients. The performance of this approach has been demonstrated through several real-world data applications [8, 9, 26, 42, 44]. The asymptotic properties have been investigated in [1, 15]. The generalization of the well-known Kimeldorf-Wahba correspondence [17] between Bayesian estimation on stochastic processes and splines for the constrained cases has been established.

In the present paper, our contributions are threefold. First, we extend this approach to handle both noisy observations and multiple and more general convex and non-convex constraints (such as boundedness within a non-convex set). Second, we propose new basis functions in order to extend the smoothness of the sample paths to differentiability of class C^p, for any $p \geq 1$. Third, the behavior of this approach is investigated in challenging situations, such as monotonicity with flat regions or boundedness where the underlying function is close to the lower or upper bounds. In that case and based on both theoretical and numerical results, the *Maximum a posteriori* (MAP) estimate behaves well and outperforms the mean a posterior (mAP) estimate (i.e., the mean of the posterior distribution) in terms of prediction accuracy. This is because the posterior distribution approximated by the efficient Hamiltonian Monte Carlo (HMC) sampler fails to capture the flat regions. To address this issue, we propose adding multiple constraints, such as monotonicity with bounded slope constraints. This leads to correction of the posterior distribution's behavior and the convergence of the mAP estimate towards the MAP estimate.

This article is structured as follows. In Sect. 2, GP regression is briefly reviewed. In Sect. 3, following the finite-dimensional GP approximation from [22], we propose a general formulation for linear inequality constraints that is capable of handling both convex and non-convex constraints. Section 4 presents the new basis functions in order to generalize the smoothness of sample paths to differentiability of class C^p, $p \geq 1$. Additionally, the asymptotic properties of the MAP estimate are investigated. Section 5 demonstrates the efficiency of the proposed framework through applications using real-world data.

2 Gaussian process regression review

A GP, namely $(Z(\boldsymbol{x}))_{\boldsymbol{x}\in\mathbb{R}^d}$, is characterized by its mean function μ and covariance function k, i.e., $Z \sim \mathcal{GP}(\mu, k)$ [33]. It can be written as follows:

$$Z(\boldsymbol{x}) = \mu(\boldsymbol{x}) + Y(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in \mathbb{R}^d,$$

where $(Y(\boldsymbol{x}))$ is a zero-mean GP with covariance function k, i.e., $Y \sim \mathcal{GP}(0, k)$, with

$$k(\boldsymbol{x}, \boldsymbol{x}') = \mathrm{Cov}(Y(\boldsymbol{x}), Y(\boldsymbol{x}')) = \mathrm{E}[Y(\boldsymbol{x})Y(\boldsymbol{x}')], \quad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d.$$

Given a dataset of size n, namely, $D = \{(\boldsymbol{x}_i, y_i), i = 1, \ldots, n\}$, where $\boldsymbol{x}_i$ denotes an input vector of dimension d and y_i denotes a scalar output. The input vectors $\{\boldsymbol{x}_i\}$ form the $n \times d$ design matrix $\mathbb{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]^\top$ and the outputs $\{y_i\}$ form the output vector $\boldsymbol{y} = [y_1, \ldots, y_n]^\top$ called data. Thus, the dataset can be written as $D = \{(\mathbb{X}, \boldsymbol{y})\}$. The following regression problem is considered

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

where f is an unknown latent function that generates the data, and ϵ_i is an additive independent identically distributed (i.i.d.) zero-mean Gaussian noise with constant variance σ^2_{noise}. A GP prior distribution on the underlying function f is assumed. Conditionally on $\boldsymbol{y} = [y_1, \ldots, y_n]^\top$, the conditional process remains a GP [33]

$$\{Y(\cdot)|\boldsymbol{y}\} \sim \mathcal{GP}\left(\mu_c(\cdot), c(\cdot, \cdot)\right), \tag{2}$$

where the conditional mean function μ_c and covariance function c are given as follows:

$$\begin{aligned} \mu_c(\boldsymbol{x}) &= \mathrm{E}\left[Y(\boldsymbol{x})|\boldsymbol{y}\right] = k(\boldsymbol{x}, \mathbb{X})^\top \left(k(\mathbb{X}, \mathbb{X}) + \sigma^2_{\text{noise}}\mathbf{I}_n\right)^{-1} \boldsymbol{y}; \\ c(\boldsymbol{x}, \boldsymbol{x}') &= k(\boldsymbol{x}, \boldsymbol{x}') - k(\boldsymbol{x}, \mathbb{X})^\top (k(\mathbb{X}, \mathbb{X}) + \sigma^2_{\text{noise}}\mathbf{I}_n)^{-1} k(\boldsymbol{x}', \mathbb{X}); \end{aligned} \tag{3}$$

with $\mathbf{I}_n$ the $n \times n$ identity matrix. Let us recall that $k(\mathbb{X}, \mathbb{X})$ is the covariance matrix of $Y(\mathbb{X})$ of dimension $n \times n$, and $k(\boldsymbol{x}, \mathbb{X}) = [k(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, k(\boldsymbol{x}, \boldsymbol{x}_n)]^\top$ is the vector of covariance between $Y(\boldsymbol{x})$ and $Y(\mathbb{X})$ of dimension n.

In the simple special case where data are noise-free [36], that is when we know $\{(\boldsymbol{x}_i, f_i)|i = 1, \ldots, n\}$, with $f_i = f(\boldsymbol{x}_i)$, the equations for GPR prediction (3) remain the same, but we have to replace the noise variance σ^2_{noise} by zero and the data vector $\boldsymbol{y}$ by $\boldsymbol{f}$, where $\boldsymbol{f} = [f_1, \ldots, f_n]^\top$.

Table 1: Some popular covariance functions with their degree of smoothness [33].

Name	Expression	Class
Squared Exponential	$\exp\left(-\frac{(x-x')^2}{2\theta^2}\right)$	C^∞
Matérn $\nu = 5/2$	$\left(1 + \frac{\sqrt{5}\lvert x-x'\rvert}{\theta} + \frac{5(x-x')^2}{3\theta^2}\right)\exp\left(-\frac{\sqrt{5}\lvert x-x'\rvert}{\theta}\right)$	C^2
Matérn $\nu = 3/2$	$\left(1 + \frac{\sqrt{3}\lvert x-x'\rvert}{\theta}\right)\exp\left(-\frac{\sqrt{3}\lvert x-x'\rvert}{\theta}\right)$	C^1
Exponential	$\exp\left(-\frac{\lvert x-x'\rvert}{\theta}\right)$	C^0

Table 1 displays some commonly used covariance functions in the machine learning community [33], ordered by decreasing degree of smoothness in the one-dimensional case, where θ is the correlation length parameter. In the present paper, we focus on the Matérn covariance function with a smoothness parameter of $\nu = 5/2$, as recommended in [33]. The exponential kernel is a Matérn covariance function with $\nu = 1/2$. The squared exponential (SE), on the other hand, is a Matérn covariance function with $\nu \to +\infty$.

Figure 1 shows three covariance functions in the left panel and their corresponding GP sample paths in the right panel.

3 Constrained Gaussian processes

3.1 C^0 approximation with Model (M_h)

In this section, the finite-dimensional approximation of GPs proposed in [22] is considered. Without loss of generality, let Y be a zero-mean GP with covariance function k, i.e., $Y \sim \mathcal{GP}(0, k)$. We

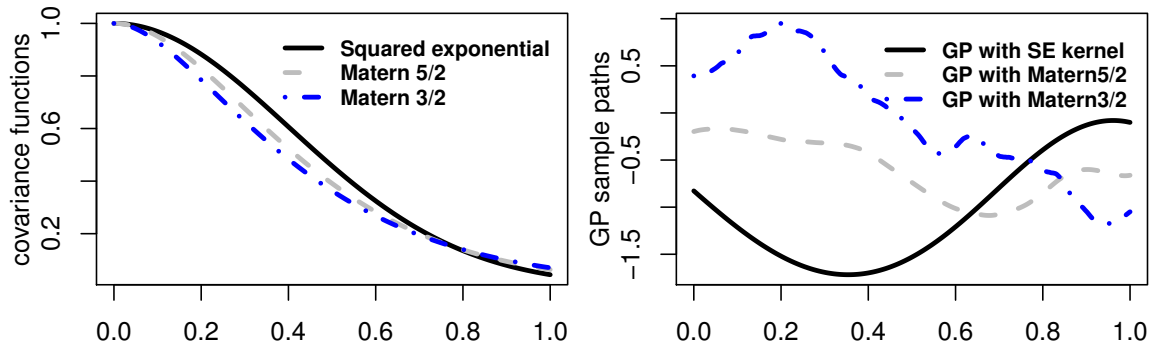


Figure 1: The left panel displays some covariance functions, while the right panel shows the corresponding GP sample paths. The correlation length parameter θ is fixed at 0.4.

first focus on the one-dimensional input case. The methodology is later extended to handle multi-dimensional input spaces (refer to Sect. 3.4 below). Let $\mathcal{D}$ be a compact set in $\mathbb{R}$. For simplicity, we suppose that $\mathcal{D}$ is the unit interval $[0,1]$. Let $t_j = (j-1)\Delta_N$, $j \in \{1,\ldots,N\}$ be a sequence of N equally spaced knots on $\mathcal{D}$, with a spacing of $\Delta_N = 1/(N-1)$. Let us mention that the methodology developed in this paper is applicable for non-uniform discretization of $\mathcal{D}$ (see the left panel of Figure 2). Let us define the three basis functions proposed in [22], which will be used in three different models, M_h, M_ϕ, and M_φ, in the present paper. These functions are given by

$$h_j(x) := h\left(\frac{x-t_j}{\Delta_N}\right), \quad \phi_j(x) := \int_0^x h_j(t)\mathrm{d}t, \quad \varphi_j(x) := \int_0^x \int_0^t h_j(u)\mathrm{d}u\mathrm{d}t, \tag{4}$$

for $j \in \{1,\ldots,N\}$, where $h(x) := (1-|x|)\mathbf{1}_{[-1,1]}(x)$ is the *hat* function on $[-1,1]$. The *hat* functions $\{h_j\}$ admit two nice properties. First, the value of any *hat* function at any knot is equal to Kronecker's delta function (i.e., $h_j(t_l) = \delta_{j,l}$), where $\delta_{j,l}$ is equal to one when $j = l$ and zero otherwise. Second, for any $x \in \mathcal{D}$, we have $\sum_{j=1}^N h_j(x) = 1$. The second property is used in the proof of Lemma 1. As mentioned in [22], 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 $\{t_j\}$,

$$\tilde{f}_N(x) = \sum_{j=1}^N f(t_j)h_j(x), \quad \forall x \in \mathcal{D}.$$

Let us recall the following well known result.

Lemma 1 (Uniform convergence C^0)**.** *Let f be a continuous function on $\mathcal{D}$, then, the piecewise linear interpolating function $\tilde{f}_N(\cdot) = \sum_{j=1}^N f(t_j)h_j(\cdot)$ converges uniformly to f when N tends to infinity.*

Proof. Indeed, for any $x \in \mathcal{D}$, we have $\sum_{j=1}^N h_j(x) = 1$ and

$$\begin{aligned}\left|\tilde{f}_N(x) - f(x)\right| &= \left|\sum_{j=1}^N \left(f(t_j) - f(x)\right) h_j(x)\right| \\ &\leq \sup_{|x-x'|\leq\Delta_N} |f(x') - f(x)| \sum_{j=1}^N h_j(x) = \sup_{|x-x'|\leq\Delta_N} |f(x') - f(x)|\,. \end{aligned} \tag{5}$$

By the uniform continuity of the function f on the compact interval $\mathcal{D}$ and the last inequality (5), we deduce that

$$\sup_{x\in\mathcal{D}} \left|\tilde{f}_N(x) - f(x)\right| \underset{N\to+\infty}{\longrightarrow} 0.$$

This concludes the proof of the lemma. □

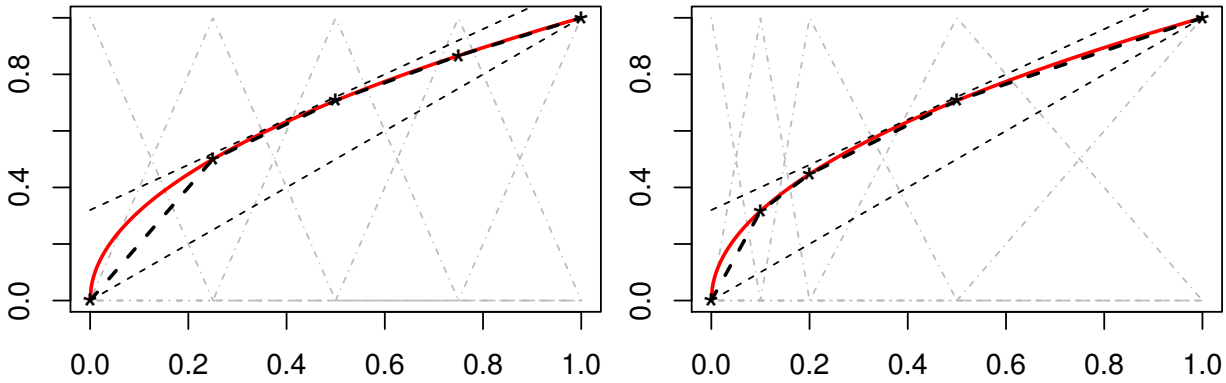


Figure 2: An example of approximating a monotone (nondecreasing) and bounded function f (red solid curve) using a piecewise linear interpolating function $\tilde{f}_N$ (black dashed curve). A uniform (resp. non-uniform) subdivision is used with $N = 5$ *hat* functions and knots in the left (resp. right) panel. The gray triangles represent the *hat* functions, while the black dashed thin lines denote the lower and upper bounds constraints.

Figure 2 shows the deterministic function $f(x) = \sqrt{x}$ (red solid curve) which verifies the monotonicity (nondecreasing) and boundedness constraints. This function is approximated by the piecewise linear interpolating function $\tilde{f}_N$ (black dashed curve) using either a uniform subdivision with $N = 5$ *hat* basis functions (left panel), or a non-uniform subdivision (right panel). The black dashed thin lines represent the lower and upper bound constraints, while the gray triangles represent the *hat* basis functions $\{h_j\}$. The black stars represent the values of the true and approximated functions at the knots $\{t_j\}$. Let us mention that the true function grows rapidly on $[0, 0.3]$. As a result, a finer discretization with $N = 3$ was used only for this interval, while $N = 2$ was used for the interval $[0.3, 1]$. This shows that a suitable subdivision can improve the accuracy of the approximation and reduce the number of knots. This also reduces the complexity of the sampling process when using an efficient HMC sampler to approximate the posterior distribution, as described in detail in Sect. 3.2.

If no additional smoothness assumptions are required, the first model can be written as follows:

$$Y^N(x) := \sum_{j=1}^{N} Y(t_j) h_j(x) = \sum_{j=1}^{N} \xi_j h_j(x), \quad x \in \mathcal{D}, \tag{M_h}$$

where we denote $\xi_j = Y(t_j)$. Since Y is a zero-mean GP with covariance function k, then, the vector $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_N]^\top$ is also zero-mean and Gaussian with covariance matrix $\boldsymbol{\Gamma}$, i.e., $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_N, \boldsymbol{\Gamma})$, where

$$\boldsymbol{\Gamma}_{j,l} = \mathrm{Cov}(Y(t_j), Y(t_l)) = k(t_j, t_l), \quad j, l \in \{1, \ldots, N\}, \tag{6}$$

and $\mathbf{0}_N = [0, \ldots, 0]^\top$ is the N-dimensional zero vector. Furthermore, the coefficients $\{\xi_j\}$ can be interpreted as the values of the original GP $(Y(x))$ evaluated at the knots $\{t_j\}$. Let $\mathcal{C}$ be the convex set of functions that verify some inequality constraints, such as monotonicity, convexity, and boundedness. The non-convex case will be investigated later in this section. For instance,

$$\mathcal{C} = \begin{cases} \mathcal{C}_b := \{f \in C^0(\mathcal{D}, \mathbb{R}) \text{ s.t. } \ell \leq f(x) \leq u, \ \forall x \in \mathcal{D}\} \\ \mathcal{C}_m := \{f \in C^0(\mathcal{D}, \mathbb{R}) \text{ s.t. } f(x) \leq f(y), \ \forall x \leq y \in \mathcal{D}\} \\ \mathcal{C}_c := \left\{f \in C^0(\mathcal{D}, \mathbb{R}) \text{ s.t. } \frac{f(y)-f(x)}{y-x} \leq \frac{f(z)-f(y)}{z-y}, \ \forall x \leq y \leq z \in \mathcal{D}\right\} \end{cases} \tag{7}$$

which corresponds to boundedness, monotonicity, and convexity constraints respectively, where the constants ℓ and u represent the lower and upper bounds, respectively, and where $C^0(\mathcal{D}, \mathbb{R})$ is the

set of continuous functions from $\mathcal{D}$ to $\mathbb{R}$. Our aim is to compute the posterior distribution of Y^N such that $Y^N \in \mathcal{C}$. The authors in [22] have shown the advantage of using the *hat* function as in Model (M_h) and more generally the basis functions defined in (4). They demonstrated that satisfying an infinite number of inequality constraints on the process $Y^N \in \mathcal{C}$ is *equivalent* to satisfying a finite number of linear inequality constraints on the coefficient vector $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_N]^\top$. To be more precise, for many various choices of $\mathcal{C}$, we have

$$Y^N \in \mathcal{C} \quad \Leftrightarrow \quad \boldsymbol{\xi} \in \mathcal{E}, \tag{8}$$

where $\mathcal{E}$ is a convex set of $\mathbb{R}^N$. For the inequality constraints given in (7), we obtain

$$\mathcal{E} = \begin{cases} \mathcal{E}_b := \left\{z \in \mathbb{R}^N : \ \ell \leq z_j \leq u, \ \forall j = 1, \ldots, N\right\} \\ \mathcal{E}_m := \left\{z \in \mathbb{R}^N : \ z_{j-1} \leq z_j, \ \forall j = 2, \ldots, N\right\} \\ \mathcal{E}_c := \left\{z \in \mathbb{R}^N : \ \dfrac{z_{j-1} - z_{j-2}}{t_{j-1} - t_{j-2}} \leq \dfrac{z_j - z_{j-1}}{t_j - t_{j-1}}, \ \forall j = 3, \ldots, N\right\} \end{cases} \tag{9}$$

which corresponds to boundedness, monotonicity, and convexity constraints respectively. The ability of these bases $\{h_j\}$ to express different constraints *equivalently* as linear restrictions on the vector of coefficients $\boldsymbol{\xi}$ is a desirable feature that may not be present in other bases such as Bernstein polynomials [12], regression splines [4, 27], and restricted splines [37]. In the following sections, we also demonstrate the attractiveness of the other basis functions defined in (4). Furthermore, new and smoother basis functions are given below (cf., Sect. 4). In the present section, we focus exclusively on Model (M_h) using the *hat* functions $\{h_j\}$. Note that the linear restrictions on the coefficients vector $\boldsymbol{\xi}$ in Eq. (9) can be expressed in a matrix form as follows:

$$\boldsymbol{\xi} \in \mathcal{E} \quad \Leftrightarrow \quad \boldsymbol{\xi} \in \mathbb{R}^N \text{ s.t. } \boldsymbol{\ell} \leq \boldsymbol{\Lambda}\boldsymbol{\xi} \leq \boldsymbol{u}, \tag{10}$$

where $\boldsymbol{\Lambda} \in \mathbb{R}^{m \times N}$ is the matrix of constraints, and $\boldsymbol{\ell}$ and $\boldsymbol{u}$ are lower and upper bounds vectors, respectively. For instance, when $\boldsymbol{\xi} \in \mathcal{E}_m$, we get

$$\boldsymbol{\Lambda}_{i,j} = \begin{cases} -1 & \text{if } j = i \qquad \text{for any } i = 1, \ldots, N-1; \\ 1 & \text{if } j = i+1 \quad \text{for any } i = 1, \ldots, N-1; \\ 0 & \text{otherwise;} \end{cases}$$

and $\boldsymbol{\ell} = [0, \ldots, 0]^\top \in \mathbb{R}^m$ and $\boldsymbol{u}$ is the vector with components $+\infty$. In that case, we get $m = N-1$ linear inequality constraints on the coefficients vector $\boldsymbol{\xi}$.

The authors of [22] demonstrated that under the representation in Model (M_h):

- Y^N is a finite-dimensional GP with covariance function

$$k_N(x, x') = \boldsymbol{h}(x)^\top \boldsymbol{\Gamma} \boldsymbol{h}(x'), \quad \forall x, x' \in \mathcal{D},$$

  where $\boldsymbol{h}(x) = [h_1(x), \ldots, h_N(x)]^\top$.

- Y^N converges uniformly to Y when N tends to infinity (with probability one).

Let us give the following results concerning Model (M_h).

Proposition 1 (Multiple constraints (M_h))**.** *Suppose that Y^N is defined as in Model (M_h).*

- *Boundedness in a convex region: Let $\mathcal{C}_{bc}$ be a set of continuous functions on $\mathcal{D}$ bounded between two functions f_ℓ and f_u such that the region between the lower bound function f_ℓ and the upper bound function f_u is convex[4]. Then, $Y^N \in \mathcal{C}_{cb}$ if and only if $f_\ell(t_j) \leq \xi_j \leq f_u(t_j)$, for any $j \in \{1, \ldots, N\}$.*

[4] A set $\mathcal{C}$ is convex if and only if for any $x, y \in \mathcal{C}$, $(1-t)x + ty \in \mathcal{C}$, where $t \in]0, 1[$.

- *Convexity: Y^N is convex (i.e., $Y^N \in \mathcal{C}_c$) if and only if*

$$\frac{\xi_{j+1} - \xi_j}{t_{j+1} - t_j} \leq \frac{\xi_{j+2} - \xi_{j+1}}{t_{j+2} - t_{j+1}}, \quad j \in \{1, \ldots, N-2\}. \tag{11}$$

In our case, $t_{j+1} - t_j = \Delta_N = 1/(N-1)$, for any $j \in \{1, \ldots, N-1\}$. Thus, inequalities (11) are equivalent to

$$\xi_{j+2} - 2\xi_{j+1} + \xi_j \geq 0, \quad j \in \{1, \ldots, N-2\},$$

which in turn is equivalent to $\mathbf{\Lambda}\boldsymbol{\xi} \geq \mathbf{0}_{N-2}$ according to the notation in (10), with $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_N]^\top$. The matrix of constraints $\mathbf{\Lambda}$ is given by

$$\mathbf{\Lambda}_{i,j} = \begin{cases} 1 & \text{if } j = i \quad \text{for any } i = 1, \ldots, N-2; \\ -2 & \text{if } j = i+1 \quad \text{for any } i = 1, \ldots, N-2; \\ 1 & \text{if } j = i+2 \quad \text{for any } i = 1, \ldots, N-2; \\ 0 & \text{otherwise.} \end{cases}$$

- *Multiple constraints: For example, Y^N is nondecreasing and bounded in a convex region (i.e., $Y^N(x) \in \mathcal{C}_m \cap \mathcal{C}_{bc}$) if and only if*

$$\begin{cases} \xi_j \leq \xi_{j+1} & j \in \{1, \ldots, N-1\}; \\ f_\ell(t_j) \leq \xi_j \leq f_u(t_j) & j \in \{1, \ldots, N\}; \end{cases} \tag{12}$$

where f_ℓ and f_u are the lower and upper bounds functions. Let us mention that if f_ℓ and f_u are constants, for example $f_\ell(x) = a < b = f_u(x)$, for any $x \in \mathcal{D}$, where $a, b \in \mathbb{R}$, then the linear constraints (12) become $a \leq \xi_1 \leq \ldots \leq \xi_N \leq b$. This leads to only $m = N + 1$ linear constraints on the coefficients vector $\boldsymbol{\xi}$ according to (10).

Proof. For the first item, if Y^N is in $\mathcal{C}_{bc}$, then in particular, $\xi_j = Y^N(t_j) \in [f_\ell(t_j), f_u(t_j)]$, for any $j \in \{1, \ldots, N\}$. Now, if $\xi_j \in [f_\ell(t_j), f_u(t_j)]$, for any $j \in \{1, \ldots, N\}$, then Y^N is in $\mathcal{C}_{bc}$, since Y^N is a piecewise linear interpolation function at the knots $\{t_j\}$.
The proof of the second item is simply a result of the fact that Y^N from Model (M_h) is a piecewise linear interpolating between the function values at the knots $\{t_j\}$ and that the convexity constraints are equivalent to having a nondecreasing slope.
The last item is evident, which completes the proof of the proposition. □

Remark 1.

- *The linear inequality constraints on the coefficients $f_\ell(t_j) \leq \xi_j \leq f_u(t_j)$, for any $j \in \{1, \ldots, N\}$ in Proposition 1 can be written in a matrix form as in (10), where $\mathbf{\Lambda} = \boldsymbol{I}_N$, $\boldsymbol{\ell} = [f_\ell(t_1), \ldots, f_\ell(t_N)]^\top$, and $\boldsymbol{u} = [f_u(t_1), \ldots, f_u(t_N)]^\top$.*
- *The three constraints introduced in Proposition 1 (boundedness, convexity, and monotonicity) can be imposed together.*

The result of Proposition 1 (boundedness constraints) remains valid if the region between the lower and upper bounds functions is non-convex, as long as it can be decomposed into convex sets at nonoverlapping subdomains (cf., Proposition 2 and Figures 5 and 7 below).

Proposition 2 (Boundedness in a non-convex region)**.** *Suppose that $\mathcal{C}_{bnc}$ is a set of continuous functions on $\mathcal{D}$ bounded between two functions f_ℓ and f_u such that the region between the lower bound function f_ℓ and the upper bound function f_u is non-convex. Suppose also that this non-convex region can be decomposed into convex regions on Q nonoverlapping input subdomains $\mathcal{D}_r$, $r \in \{1, \ldots, Q\}$. If the intersection extremities $t_{1,N_1}, \ldots, t_{Q-1,N_{Q-1}}$ of each subdomain are elements of the subdivision*

of $\mathcal{D}$, then Y^N is in $\mathcal{C}_{bnc}$ if and only if $f_\ell(t_{r,j_r}) \leq \xi_{r,j_r} \leq f_u(t_{r,j_r})$, for all $j_r \in \{1,\ldots,N_r\}$ and $r = \{1,\ldots,Q\}$. These linear inequalities on ξ_{r,j_r} can be expressed in a matrix form as follows:

$$\boldsymbol{\ell} \leq \boldsymbol{\Lambda}\boldsymbol{\xi} \leq \boldsymbol{u},$$

where $\boldsymbol{\ell} = [\boldsymbol{\ell}_1^\top,\ldots,\boldsymbol{\ell}_Q^\top]^\top$, $\boldsymbol{u} = [\boldsymbol{u}_1^\top,\ldots,\boldsymbol{u}_Q^\top]^\top$, and $\boldsymbol{\Lambda} = [\boldsymbol{\Lambda}_1^\top,\ldots,\boldsymbol{\Lambda}_Q^\top]^\top$ (block diagonal matrix), with for example, $\boldsymbol{\ell}_1 = [f_\ell(t_{1,1}),\ldots,f_\ell(t_{1,N_1})]^\top$.

Proof. We denote by $\mathcal{C}_1 \cup \ldots \cup \mathcal{C}_Q = \cup_r \mathcal{C}_r = \mathcal{C}_{bnc}$, where $\mathcal{C}_r$ is the convex region on the subdomain $\mathcal{D}_r$ for any $r \in \{1,\ldots,Q\}$. Thanks to Proposition 1, for each $r \in \{1,\ldots,Q\}$, the boundedness constraint $f_\ell(t_{r,j_r}) \leq \xi_{r,j_r} \leq f_u(t_{r,j_r})$ for all $j_r \in \{1,\ldots,N_r\}$ is *equivalent* to $Y^N \in \mathcal{C}_r$. The following condition: 'the intersection extremities $t_{1,N_1},\ldots,t_{Q-1,N_{Q-1}}$ of each subdomain are elements of the subdivision of $\mathcal{D}$' completes the proof of the proposition. □

3.2 Constrained Gaussian process with noisy observations

In this section, we consider the finite-dimensional GP approximation defined in (M_h) given both noisy observations and inequality constraints:

$$Y^N(x) = \sum_{j=1}^{N} \xi_j h_j(x) \quad \text{s.t.} \quad \left\{ \begin{array}{ll} Y^N(x_i) + \epsilon_i = y_i & \text{(noisy observations)}, \\ Y^N \in \mathcal{C} & \text{(inequality constraints)}, \end{array} \right. \tag{13}$$

where $x_i \in \mathcal{D}$ is the design point, $y_i \in \mathbb{R}$ is the data and $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,\sigma^2_{\text{noise}})$, with σ^2_{noise} the noise variance. Given a set of design points $\mathbb{X} = [x_1,\ldots,x_n]^\top \in \mathcal{D}^n$, the noisy observations can be written a matrix form as follows:

$$\mathbf{H}\boldsymbol{\xi} + \boldsymbol{\epsilon} = \boldsymbol{y},$$

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 Gaussian vector and $\mathbf{H}$ is the $n \times N$ matrix defined by $\mathbf{H}_{i,j} := h_j(x_i)$. Following the strategy in [21] and the equivalent in (8), the conditional distribution of Y^N given both noisy observations $\{Y^N(\mathbb{X}) + \boldsymbol{\epsilon} = \boldsymbol{y}\}$ and inequality constraints $\{Y^N \in \mathcal{C}\}$ can be obtained from the conditional distribution of $\boldsymbol{\xi}$ given $\{\mathbf{H}\boldsymbol{\xi} + \boldsymbol{\epsilon} = \boldsymbol{y}\}$ and $\{\boldsymbol{\xi} \in \mathcal{E}\}$

$$\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_N, \boldsymbol{\Gamma}) \quad \text{s.t.} \quad \left\{ \begin{array}{ll} \mathbf{H}\boldsymbol{\xi} + \boldsymbol{\epsilon} = \boldsymbol{y} & \text{(noisy observations)} \\ \boldsymbol{\xi} \in \mathcal{E} & \text{(linear inequality constraints)} \end{array} \right. \tag{14}$$

Hereafter, the linear inequality constraints $\{\boldsymbol{\xi} \in \mathcal{E}\}$ is reformulated as $\boldsymbol{\ell} \leq \boldsymbol{\Lambda}\boldsymbol{\xi} \leq \boldsymbol{u}$, where $\boldsymbol{\ell}$ and $\boldsymbol{u}$ are the lower and upper bounds vectors of dimension m. Now, we will explain the procedure for sampling from the posterior distribution as stated in (14). Since $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_N, \boldsymbol{\Gamma})$, then $\mathbf{H}\boldsymbol{\xi} + \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}_N, \mathbf{H}\boldsymbol{\Gamma}\mathbf{H}^\top + \sigma^2_{\text{noise}}\mathbf{I}_n)$. Under only noisy observations, the conditional distribution of $\boldsymbol{\xi}$ is a multivariate normal (MVN) [23, 33]:

$$\{\boldsymbol{\xi} | \mathbf{H}\boldsymbol{\xi} + \boldsymbol{\epsilon} = \boldsymbol{y}\} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad \text{where}$$

$$\left\{ \begin{array}{rcl} \boldsymbol{\mu} & = & (\mathbf{H}\boldsymbol{\Gamma})^\top (\mathbf{H}\boldsymbol{\Gamma}\mathbf{H}^\top + \sigma^2_{\text{noise}}\mathbf{I}_n)^{-1}\boldsymbol{y} \\ \boldsymbol{\Sigma} & = & \boldsymbol{\Gamma} - (\mathbf{H}\boldsymbol{\Gamma})^\top (\mathbf{H}\boldsymbol{\Gamma}\mathbf{H}^\top + \sigma^2_{\text{noise}}\mathbf{I}_n)^{-1}\mathbf{H}\boldsymbol{\Gamma} \end{array} \right. \tag{15}$$

with $\mathbf{I}_n$ the $n \times n$ identity matrix. Note that this problem is called *hyperplane-truncated* MVN distribution [7, 23, 24, 25]. The consideration of noisy observations in (13) has a *relaxing* effect on the interpolation conditions, as the number of knots and basis functions N does not need to be larger than the size n of the samples (condition required in [22] for the interpolation of noise-free observations called *degree of freedom*). This leads to less restrictive sample spaces and less

expensive MCMC samplers when $N \ll n$ as it is performed on $\mathbb{R}^N$ and independent of the number of observations n. Additionally, it should be mentioned that, unlike interpolation with noise-free observations, the given data does not need to satisfy inequality constraints (see, for example, Figure 3). The posterior distribution (14) is the following truncated MVN distribution:

$$\{\boldsymbol{\xi}|\mathbf{H}\boldsymbol{\xi} = \boldsymbol{y} + \boldsymbol{\epsilon}, \boldsymbol{\ell} \leq \boldsymbol{\Lambda}\boldsymbol{\xi} \leq \boldsymbol{u}\} \sim \mathcal{N}_{\mathcal{T}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\ell}, \boldsymbol{u}), \tag{16}$$

where $\mathcal{N}_{\mathcal{T}}(\boldsymbol{m}, \boldsymbol{C}, \boldsymbol{a}, \boldsymbol{b})$ is the truncated MVN distribution with mean vector $\boldsymbol{m}$, covariance matrix $\boldsymbol{C}$, and lower and upper bounds constraints $\boldsymbol{a}$ and $\boldsymbol{b}$ respectively. Recently, several efficient MCMC algorithms have been proposed to approximate the truncated posterior distribution (16), such as Gibbs sampling [39], Metropolis-Hastings (MH) [28], HMC [30] and the minimax tilting method accept-reject sampler [2]. In the present paper, the fast HMC sampler developed in [30] and implemented in the R package tmg is used.

Let us mention that the posterior mode, which corresponds to the maximum of the posterior probability density function (pdf), i.e.,

$$\boldsymbol{\mu}^* := \arg\min_{\boldsymbol{z}\in\mathbb{R}^N} \{\boldsymbol{z}^\top \boldsymbol{\Gamma}^{-1}\boldsymbol{z} | \mathbf{H}\boldsymbol{z} + \boldsymbol{\epsilon} = \boldsymbol{y}, \boldsymbol{\ell} \leq \boldsymbol{\Lambda}\boldsymbol{z} \leq \boldsymbol{u}\}$$

can be computed. This problem is equivalent to the following quadratic optimization problem subject to convex constraints (see [3, 14])

$$\boldsymbol{\mu}^* := \arg\min_{\substack{\boldsymbol{z}\in\mathbb{R}^N \\ \boldsymbol{\ell}\leq\boldsymbol{\Lambda}\boldsymbol{z}\leq\boldsymbol{u}}} \left\{(\boldsymbol{z}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{z}-\boldsymbol{\mu})\right\} = \arg\min_{\substack{\boldsymbol{z}\in\mathbb{R}^N \\ \boldsymbol{\ell}\leq\boldsymbol{\Lambda}\boldsymbol{z}\leq\boldsymbol{u}}} \left\{\boldsymbol{z}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{z} - \frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Sigma}^{-1}\boldsymbol{z}\right\}, \tag{17}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are given in (15). In the numerical examples of the present paper, the solve.QP function from the R package quadprog is used to compute the posterior mode $\boldsymbol{\mu}^*$. The next sections highlight the advantages of the posterior mode $\boldsymbol{\mu}^*$ over the posterior mean.

Algorithm 1: Sampling scheme of $\{\boldsymbol{\xi}|\mathbf{H}\boldsymbol{\xi} + \boldsymbol{\epsilon} = \boldsymbol{y}, \boldsymbol{\ell} \leq \boldsymbol{\Lambda}\boldsymbol{\xi} \leq \boldsymbol{u}\}$, where $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_N, \boldsymbol{\Gamma})$.

- Initialization: $\boldsymbol{y}$, $\boldsymbol{\Gamma} \in \mathbb{R}^{N\times N}$, $\mathbf{H} \in \mathbb{R}^{n\times N}$, $\boldsymbol{\Lambda} \in \mathbb{R}^{m\times N}$, $\boldsymbol{\ell}$, $\boldsymbol{u}$.
- Compute the conditional mean and covariance of $\{\boldsymbol{\xi}|\mathbf{H}\boldsymbol{\xi} + \boldsymbol{\epsilon} = \boldsymbol{y}\}$

$$\begin{aligned} \boldsymbol{\mu} &= (\mathbf{H}\boldsymbol{\Gamma})^\top (\mathbf{H}\boldsymbol{\Gamma}\mathbf{H}^\top + \sigma^2_{\text{noise}}\mathbf{I}_n)^{-1}\boldsymbol{y}; \\ \boldsymbol{\Sigma} &= \boldsymbol{\Gamma} - (\mathbf{H}\boldsymbol{\Gamma})^\top (\mathbf{H}\boldsymbol{\Gamma}\mathbf{H}^\top + \sigma^2_{\text{noise}}\mathbf{I}_n)^{-1}\mathbf{H}\boldsymbol{\Gamma}. \end{aligned}$$

- Compute the posterior mode by solving the quadratic optimization problem subject to linear inequality constraints (the solve.QP function from the R package quadprog is used in this paper):

$$\boldsymbol{\mu}^* := \arg\min_{\boldsymbol{z}\in\mathbb{R}^N} \{\boldsymbol{z}^\top \boldsymbol{\Gamma}^{-1}\boldsymbol{z} | \mathbf{H}\boldsymbol{z} + \boldsymbol{\epsilon} = \boldsymbol{y}, \boldsymbol{\ell} \leq \boldsymbol{\Lambda}\boldsymbol{z} \leq \boldsymbol{u}\}.$$

- Sample from the truncated MVN distribution (HMC tmg is used in this paper)

$$\{\boldsymbol{\xi}|\mathbf{H}\boldsymbol{\xi} + \boldsymbol{\epsilon} = \boldsymbol{y}, \boldsymbol{\ell} \leq \boldsymbol{\Lambda}\boldsymbol{\xi} \leq \boldsymbol{u}\} \sim \mathcal{N}_{\mathcal{T}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\ell}, \boldsymbol{u}).$$

Remark 2. *Algorithm 1 generates the posterior distribution of the coefficients vector $\boldsymbol{\xi}$ conditionally on noisy observations $\{\boldsymbol{H\xi} + \boldsymbol{\epsilon} = \boldsymbol{y}\}$ and linear inequality constraints $\{\boldsymbol{\xi} \in \mathbb{R}^N$ s.t. $\boldsymbol{\ell} \leq \boldsymbol{\Lambda\xi} \leq \boldsymbol{u}\}$. To get the posterior distribution of Y^N, i.e., $\{Y^N|Y^N(\mathbb{X}) + \boldsymbol{\epsilon} = \boldsymbol{y}, Y^N \in \mathcal{C}\}$, one can substitute the generated samples from Algorithm 1 into Model (M_h). Let us note that the posterior mode $\boldsymbol{\mu}^*$ can serve as a suitable starting point for the HMC sampler.*

Before presenting numerical examples of the proposed approach for various types of inequality constraints, we define the following two estimators.

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

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

where $\boldsymbol{\mu}^* = [\mu_1^*, \dots, \mu_N^*]^\top \in \mathbb{R}^N$ is the posterior mode computed by (17) *and $\boldsymbol{h}(x) = [h_1(x), \dots, h_N(x)]^\top$.*

Let us provide some comments: the MAP estimate in Algorithm 1 is independent of the sampling process. It is determined only by solving a quadratic optimization problem with linear inequality constraints (17). Furthermore, it has been shown that the MAP estimate M^N converges to the optimization spline problem in both noise and noise-free observation cases when using the *hat* basis functions, i.e., Model (M_h) (see [1, 15]). These two results can be seen as a generalization of the Kimeldorf-Wahba correspondence [17] between Bayesian estimation on stochastic processes and smoothing by splines.

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

$$m^N(x) := \mathrm{E}\left[Y^N(x) \middle| Y^N(\mathbb{X}) + \boldsymbol{\epsilon} = \boldsymbol{y}, Y^N \in \mathcal{C}\right] = \boldsymbol{h}(x)^\top \boldsymbol{\mu}_c,$$

where $\boldsymbol{\mu}_c := \mathrm{E}\left[\boldsymbol{\xi} | \boldsymbol{H}\boldsymbol{\xi} + \boldsymbol{\epsilon} = \boldsymbol{y}, \boldsymbol{\ell} \leq \boldsymbol{\Lambda}\boldsymbol{\xi} \leq \boldsymbol{u}\right]$ is the posterior mean which is computed from simulations and $\boldsymbol{h}(x) = [h_1(x), \dots, h_N(x)]^\top$.

3.3 Illustrative examples of Model (M_h)

In the following examples, the performance of the MAP estimate using Model (M_h) is highlighted, and the flexibility of Model (M_h) to incorporate multiple types of convex and non-convex inequality constraints is shown.

Example 1 (Boundedness in a convex set)**.** *We consider the function $f_1(x) = 0.8x\sin(5x)$ for any $x \in \mathcal{D}$. This function is bounded on $\mathcal{D}$ between convex and concave functions f_ℓ and f_u respectively:*

$$f_\ell(x) = (x - 0.5)^2 - 1.2 \quad \textit{and} \quad f_u(x) = -(x - 0.5)^2 + 0.3.$$

Additionally, this function is slightly flat and close to the upper bound function f_u on the interval $[0, 0.5]$.

Figure 3 illustrates the GP approximation from Model (M_h) with and without boundedness constraints. We use the Matérn covariance function with $\nu = 5/2$ and $\theta = 0.4$. We fix $N = 30$ to avoid the possibility of overfitting [29]. Our numerical analysis indicates that this value of N provides a satisfactory approximation, as shown in Figure 13 in Sect. 4.5. In the right panel, we use the HMC sampler [30] to sample from the posterior distribution of the coefficients $\{\xi_j\}$ as in Algorithm 1. The black solid curve represents the true bounded function f_1. The two black dashed thin curves correspond to the lower and upper bounds functions f_ℓ and f_u respectively. The red dashed (resp. blue dashed-dotted) curve represents the MAP (resp. mAP) estimate (cf., Definitions 1 and 2). The gray shaded area is the 95% pointwise confidence interval. The black stars represent the 50 training data generated from (1) *using the true function f_1 and a true noise standard deviation of $\sigma_{noise} = 0.25$. First, we observe that both the prediction estimates and the confidence intervals in*

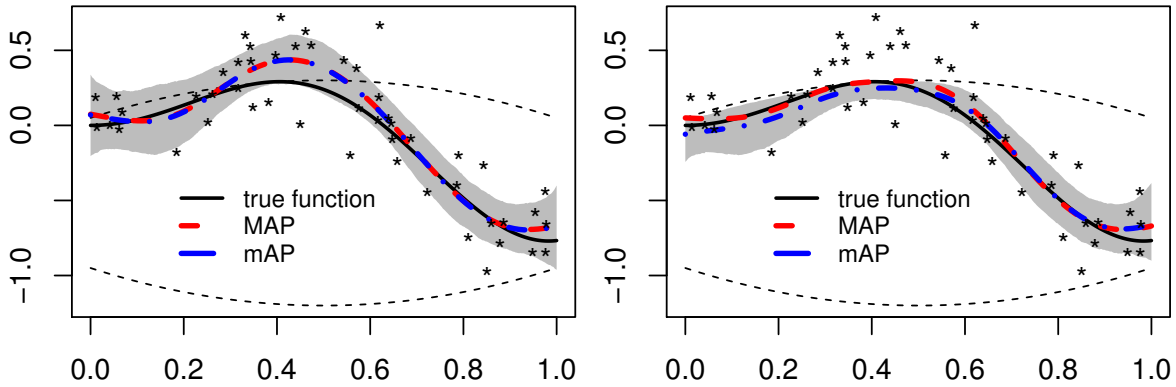
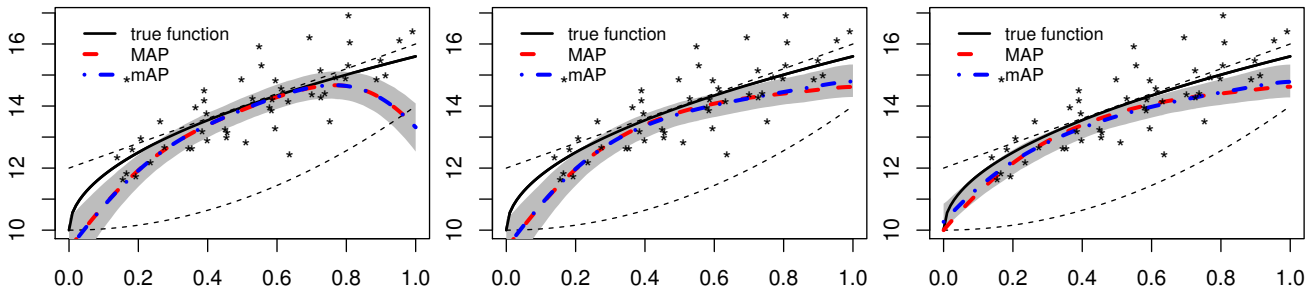


Figure 3: In the left (resp. right) panel, the true function and its estimates (MAP and mAP) are plotted for the GP approximation of Model (M_h) without (resp. with) boundedness constraints. The 50 training data are represented by black stars, and the 95% pointwise confidence interval is shown in gray. The lower and upper bounds functions are indicated by the two black dashed thin curves.

the left panel do not satisfy boundedness constraints. Second, we observe that including the boundedness constraints into the posterior distribution (right panel) results in more accurate predictions and smaller confidence intervals compared to those produced by the unconstrained GP model (left panel). Let us note that in the left panel, the MAP and mAP estimates coincide in the unconstrained case according to the result in [17]. However, for the constrained case on the right panel, the MAP and mAP no longer coincide. In this paper, the performance of the MAP estimate is further discussed, with a particular focus on its behavior compared to that of the other estimate. Let us conclude this example by noting that, based on this numerical experiment, the MAP estimate appears to perform better visually than the mAP estimate. It is closer to the observed values than the mAP estimate.

Example 2 (Monotonicity and boundedness constraints)**.** *In this example, our aim is to show the flexibility of the proposed Model (M_h) in incorporating multiple types of inequality constraints simultaneously. Additionally, we examine the behavior of the MAP estimate in terms of prediction accuracy when the function is flat and close to upper and/or lower bounds at certain regions of $\mathcal{D}$. To do this, we consider the function $f_2(x) = 5.6\sqrt{x} + 10$, which is increasing and bounded between the following convex and linear functions on $\mathcal{D}$:*

$$f_\ell(x) = 4x^2 + 10 \quad \textit{and} \quad f_u(x) = 4x + 12.$$



(a) Unconstrained (b) Monotonicity constraints (c) Monotonicity and boundedness

Figure 4: GP approximation from Model (M_h) satisfying different constraints.

Figure 4 shows side by side the unconstrained GP (left), Model (M_h) under monotonicity constraints (middle), and Model (M_h) under both monotonicity and boundedness constraints (right).

The Matérn covariance function with $\nu = 5/2$ and $\theta = 0.5$ is used. The black stars represent the training data generated from (1) using f_2 and a true noise standard deviation $\sigma_{noise} = 1$. We set $N = 30$ in (M_h) and we use the HMC technique to sample from the posterior distribution of $\{\xi_j\}$ as in Algorithm 1. The descriptions of the panels are identical to those in Figure 3. We observe that including monotonicity constraints improves the predictive accuracy and reduces the confidence intervals in comparison to those generated by the GP without constraints (left panel). However, they do not satisfy the boundedness constraints provided by the lower and upper bounds functions (black dashed thin curves). Adding the boundedness constraints, as shown in the right panel, leads to more accurate predictions and more realistic confidence intervals.

Example 3 (Boundedness in a non-convex region)**.** *In this illustrative example, we consider the scenario where the true underlying function is bounded between two functions (a lower and an upper bound) f_ℓ and f_u, and the region between these bounding functions is non-convex. For example, consider the case where the lower and upper bound functions are as follows:*

$$f_\ell(x) = (x - 0.5)^2 + 0.1 \quad \textit{and} \quad f_u(x) = \begin{cases} -x + 0.8 & \textit{if } x \in [0, 0.4] \\ 0.5x + 0.2 & \textit{if } x \in (0.4, 1] \end{cases}$$

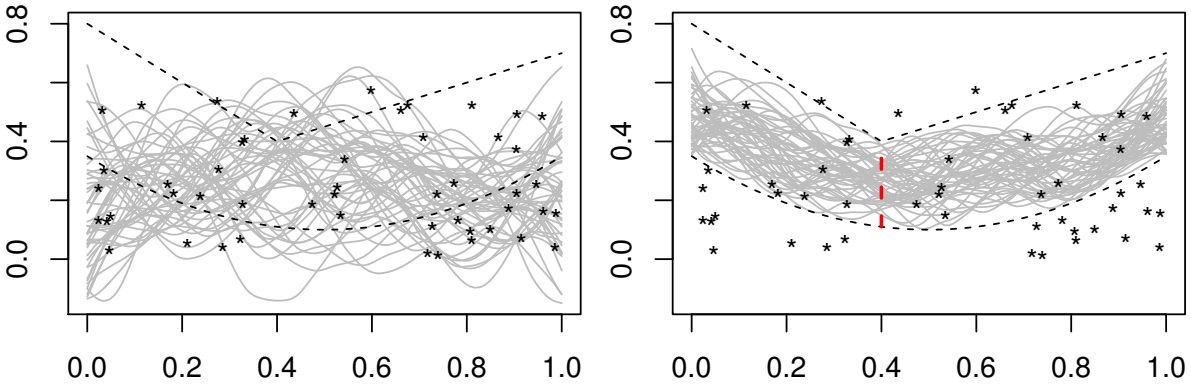


Figure 5: GP approximation from Model (M_h) with and without boundedness constraints at the right and left panels, respectively. The input domain is divided into two nonoverlapping subdomains by a vertical red dashed line on the right, resulting in two bounded convex regions.

Figure 5 shows fifty GP sample paths from Model (M_h) with and without boundedness constraints at the right and left panels, respectively. We use the Matérn covariance function with $\nu = 5/2$ and $\theta = 0.4$. The black stars represent the training observations, where $\{x_i\}$ are generated uniformly on $[0, 1]$ and the data $\{y_i\}$ are generated uniformly on $[0, 0.6]$. The region between the lower and upper bounds functions f_ℓ and f_u is non-convex. However, the input domain, $\mathcal{D}$, can be divided into two nonoverlapping subdomains at $x = 0.4$, yielding two convex regions. In that case, we apply Proposition 2 with $Q = 2$ to get the matrix of constraints $\boldsymbol{\Lambda} = [\boldsymbol{\Lambda}_1^\top, \boldsymbol{\Lambda}_2^\top]^\top$, and the lower and upper bounds vectors:

$$\boldsymbol{\ell} = [\boldsymbol{\ell}_1^\top, \boldsymbol{\ell}_2^\top]^\top \quad \textit{and} \quad \boldsymbol{u} = [\boldsymbol{u}_1^\top, \boldsymbol{u}_2^\top]^\top.$$

As required in Proposition 2, we impose $x = 0.4$ to be an element of the subdivision $\{t_j\}$. We use Algorithm 1 to get the posterior distribution of $\{\xi_j\}$. Unlike the left panel, the sample paths of the GP satisfy boundedness constraints everywhere. Let us note that these types of boundedness constraints can be integrated with other inequality constraints, such as monotonicity or convexity constraints, as shown in Example 4.

Example 4 (Sequential constraints)**.** *Model (M_h) is capable of incorporating different inequality constraints sequentially at nonoverlapping intervals as shown in Figures 6 and 7 below.*

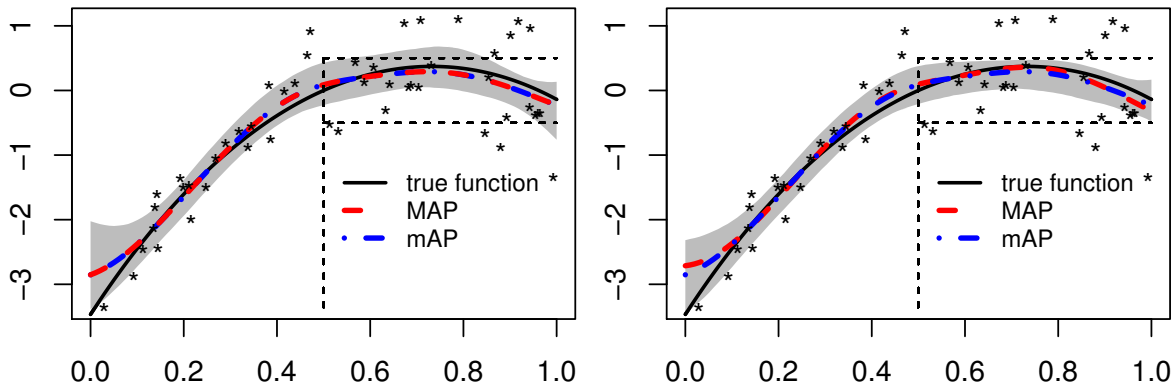


Figure 6: Left: GP model without any constraints. Right: GP approximation from Model (M_h) under monotonicity constraints on $[0, 0.5]$ and under boundedness constraints on $[0.5, 1]$. The vertical dashed line splits the input set into two nonoverlapping intervals at $x = 0.5$. The two horizontal dashed lines represent the lower and upper bounds.

In Figure 6, the true underlying function is defined as $f_4(x) = (5 - 5.2x)\log((x+1)/(2-x))$. *It admits two different behaviors at two nonoverlapping intervals. Indeed, it is increasing on* $[0, 0.5]$ *and bounded between* -0.5 *and* 0.5 *on* $[0.5, 1]$. *We observe that adding sequential constraints significantly reduces the 95% pointwise confidence intervals. The monotonicity and boundedness constraints are satisfied at their respective subintervals when using the proposed approach under sequential constraints, as shown in the right panel.*

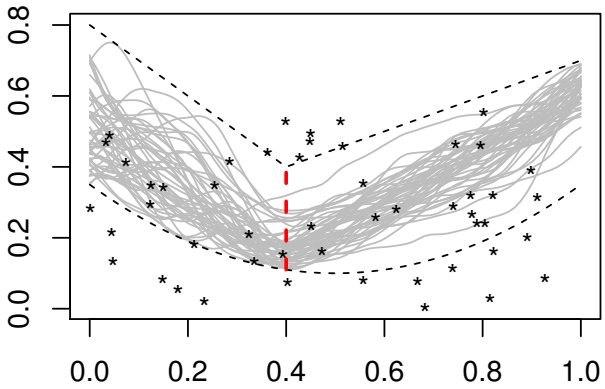


Figure 7: GP approximation from Model (M_h) under boundedness constraints on $[0, 0.4]$ and under both boundedness and monotonicity constraints on $[0.4, 1]$. The vertical dashed line divides the interval $[0, 1]$ into two nonoverlapping subintervals at $x = 0.4$.

In Figure 7, the lower and upper bounds functions f_ℓ *and* f_u *are those used in Figure 5. The input domain is split into two nonoverlapping subdomains by a red vertical dashed line at* $x = 0.4$. *We suppose that the underlying function admits two different behaviors within these two subdomains. In the first subdomain* $[0, 0.4]$, *the function is bounded. In the second subdomain* $[0.4, 1]$, *it is both nondecreasing and bounded. Again, we applied Proposition 2 with* $Q = 2$, *ensuring that* $x = 0.4$ *is included as an element in the subdivision* $\{t_j\}$. *The gray curves represent the GP sample paths from Model* (M_h) *under sequential constraints. The fifty black stars are the training data and the black dashed thin curves are the lower and upper bounds (identical to those in Figure 5). The sample paths respect boundedness constraints on the entire domain and nondecreasing constraints on* $[0.4, 1]$.

Let us recall that Model (M_h) has been generalized to multidimensional input spaces, as detailed in [22]. For example, the convexity constraints with respect to two input variables or to only one

variable are given in [21]. For simplicity, the two-dimensional case is considered. Next section shows the performance and the flexibility of the proposed approach to incorporate multiple constraints, such as monotonicity and boundedness (with lower and upper bound functions), as well as when the bounded region is non-convex.

3.4 Multidimensional input spaces

The finite-dimensional GP approximation defined in (M_h) can be extended to d-dimensional input spaces by tensorization [22]. For simplicity of notations, we focus on the case $d = 2$, with $\mathcal{D}^2 = [0,1]^2$ and $N_1 \times N_2$ knots located on a regular (or non regular) grid. Then, for any $\boldsymbol{x} = (x_1, x_2) \in \mathcal{D}^2$, the finite-dimensional Gaussian approximation is given by

$$Y^{N_1,N_2}(x_1, x_2) := \sum_{j_1=1}^{N_1} \sum_{j_2=1}^{N_2} \xi_{j_1,j_2} h_{j_1}^1(x_1) h_{j_2}^2(x_2), \tag{18}$$

where $\{h_{j_1}^1\}$ and $\{h_{j_2}^2\}$ are the *hat* functions defined in (4), and $\xi_{j_1,j_2} = Y(t_{j_1}, t_{j_2})$, with $\{(t_{j_1}, t_{j_2})\}$ the knots. Similarly to the one-dimensional case, $\boldsymbol{\xi} = (\xi_{j_1,j_2}) \in \mathbb{R}^{N_1 \times N_2}$ is a zero-mean Gaussian vector with covariance matrix $\boldsymbol{\Gamma}$ as in (6). Thus, we have the following results:

- For the case of monotonicity in two dimensions, the constraints to be satisfied are given by $\xi_{j+1,l} \geq \xi_{j,l}$ and $\xi_{j,l+1} \geq \xi_{j,l}$ for any $j \in \{1, \ldots, N_1 - 1\}$ and $l \in \{1, \ldots, N_2 - 1\}$. The constraints for the monotonicity with respect to one of the two input variables can be computed in a similar way [22].

- For the convexity in two dimensions, the constraints to be satisfied are given by

$$\frac{\xi_{j+1,l} - \xi_{j,l}}{t_{j+1} - t_j} \leq \frac{\xi_{j+2,l} - \xi_{j+1,l}}{t_{j+2} - t_{j+1}} \text{ and } \frac{\xi_{j,l+1} - \xi_{j,l}}{t_{l+1} - t_l} \leq \frac{\xi_{j,l+2} - \xi_{j,l+1}}{t_{l+2} - t_{l+1}},$$

  for any $j \in \{1, \ldots, N_1 - 2\}$ and $l \in \{1, \ldots, N_2 - 2\}$. The constraints for the convexity with respect to one of the two input variables can be computed in a similar way.

- For the upper and lower bound functions, f_u and f_ℓ respectively, that define a convex set, the constraints are expressed as $f_\ell(t_j, t_l) \leq \xi_{j,l} \leq f_u(t_j, t_l)$, for all $j \in \{1, \ldots, N_1\}$ and $l \in \{1, \ldots, N_2\}$.

Remark 3. *The boundedness constraints in the last item above can be extended to the case where the region between the lower and upper bounds is non-convex. As in the one-dimensional case, the only requirement is that the points where the input domain $\mathcal{D}^2$ is divided into convex subsets must be part of the discretization grid (see Example 6).*

Example 5 (Numerical illustrations in 2D)**.** *The purpose of this numerical example is to demonstrate the effectiveness of the proposed approach in the two-dimensional scenario. The flexibility of Model* (18) *in incorporating multiple constraints in two dimensions is highlighted.*

Figure 8 shows an example where boundedness and monotonicity constraints in two dimensions are imposed together. The two-dimensional squared exponential (SE) covariance function is used

$$k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{(x_1 - x_1')^2}{2\theta_1} - \frac{(x_2 - x_2')^2}{2\theta_2}\right), \quad \boldsymbol{x} = (x_1, x_2) \in \mathcal{D}^2, \tag{19}$$

where θ_1 and θ_2 are the correlation length hyperparameters. The one hundred training observations (black stars) were generated using the Hypercube Latin from the R package lhs, Eq. (1)*, the function $f(x_1, x_2) = 5.6\sqrt{x_1} + x_2 + 10$, and a true $\sigma_{noise} = 1$. This function is monotone nondecreasing with*

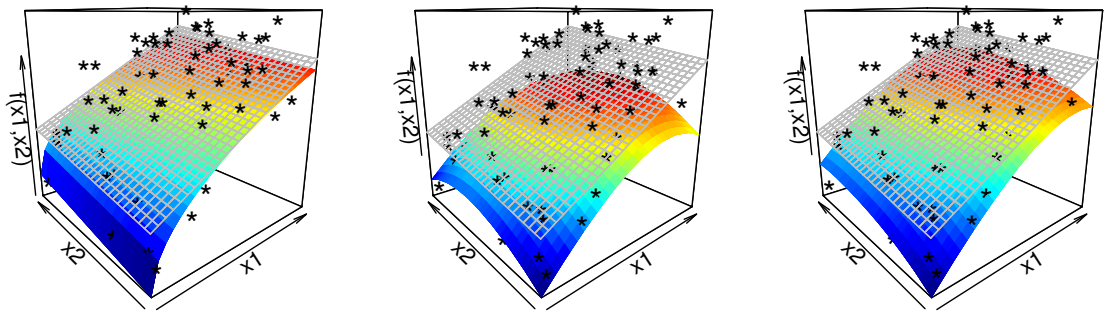


Figure 8: The true nondecreasing and bounded function (left). The MAP estimate from Model (18) under boundedness constraints (middle) and under both boundedness and monotonicity constraints (right). The gray surface represents the upper bound constraint, and the black stars represent the training data.

respect to the two input variables and bounded from above by $f_u(x_1, x_2) = 4x_1 + x_2 + 12$. *In this figure, we illustrate the true function with the upper bound constraints on the left, the MAP estimate with boundedness constraints in the middle, and the MAP with both boundedness and monotonicity constraints on the right, side by side. The addition of multiple constraints enhances the accuracy of predictions.*

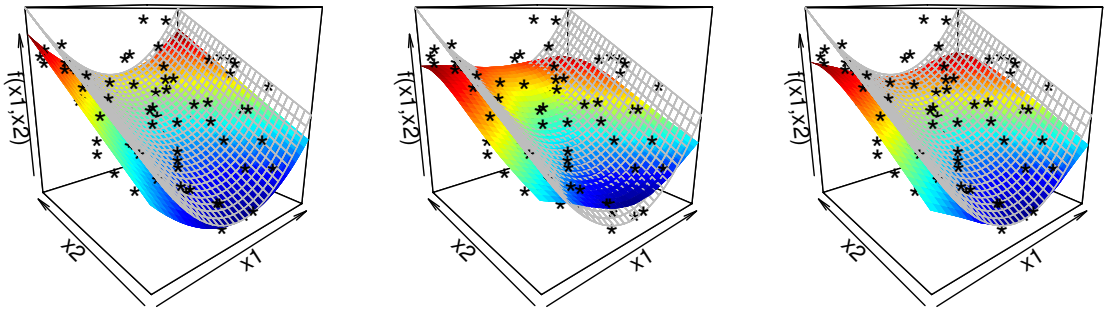


Figure 9: In the left, middle, and right panels, the true bounded function (as well as the unconstrained mean and MAP) are displayed along with the upper bound constraint function (gray surface) and the training observations (black stars).

Example 6 (Boundedness in a non-convex region 2D)**.** *Figure 9 illustrates an example where boundedness constraints form a non-convex region. As in Figure 8, the two-dimensional SE covariance function* (19) *is used. The one hundred training observations (black stars) are generated using the Hypercube Latin from the* R *package* lhs*, Eq.* (1)*, the true bounded function* $f(x_1, x_2) = 2(x_1 - 0.5)^2 + x_2$*, and a true* σ_{noise} *fixed at* 0.25*. This function is bounded from above on the unit square* $\mathcal{D}^2$ *by the function* $f_u(x_1, x_2) = 3(x_1 - 0.5)^2 + x_2$*. The boundedness constraints in this example form a non-convex set. However, this set can be decomposed into two convex subsets at two nonoverlapping subdomains obtained by splitting* $\mathcal{D}^2$ *at* $x_1 = 0.5$*. As in the one-dimensional case, the only requirement is to include the knots at* $x_1 = 0.5$ *in the subdivision grid. Model* (18) *is used, where* N_1 *and* N_2 *are fixed at 9. Thus, we have 81 knots and basis functions. We illustrate, side by side, the true bounded function with the upper bound constraints and training samples on the left, the unconstrained mean, i.e.,* μ_c *in* (3) *in the middle, and , the MAP estimate on the right. Contrary to the MAP estimate, the unconstrained mean violates the upper bound constraints in certain regions.*

4 Constrained GPs: C^p approximation, $p \geq 1$

In this section, we generalize Model (M_h) in order to provide smoother sample paths by proposing new basis functions. The capability of this new model to incorporate multiple constraints such as monotonicity with bounded slope constraints is investigated. Furthermore, a comparison of the prediction accuracy between these models is included.

4.1 C^1 approximation with Model (M_ϕ)

In this section, shape constraints for continuous and differentiable functions $f \in C^1(\mathcal{D}, \mathbb{R})$ is considered. For example, the convex set $\mathcal{C}_m$ of nondecreasing functions is given by

$$\mathcal{C}_m = \left\{ f \in C^1(\mathcal{D}, \mathbb{R}) \text{ s.t. } f'(x) \geq 0,\ x \in \mathcal{D} \right\}.$$

As stated in [22], any at least differentiable function f can be written as $f(x) = f(0) + \int_0^x f'(t)\mathrm{d}t$, where $f'(t)$ represents the derivative of f at t. Following the strategy of Sect. 3.1, any differentiable function f can be approximated by $\tilde{f}_N(x) = f(0) + \sum_{j=1}^N f'(t_j)\phi_j(x)$, for any $x \in \mathcal{D}$, where we recall that $\phi_j(x) = \int_0^x h_j(x)\mathrm{d}x$, for any $x \in \mathcal{D}$.

Lemma 2 (Uniform convergence C^1)**.** *For any $f \in C^1(\mathcal{D}, \mathbb{R})$, the function $\tilde{f}_N := f(0) + \sum_{j=1}^N f'(t_j)\phi_j(x)$, where $\{\phi_j\}$ are defined in (4) converges uniformly to f when N tends to infinity.*

Proof. For any $x \in \mathcal{D}$, we have

$$\begin{aligned}
|\tilde{f}_N(x) - f(x)| &= \left| f(0) + \sum_{j=1}^N f'(t_j)\phi_j(x) - \left(f(0) + \int_0^x f'(t)\mathrm{d}t \right) \right| \\
&= \left| \sum_{j=1}^N f'(t_j) \int_0^x h_j(t)\mathrm{d}t - \int_0^x f'(t)\mathrm{d}t \right| \\
&= \left| \int_0^x \sum_{j=1}^N \left(f'(t_j) - f'(t) \right) h_j(t)\mathrm{d}t \right| \leq \int_0^x \sup_{|t'-t| \leq \Delta_N} |f'(t') - f'(t)|\mathrm{d}t.
\end{aligned}$$

The proof is done by following the reasoning of Lemma 1. □

In Figure 10 right panel, the red solid curve is the deterministic function $f(x) = x^3$, which verifies monotonicity (nondecreasing) constraints. The black dashed curve, represented by $\tilde{f}_N(x) = f(0) + \sum_{j=1}^N f'(t_j)\phi_j(x)$, is an approximation of $f(x)$ using a uniform subdivision with $N = 5$ basis functions. The gray curves represent the basis functions $\{\phi_j\}$ defined in (4). The black stars represent the value of the true function at knots $\{t_j\}$. In contrast, the left panel shows the approximation of $f'(x) = 3x^2$ (red solid curve) using the derivative of our proposed approach $\tilde{f}'_N$ (black dashed curve) and the hat functions $\{h_j\}$.

Now we consider the second model proposed in [22]

$$Y^N(x) := Y(0) + \sum_{j=1}^N Y'(t_j)\phi_j(x) = \xi_0 + \sum_{j=1}^N \xi_j \phi_j(x), \quad x \in \mathcal{D}, \tag{M_ϕ}$$

where, we denote by $\xi_j = Y'(t_j)$, for any $j \in \{1, \ldots, N\}$, $\xi_0 = Y(0)$, and $\{\phi_j\}$ are the basis functions defined in (4). A comparison between Models (M_h) and (M_ϕ) is studied later in this section. The proposed approach, i.e., Model (M_ϕ) is also applicable for non-uniform subdivision as in Sect. 3.1. This model has been considered in [44] for revisiting the proton-radius problem, and more recently

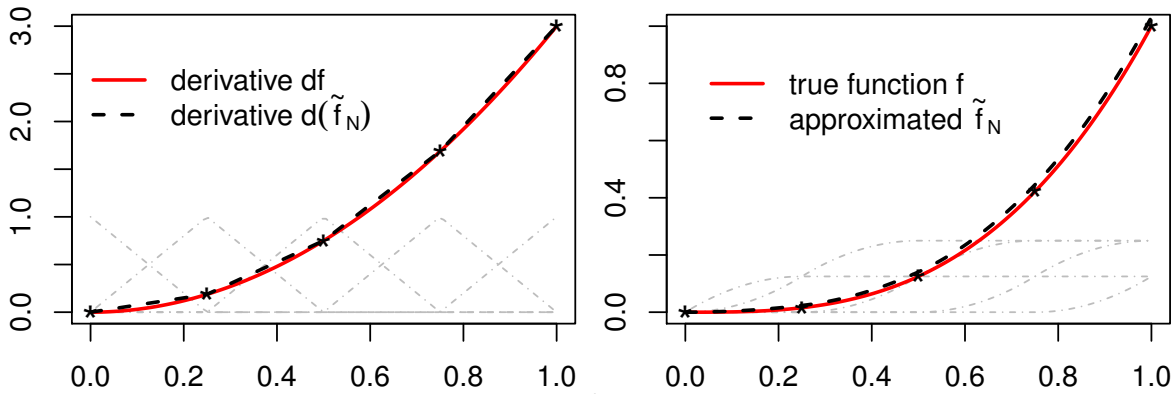


Figure 10: Left: the approximation of $f'(x) = 3x^2$ (red solid curve) by the derivative of the proposed approach $\tilde{f}'_N$ (black dashed curve) together with the hat functions $\{h_j\}$. Right: the approximation of the monotone nondecreasing function $f(x) = x^3$ (red solid curve) by the proposed approach $\tilde{f}_N$ (black dashed curve) together with the basis functions $\{\phi_j\}$ (gray curves). A uniform subdivision is used with $N = 5$ basis functions and knots.

in [45] to describe the *mass-shifting* phenomenon of the truncated MVN distribution for a flat region problem. Since $(Y(x))_{x\in\mathcal{D}}$ is a zero-mean GP, then the vector $[Y(t_1), \ldots, Y(t_N)]^\top$ is also zero-mean, and Gaussian [33]. By [10, 31], we know that $[Y'(t_1), \ldots, Y'(t_N)]^\top$ is still a Gaussian vector with covariance matrix

$$\mathbf{G}_{j,l} = \mathrm{Cov}(\xi_j, \xi_l) = \frac{\partial^2}{\partial x \partial x'} k(t_j, t_l), \quad \forall j, l \in \{1, \ldots, N\},$$

where we recall that k is the covariance function of the original GP Y. Therefore, the covariance matrix of the Gaussian vector $\boldsymbol{\xi} = [\xi_0, \xi_1, \ldots, \xi_N]^\top$ is given by

$$\boldsymbol{\Gamma} = \begin{pmatrix} k(0,0) & \frac{\partial}{\partial x'} k(0, t_l) \\ \frac{\partial}{\partial x} k(t_j, 0) & \mathbf{G}_{j,l} \end{pmatrix}_{1 \le j,l \le N} \in \mathbb{R}^{(N+1)^2}.$$

Thus, $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_{N+1}, \boldsymbol{\Gamma})$ as in Sect. 3.1, where, $\mathbf{0}_{N+1}$ is the $(N+1)$-dimensional zero vector.

Proposition 3 (Monotonicity and bounded slope)**.** *If the GP approximation Y^N is defined as in (M_ϕ), then, Y^N is nondecreasing (resp. nonincreasing) and $(Y^N)'(x) \in [\ell, u]$, for any $x \in \mathcal{D}$ if and only if $\xi_j \in [\ell, u]$ for any $j \in \{1, \ldots, N\}$, where the lower bound ℓ is nonnegative (resp. the upper bound u is nonpositive). This property can be extended to bounded slope constraints at a subset of $\mathcal{D}$ (see the right panel in Figure 16).*

The linear inequality constraints on the coefficients $\{\xi_j\}$ in Proposition 3 can be written in a matrix form as follows

$$\boldsymbol{\ell} \le \boldsymbol{\Lambda}\boldsymbol{\xi} \le \boldsymbol{u},$$

where $\boldsymbol{\ell}$ and $\boldsymbol{u}$ are the m-dimensional vectors representing the lower and upper bounds, and $\boldsymbol{\Lambda}$ is the $m \times (N+1)$ matrix of constraints, with m number of linear constraints. Since the lower bound vector $\boldsymbol{\ell}$ is nonnegative, incorporating monotonicity and bounded slope constraints requires only $m = N$ linear constraints on the basis coefficients $\{\xi_j\}$.

Proof of Proposition 3. We have for any $x \in \mathcal{D}$

$$Y^N(x) = \xi_0 + \sum_{j=1}^{N} \xi_j \phi_j(x) \quad \Rightarrow \quad (Y^N)'(x) = \sum_{j=1}^{N} \xi_j h_j(x) = \sum_{j=1}^{N} Y^N(t_j) h_j(x),$$

where $\{h_j\}$ are the *hat* functions defined in (4). The above right hand side equation corresponds to Model (M_h). So, the bounded slope constraints $(Y^N)'(x) \in [\ell, u]$, for any $x \in \mathcal{D}$ is *equivalent* to $\xi_j \in [\ell, u]$, for any $j \in \{1, \ldots, N\}$. This concludes the proof of the proposition since the monotonicity (nondecreasing) constraints are equivalent to the nonnegativity of the coefficients $\{\xi_j\}$. □

Corollary 1. *If Y^N is defined as in Model (M_ϕ), and f_ℓ and f_u are lower and upper bounds functions such that the region between these two functions is convex, then, $(Y^N)'(x) \in [f_\ell(x), f_u(x)]$ for any $x \in \mathcal{D}$ if and only if $\xi_j \in [f_\ell(t_j), f_u(t_j)]$, for any $j \in \{1, \ldots, N\}$.*

Proof. The proof is similar to the one given in the first item Proposition 1. □

Remark 4. *The result in Corollary 1 can be extended to include bounded slope constraints on subsets of $\mathcal{D}$, as well as non-convex regions between lower and upper bounds functions f_ℓ and f_u, by decomposing the input domain $\mathcal{D}$ into nonoverlapping subdomains with convex regions. As in Proposition 2, the only requirement is to include the intersection extremities of the subdomains in the subdivision $\{t_j\}$.*

Proposition 4 (Multiple constraints (M_ϕ))**.** *If the GP approximation Y^N is defined as in (M_ϕ), then,*

- *Monotonicity and boundedness: Y^N is nondecreasing and nonnegative (resp. nonincreasing and nonpositive) if and only if $\{\xi_j\}$ are nonnegative (resp. nonpositive) for any $j \in \{0, \ldots, N\}$.*
- *Convexity: Y^N is convex if and only if $\xi_j \leq \xi_{j+1}$, for any $j \in \{1, \ldots, N-1\}$.*
- *Monotonicity and convexity: Y^N is monotone (nondecreasing) and convex in the entire domain if and only if $0 \leq \xi_1 \leq \ldots \leq \xi_N$.*
  This leads to only $m = N$ linear constraints on the basis coefficients $\{\xi_j\}$ according to the notation in (10).
- *Monotonicity, convexity and boundedness: Y^N is nondecreasing, convex, and nonnegative if and only if $0 \leq \xi_1 \leq \ldots \leq \xi_N$ and $\xi_0 \geq 0$.*

Proof. If Y^N is nondecreasing then from Proposition 3 the coefficients $\{\xi_j\}$ are nonnegative, for any $j \in \{1, \ldots, N\}$. Since $Y^N(x)$ is nonnegative for any x in $\mathcal{D}$, in particular $\xi_0 = Y^N(0)$ is nonnegative. Now, if $\{\xi_j\}$ are nonnegative, for any $j \in \{0, \ldots, N\}$, then Y^N is nondecreasing (from Proposition 3) and nonnegative since the basis functions $\{\phi_j\}$ are nonnegative.
In one hand, if Y^N is convex, then $(Y^N)'$ is nondecreasing. In particular, $(Y^N)'(t_j) \leq (Y^N)'(t_{j+1})$ which implies that $\xi_j \leq \xi_{j+1}$ for any $j \in \{1, \ldots, N-1\}$. In the second hand, if $\xi_j \leq \xi_{j+1}$ for any $j \in \{1, \ldots, N-1\}$, then $(Y^N)'$ is nondecreasing since $(Y^N)'(x) = \sum_{j=1}^N \xi_j h_j(x)$, which is a piecewise linear function. The last two items are obvious. □

Corollary 2. *The bounded slope constraints from Proposition 3 can be combined with the ones provided in Proposition 4.*

As in Sect. 3.2, the GP approximation defined in (M_ϕ) with both noisy observations and inequality constraints is considered:

$$Y^N(x) = \xi_0 + \sum_{j=1}^{N} \xi_j \phi_j(x) \quad \text{s.t.} \quad \begin{cases} Y^N(x_i) + \epsilon_i = y_i & \text{(noisy observations);} \\ Y^N \in \mathcal{C} & \text{(inequality constraints).} \end{cases}$$

where $x_i \in \mathcal{D}$ is the input, $y_i \in \mathbb{R}$ is the output and $\epsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2_{\text{noise}})$, with σ^2_{noise} the noise variance. Given a set of design points $\mathbb{X} = [x_1, \ldots, x_n]^\top$, the noisy observations can be written in a matrix form as follows:

$$\boldsymbol{\Phi}\boldsymbol{\xi} + \boldsymbol{\epsilon} = \boldsymbol{y},$$

where $\boldsymbol{y} = [y_1, \ldots, y_n]^\top$ is the vector of observations, $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_n]^\top$ is the noise Gaussian vector and $\boldsymbol{\Phi}$ is the $n \times (N+1)$ matrix defined by

$$\boldsymbol{\Phi}_{i,j} := \begin{cases} 1 & \text{if } j = 1; \\ \phi_{j-1}(x_i) & \text{for } j = 2, \ldots, N+1. \end{cases}$$

Following the strategy in [22] and (8), the conditional distribution of Y^N under both noisy observations $Y^N(\mathbb{X}) + \boldsymbol{\epsilon} = \boldsymbol{y}$ and inequality constraints $Y^N \in \mathcal{C}$ can be obtained from the conditional distribution of $\boldsymbol{\xi}$ given $\boldsymbol{\Phi}\boldsymbol{\xi} + \boldsymbol{\epsilon} = \boldsymbol{y}$ and $\boldsymbol{\ell} \leq \boldsymbol{\Lambda}\boldsymbol{\xi} \leq \boldsymbol{u}$

$$\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}_{N+1}, \boldsymbol{\Gamma}) \quad \text{s.t.} \quad \begin{cases} \boldsymbol{\Phi}\boldsymbol{\xi} + \boldsymbol{\epsilon} = \boldsymbol{y} & \text{(noisy observations);} \\ \boldsymbol{\ell} \leq \boldsymbol{\Lambda}\boldsymbol{\xi} \leq \boldsymbol{u} & \text{(linear inequality constraints).} \end{cases} \tag{20}$$

As stated in Sect. 3.2, the performance of samplers depends only on the number of knots N, and therefore, for a large number of observations, it is possible to keep N much smaller than n. One can follow the procedure in Sect. 3.2 and Algorithm 1 to sample from the posterior distribution (20) as well as to get the MAP and mAP estimates:

$$\begin{aligned} M^N(x) &:= \mu_0^* + \sum_{j=1}^{N} \mu_j^* \phi_j(x) = [1, \boldsymbol{\phi}(x)^\top]\boldsymbol{\mu}^*; \\ m^N(x) &:= \mathrm{E}\left[Y^N(x) \middle| Y^N(\mathbb{X}) + \boldsymbol{\epsilon} = \boldsymbol{y}, Y^N \in \mathcal{C}\right] = [1, \boldsymbol{\phi}(x)^\top]\boldsymbol{\mu}_c, \end{aligned} \tag{21}$$

for any $x \in \mathcal{D}$, where $\boldsymbol{\phi}(x) = [\phi_1(x), \ldots, \phi_N(x)]^\top$, $\boldsymbol{\mu}^* = [\mu_0^*, \ldots, \mu_N^*]^\top := \arg\min_{\boldsymbol{z} \in \mathbb{R}^{N+1}} \{\boldsymbol{z}^\top \boldsymbol{\Gamma}^{-1} \boldsymbol{z} | \boldsymbol{\Phi}\boldsymbol{z} + \boldsymbol{\epsilon} = \boldsymbol{y}, \boldsymbol{\ell} \leq \boldsymbol{\Lambda}\boldsymbol{z} \leq \boldsymbol{u}\}$ is the posterior mode computed as in (17) and $\boldsymbol{\mu}_c := \mathrm{E}\left[\boldsymbol{\xi} | \boldsymbol{\Phi}\boldsymbol{\xi} + \boldsymbol{\epsilon} = \boldsymbol{y}, \boldsymbol{\ell} \leq \boldsymbol{\Lambda}\boldsymbol{\xi} \leq \boldsymbol{u}\right]$ is the posterior mean which is computed from simulations.

Remark 5 (Model (M_h) versus Model (M_ϕ))**.** *Before summarizing the differences between both models (M_h) and (M_ϕ), let us note that they provide similar results in terms of prediction accuracy, particularly when using the MAP estimate.*

- *Model (M_h) is more flexible than Model (M_ϕ). Indeed, with Model (M_h) one can incorporate different type of constraints like monotonicity, convexity and boundedness in convex or non-convex regions. Additionally, these constraints can be incorporated either together or sequentially. Furthermore, Model (M_h) is generalized to multidimensional cases [22].*

- *With Model (M_ϕ), the monotonicity (nondecreasing) constraints are equivalent to the nonnegativity of the basis coefficients $\{\xi_j\}$, which can be an advantage for the sampling procedure. For example, the authors in [34] proposed an efficient algorithm for sampling a Gaussian vector truncated on the positive orthant. Additionally, Model (M_ϕ) provides smoother sample paths compared to Model (M_h), which only provides continuous sample paths. Furthermore, as shown in Proposition 3, the slope of the sample paths can be controlled. Model (M_ϕ) allows the imposition of multiple constraints, such as monotonicity, convexity, negativity/positivity, and bounded slope constraints, leading to improved prediction accuracy and less restricted credible intervals. In terms of prediction accuracy, both models provide similar results under the same parameters, but Model (M_ϕ) has more stability (see Figure 13).*

4.2 C^p approximation, $p \geq 2$ with Model (M_ψ)

The sample paths generated from Model (M_ϕ) are differentiable. This means that the derivatives of order zero and one are continuous. This is because the basis functions $\{\phi_j\}$ are the primitive of the *hat* functions $\{h_j\}$ which are only continuous. The differentiability of the sample paths generated

from (M_ϕ) can be generalized to any class C^p, $p \geq 1$. For example, to obtain sample paths that are twice differentiable, it is sufficient to define a *hat* basis function that is differentiable, as follows:

$$\kappa_j(x) = \kappa(x - t_j/\Delta_N) \quad \text{and} \quad \psi_j(x) = \int_0^x \kappa_j(t)\mathrm{d}t, \text{ where} \tag{22}$$

$$\kappa(x) = \begin{cases} -2x^3 - 3x^2 + 1 & \text{if } x \in [-1, 0]; \\ 2x^3 - 3x^2 + 1 & \text{if } x \in (0, 1]; \\ 0 & \text{otherwise.} \end{cases} \tag{23}$$

The function κ can be seen as the cubic Hermite spline defined on $[-1, 1]$. It is clear that κ is a differentiable function on $\mathbb{R}$ and that $\{\kappa_j\}$ is also differentiable. This implies that $\{\psi_j\}$ are twice differentiable. Additionally, $\kappa_j(t_l) = \delta_{j,l}$, where $\delta_{j,l}$ is the Kronecker's delta function equal to one if $j = l$ and zero otherwise. Furthermore, the 'new' *hat* functions $\{\kappa_j\}$ admit the following nice property $\sum_{j=1}^N \kappa_j(x) = 1$, for any $x \in \mathcal{D}$. This property plays an important role in the bounded slope constraints that are added to the proposed approach, as well as in the convexity constraints (see Proposition 6 below). Following the strategy of Sect. 4, any differentiable function f can be approximated by

$$\tilde{f}_N(x) = f(0) + \sum_{j=1}^N f'(t_j)\psi_j(x), \tag{24}$$

for any $x \in \mathcal{D}$.

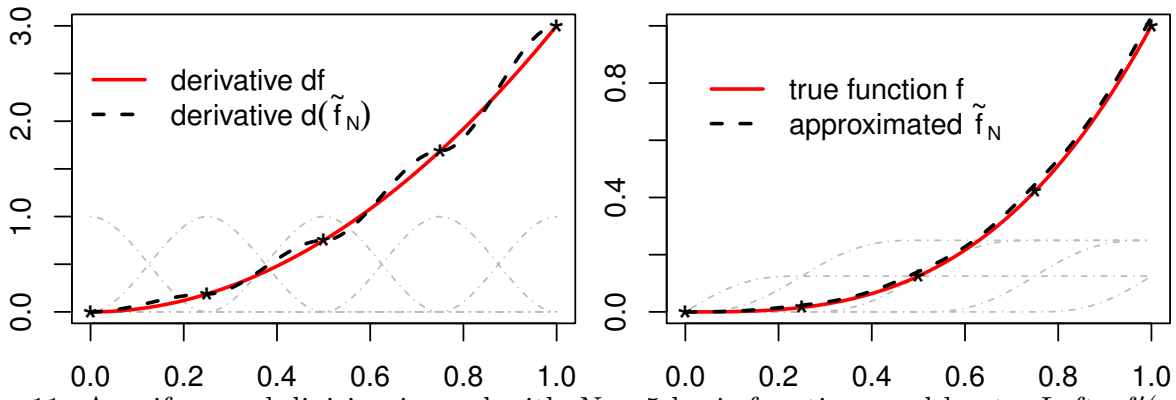


Figure 11: A uniform subdivision is used with $N = 5$ basis functions and knots. Left: $f'(x) = 3x^2$ (red solid curve) and $\tilde{f}'_N$ (black dashed curve) together with the 'new' *hat* functions $\{\kappa_j\}$ (gray curves). Right: the monotone nondecreasing function $f(x) = x^3$ (red solid curve) and the proposed approach $\tilde{f}_N$ in (24) represented by the black dashed curve, together with the 'new' basis functions $\{\psi_j\}$ (gray curves).

In Figure 11, a uniform subdivision of the domain $\mathcal{D} = [0, 1]$ is used with $N = 5$ basis functions and knots. The non-uniform subdivision case is straightforward. In the left panel, we show the 'new' *hat* functions $\{\kappa_j\}$ as well as the derivative function $f'(x) = 3x^2$ approximated by $\tilde{f}'_N(x) = \sum_{j=1}^N f'(t_j)\kappa_j(x)$. The right panel shows the deterministic function $f(x) = x^3$ (red solid curve), which verifies monotonicity (nondecreasing) constraints and the proposed approach $\tilde{f}_N(x) = f(0) + \sum_{j=1}^N f'(t_j)\psi_j(x)$ (black dashed curve). The gray curves are the basis functions $\{\psi_j\}$ defined in (22). The use of the basis functions $\{\psi_j\}$ as in (M_ψ) leads to differentiable GP sample paths of class C^2. The slope of the sample paths can also be controlled, as show in Proposition 6.

Proposition 5 (Bounded slope C^2)**.** *Let f be a continuous and differentiable function on $\mathcal{D}$ over $\mathbb{R}$, i.e., $f \in C^1(\mathcal{D}, \mathbb{R})$, and $\tilde{f}_N(x) := f(0) + \sum_{j=1}^N f'(t_j)\psi_j(x)$, for any $x \in \mathcal{D}$. Then*

$$\tilde{f}'_N(x) \in [\ell, u], \ \forall x \in \mathcal{D} \quad \Leftrightarrow \quad f'(t_j) \in [\ell, u], \ \forall j \in \{1, \ldots, N\},$$

where ℓ and u are the lower and upper bounds respectively.

Proof. In one hand, if $\tilde{f}'_N(x) \in [\ell, u]$ for any $x \in \mathcal{D}$, then, in particular $f'(t_j) = \tilde{f}'_N(t_j) \in [\ell, u]$. In the other hand, if $f'(t_j) \in [\ell, u]$ for all $j \in \{1, \ldots, N\}$, then, for any $x \in \mathcal{D}$ their exists $l \in \{1, \ldots, N-1\}$ such that $x \in [t_l, t_{l+1}]$. Therefore,

$$\tilde{f}'_N(x) = \kappa_l(x) f'(t_l) + \kappa_{l+1}(x) f'(t_{l+1}).$$

This is because $\kappa_j(x) = 0$ for any $j \neq \{l, l+1\}$. Thus, the prove of the upper bound is as follows:

$$\tilde{f}'_N(x) = \kappa_l(x) f'(t_l) + \kappa_{l+1}(x) f'(t_{l+1}) \leq (\kappa_l(x) + \kappa_{l+1}(x))u = u.$$

The lower bound is done similarly, which completes the proof of the proposition. □

In that case, the proposed approach is defined as follows:

$$Y^N(x) := Y(0) + \sum_{j=1}^{N} Y'(t_j)\psi_j(x) = \xi_0 + \sum_{j=1}^{N} \xi_j \psi_j(x), \quad \forall x \in \mathcal{D}, \tag{M_ψ}$$

where, the basis functions $\{\psi_j\}$ are defined in (22), $\xi_0 = Y(0)$ and $\xi_j = Y'(t_j)$, for any $j \in \{1, \ldots, N\}$.

Proposition 6 (Multiple constraints (M_ψ))**.** *If Y^N is defined as in (M_ψ), then*

- *the sample paths generated from Y^N are twice differentiable.*
- *Monotonicity: $Y^N \in \mathcal{C}_m$ if and only if $\{\xi_j\}$ are nonnegative, for any $j \in \{1, \ldots, N\}$.*
- *Monotonicity and bounded slope: $Y^N \in \mathcal{C}_m$ and $(Y^N)'(x) \in [\ell, u]$, for any $x \in \mathcal{D}$ if and only if $\{\xi_j\}$ are in $[\ell, u]$, for any $j \in \{1, \ldots, N\}$, where $\ell, u \in \mathbb{R}_+$ are the lower and upper bounds respectively.*
- *Monotonicity and boundedness: Y^N is nondecreasing and nonnegative if and only if $\{\xi_j\}$ are nonnegative, for any $j \in \{0, \ldots, N\}$.*
- *Convexity: $Y^N \in \mathcal{C}_c$ if and only if $\xi_j \leq \xi_{j+1}$, for any $j \in \{1, \ldots, N-1\}$.*
- *Monotonicity and convexity: $Y^N \in \mathcal{C}_m \cap \mathcal{C}_c$ if and only if $0 \leq \xi_1 \leq \ldots \leq \xi_N$.*
- *Monotonicity, convexity and boundedness: Y^N is nondecreasing, convex, and nonnegative if and only if $0 \leq \xi_1 \leq \ldots \leq \xi_N$ and $\xi_0 \geq 0$.*

Proof. The first item is done by the fact that the basis functions $\{\psi_j\}$ are twice differentiable. For the second one, if Y^N is nondecreasing, then $(Y^N)'(x) \geq 0$ for any $x \in \mathcal{D}$. In particular $(Y^N)'(t_j) = \xi_j \geq 0$ for any $j \in \{1, \ldots, N\}$. Now, if $\xi_j \geq 0$ for any $j \in \{1, \ldots, N\}$, then, Y^N is nondecreasing since the basis functions $\{\psi_j\}$ are nondecreasing too. For the third one, if Y^N is nondecreasing and nonnegative then $\{\xi_j\}$ are nonnegative for any $j \in \{1, \ldots, N\}$ and in particular $\xi_0 = Y^N(0) \geq 0$. Now, if $\{\xi_j\}$ are nonnegative for any $j \in \{0, \ldots, N\}$, then Y^N is nondecreasing and nonnegative since the basis functions are nonnegative. The fourth one is a simple consequence of Proposition 5. For the fifth one, in one hand if Y^N is convex then $(Y^N)'$ is nondecreasing. In particular, $(Y^N)'(t_j)$ is nondecreasing which implies that $\xi_j \leq \xi_{j+1}$, for any $j \in \{1, \ldots, N-1\}$. In the other hand, if $\xi_j \leq \xi_{j+1}$ for any $j \in \{1, \ldots, N-1\}$, then it is sufficient to prove that $(Y^N)'$ is nondecreasing. Two cases should be verified. The first one, their exists $j < l$ such that $x \in [t_j, t_{j+1}]$

and $x' \in [t_l, t_{l+1}]$. This is the simple case since $(Y^N)'(x) \in [\xi_j, \xi_{j+1}]$ and $(Y^N)'(x') \in [\xi_l, \xi_{l+1}]$. Indeed,

$$\xi_j = \xi_j \underbrace{(\kappa_j(x) + \kappa_{j+1}(x))}_{=1} \leq \xi_j\kappa_j(x) + \xi_{j+1}\kappa_{j+1}(x) = (Y^N)'(x) \leq \xi_{j+1}.$$

The second one, their exists $j \in \{1, \ldots, N-1\}$ such that $x \leq x' \in [t_j, t_{j+1}]$. In that case, we have

$$\begin{aligned}
(Y^N)'(x') - (Y^N)'(x) &= \xi_j\kappa_j(x') + \xi_{j+1}\kappa_{j+1}(x') - (\xi_j\kappa_j(x) + \xi_{j+1}\kappa_{j+1}(x)) \\
&= \xi_j(\kappa_j(x') - \kappa_j(x)) + \xi_{j+1}(\kappa_{j+1}(x') - \kappa_{j+1}(x)) \\
&= \xi_j(\kappa_{j+1}(x) - \kappa_{j+1}(x')) - \xi_{j+1}(\kappa_{j+1}(x) - \kappa_{j+1}(x')) \\
&= \underbrace{(\kappa_{j+1}(x) - \kappa_{j+1}(x'))}_{\leq 0}\underbrace{(\xi_j - \xi_{j+1})}_{<0} \geq 0.
\end{aligned}$$

The sixth is a simple consequence of the previous two items. For the last one, it is enough to prove that if $0 \leq \xi_1 \leq \ldots \leq \xi_N$ and $\xi_0 \geq 0$, then $Y^N \in \mathcal{C}_m \cap \mathcal{C}_c$ and is nonnegative as the basis functions are nonnegative. Conversely, if Y^N is nondecreasing, convex, and nonnegative, then $0 \leq \xi_1 \leq \ldots \leq \xi_N$ and, in particular, $\xi_0 = Y^N(0) \geq 0$. □

Remark 6. *In this section, the smoothness of the sample paths of Model (M_ϕ) has been investigated, with a focus on its differentiability in class C^p, for $p \geq 1$. The function κ, as given in (23), has a differentiability of class C^1. This implies that the sample paths from (M_ψ) are twice differentiable. For example, to obtain sample paths that are differentiable up to order three (i.e., class C^3), it is sufficient to define κ as follows*

$$\kappa(x) = \begin{cases} -3x^4 - 8x^3 - 6x^2 + 1 & \text{if } x \in [-1, 0]; \\ -3x^4 + 8x^3 - 6x^2 + 1 & \text{if } x \in (0, 1]; \\ 0 & \text{otherwise.} \end{cases} \tag{25}$$

It is straightforward to verify that κ is a differentiable function of class C^2. Adopting the same approach, we define κ_j and ψ_j as follows:

$$\kappa_j(x) = \kappa(x - t_j/\Delta_N) \quad \text{and} \quad \psi_j(x) = \int_0^x \kappa_j(t)\mathrm{d}t.$$

Finally, let us recall that the functions $\{\kappa_j\}$ verify the following two fundamental properties:

$$\begin{aligned}
&\kappa_j(t_l) = \delta_{j,l}, \quad \forall j, l = 1, \ldots, N; \\
&\sum_{j=1}^N \kappa_j(x) = 1, \quad \forall x \in \mathcal{D},
\end{aligned}$$

where we recall that $\delta_{j,l}$ is the Kronecker's delta function equal to one if $j = l$ and zero otherwise.

4.3 C^p approximation, $p \geq 2$ with Model (M_φ)

In this section, we first consider the convexity constraint for continuous and twice differentiable functions $f \in C^2$. Thus, the convex set $\mathcal{C}_c$ is given by

$$\mathcal{C}_c = \left\{ f \in C^2(\mathcal{D}, \mathbb{R}) \text{ s.t. } f''(x) \geq 0, \ x \in \mathcal{D} \right\},$$

where f'' represents the second-order derivative of f. As stated in [22], any at least twice differentiable function f can be written as $f(x) = f(0) + xf'(0) + \int_0^x \int_0^t f''(u)\mathrm{d}u\mathrm{d}t$. Following the strategy of Sect. 3.1, any twice differentiable function f can be approximated by

$$\tilde{f}_N(x) = f(0) + xf'(0) + \sum_{j=1}^N f''(t_j)\varphi_j(x) \tag{26}$$

for any $x \in \mathcal{D}$, where $\{\varphi_j\}$ are given in (4). As in Sect. 3.1 and 4, we recall the following result:

Lemma 3 (Uniform convergence C^2)**.** *For any $f \in C^2(\mathcal{D}, \mathbb{R})$, the function $\tilde{f}_N$ defined in* (26) *converges uniformly to f when N tends to infinity.*

Proof. The proof is similar to the one given in Lemma 2. □

In that case, we consider the following model

$$Y^N(x) := Y(0) + Y'(0)x + \sum_{j=1}^{N} Y''(t_j)\varphi_j(x) = \xi_0^* + \xi_0 x + \sum_{j=1}^{N} \xi_j \varphi_j(x), \quad x \in \mathcal{D}, \qquad (M_\varphi)$$

where we denote by $\xi_0^* = Y(0)$, $\xi_0 = Y'(0)$, and $\xi_j = Y''(t_j)$, for any $j \in \{1, \ldots, N\}$. In this case, Y^N is convex (resp. concave) on $\mathcal{D}$ *if and only if* $\{\xi_j\}$ are nonnegative (resp. nonpositive) for any $j \in \{1, \ldots, N\}$.

Proposition 7 (Multiple constraints (M_φ))**.** *If Y^N is defined as in (M_φ), then*

- *Monotonicity and convexity: $Y^N \in \mathcal{C}_m \cap \mathcal{C}_c$ if and only if $\xi_j \geq 0$ for any $j \in \{0, \ldots, N\}$.*
- *Monotonicity, convexity and boundedness: $Y^N \in \mathcal{C}_m \cap \mathcal{C}_c$ and Y^N is nonnegative if and only if $\xi_0^* \geq 0$ and $\xi_j \geq 0$, for any $j \in \{0, \ldots, N\}$.*

Proof. On one hand, if $Y^N \in \mathcal{C}_m \cap \mathcal{C}_c$, then for any $j \in \{1, \ldots, N\}$, $\{\xi_j\}$ are nonnegative and $(Y^N)'(0) = \xi_0 \geq 0$. On the other hand, if $\{\xi_j\}$ are nonnegative for any $j \in \{0, \ldots, N\}$, then for any $x \in \mathcal{D}$, $(Y^N)''(x) = \sum_{j=1}^N \xi_j h_j(x) \geq 0$ and $(Y^N)'(x) = \xi_0 + \sum_{j=1}^N \xi_j \phi_j(x) \geq 0$ since $\phi_j(x) \geq 0$ for any $x \in \mathcal{D}$. The second item is obvious. □

Remark 7. *The sample paths generated from Model (M_φ) are twice differentiable. As in Sect. 4.2, the smoothness of the sample paths can be generalized to class C^p, for any $p \geq 2$ by defining smoother hat functions like ones given in* (22) *and* (25)*.*

Performance illustrations of Model (M_φ)

The aim is to illustrate Model (M_φ) and to demonstrate the superior prediction accuracy of the MAP estimate over the mAP estimate.

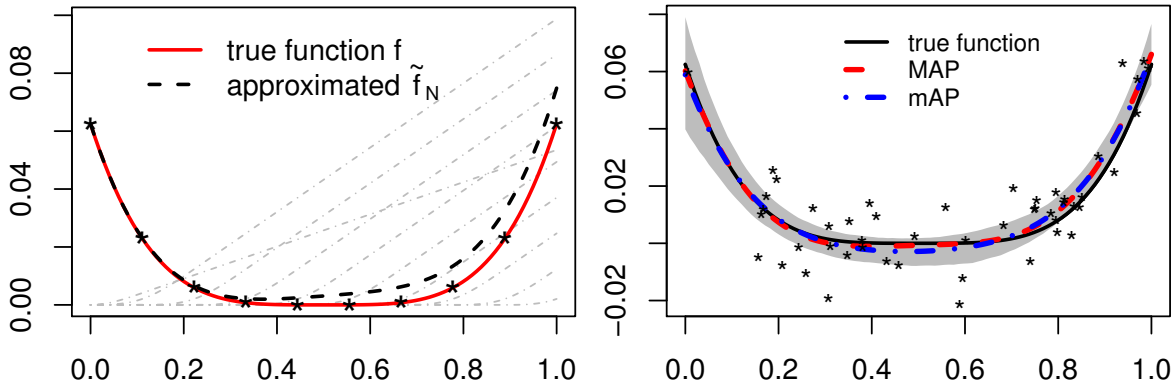


Figure 12: Left: The approximation of the convex function $f_3(x) = (x - 0.5)^4$ is shown by the red solid curve, while the proposed approach $\tilde{f}_N$ is represented by the black dashed curve, along with the basis functions $\{\varphi_j\}$ shown in gray curves. The approximation is obtained using a uniform subdivision with $N = 10$ basis functions and knots. Right: The GP approximation obtained from Model (M_φ) under convexity constraints is shown.

Figure 12 (left) shows the true convex function $f_3(x) = (x-0.5)^4$ (red solid curve), approximated by the proposed approach $\tilde{f}_N$ (26) (black dashed curve) using a uniform subdivision with $N = 10$

basis functions. The black stars represent the values of the true function f_3 at the knots $\{t_j\}$. Figure 12 (right) shows the GP approximation from Model (M_φ) with convexity constraints. We fix $N = 30$ and use the SE covariance function (see Table 1), with a nugget effect of order 10^{-8}. The black stars represent the 50 training data generated from (1) using f_3 and a true $\sigma_{\text{noise}} = 0.01$. To sample from the posterior distribution of the coefficients $\{\xi_j\}$ as in Algorithm 1, we use the efficient HMC technique developed in [30].

To rigorously compare the MAP and mAP estimates in terms of prediction accuracy, we propose to generate a dataset of size $n = 500$ from (1) with f_3 and a true $\sigma_{\text{noise}} = 0.01$. This dataset is randomly split into training set of size 300 and testing set of size 200. We propose to place a uniform prior distribution on the correlation length parameter $\theta \sim \mathcal{U}(0.1, 1)$. In that case, we get an average mean squared prediction error (MSPE) over 1,000 replicates equal to 1.85×10^{-3} when using the MAP estimate, and equal to 1.92×10^{-3} when using the mAP estimate. Let us mention that the mAP estimate is computed by averaging 1,000 samples generated from the efficient HMC technique.

4.4 Comparison between Models (M_h), (M_ϕ) and (M_φ)

In this section, a comparison between Models (M_h), (M_ϕ) and (M_φ) in terms of prediction accuracy is investigated. To do this, we consider the monotone nondecreasing function f_{m_2} given in (27) and the convex function $f_3(x) = (x - 0.5)^4$ (shown in Figure 12).

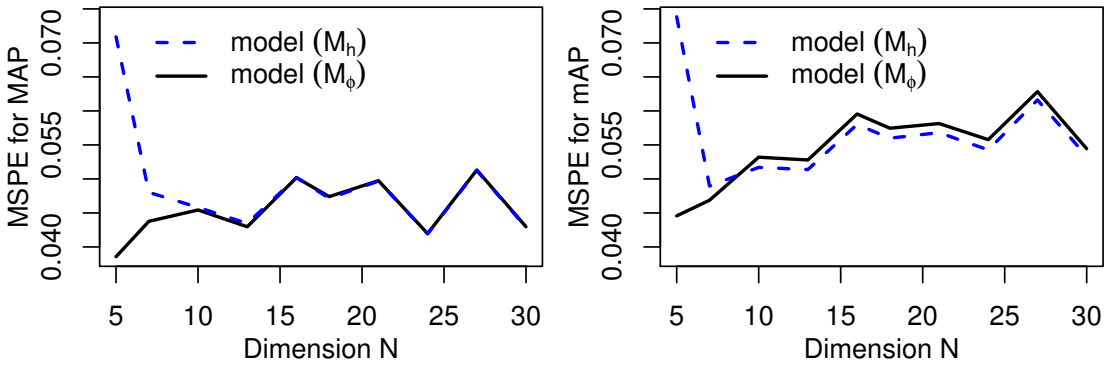


Figure 13: The average MSPE over 25 replicates as a function of the dimension N is shown using the MAP (resp. mAP) estimate for Models (M_h) and (M_ϕ) in the left (resp. right) panel.

Figure 13 shows the average MSPE over 25 replicates as a function of the dimension N. A dataset of size $n = 500$ generated from (1) using f_{m_2} and a true $\sigma_{\text{noise}} = 0.5$ is randomly split into training set of size 300 and testing set of size 200. The Matérn covariance function with $\nu = 5/2$ and $\theta = 0.5$ has been used. The left panel shows results using the MAP estimate, while the right panel shows results using the mAP estimate. We observe that the average MSPE obtained with the MAP estimate is smaller than that obtained with the mAP estimate. The difference between the two models is more pronounced when N is small, but becomes less significant as N increases. For N around 10, both models provide approximately the same MSPE.

In Figure 14, the three Models (M_h) ,(M_ϕ) and (M_φ) are compared in terms of MSPE. A dataset of size 500 generated randomly from (1) with f_3 and a true $\sigma_{\text{noise}} = 0.01$ is split randomly into 300 training samples and 200 testing samples. The SE covariance function has been used with correlation length parameter θ fixed at 0.5. As for the monotonicity case (Figure 13), we observe that the average MSPE obtained with the MAP estimate is smaller than that obtained with the mAP estimate. The difference between the three models is more pronounced when N is small, but becomes less significant as N increases. As expected, Model (M_φ) provides the least average MSPE for the MAP estimate.

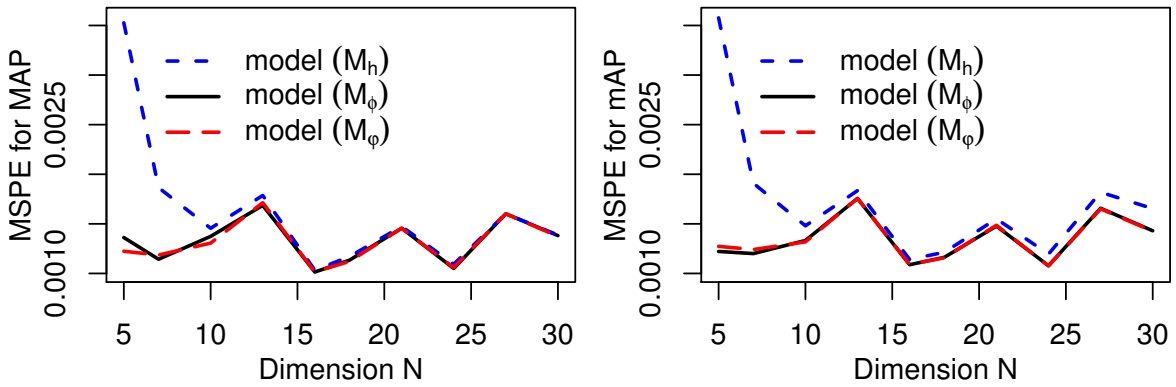


Figure 14: The average MSPE over 25 replicates as a function of the dimension N is shown using the MAP (resp. mAP) estimate for Models (M_h), (M_ϕ), and (M_φ) in the left (resp. right) panel.

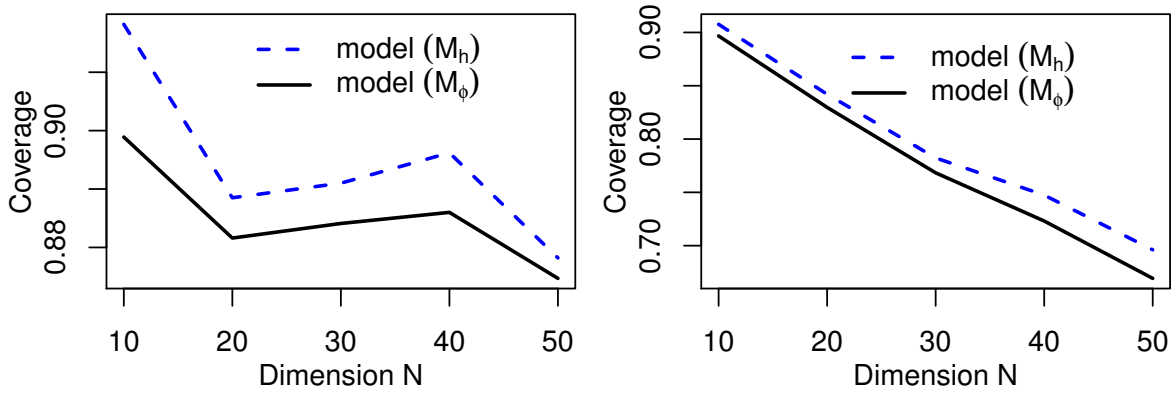


Figure 15: The average 95% posterior coverage over 25 replicates as a function of the dimension N is shown using the Matérn $\nu = 5/2$ (resp. Matérn $\nu = 3/2$) for Models (M_h) and (M_ϕ) in the left (resp. right) panel.

Finally, in Figure 15, the average 95% posterior coverage (the estimated value should include a 95% chance that the true value lies within it) over 25 replicates for Models (M_h) and (M_ϕ) is displayed as a function of the dimension N. As before, the monotone nondecreasing function f_{m_2} given in (27) is used as a test function which presents a challenge situation since this function is approximately flat on $[0.7, 1]$. The Matérn covariance function with $\nu = 5/2$ is applied in the left panel, and $\nu = 3/2$ is applied in the right panel. As expected, Model (M_h) performs better than Model (M_ϕ) in terms of posterior coverage for both scenarios, as the corresponding coverage is closer to 95%. This is related to the smoothness of the sample paths generated by both models, which confirms the frequentist coverage results presented in [43]. We also observe that, on average, the Matérn covariance function with a smoothness parameter of $\nu = 5/2$ provides posterior coverage closer to 95% than that provided by the Matérn covariance function with $\nu = 3/2$. Additionally, both cases show a decreasing posterior coverage as the dimension N of the basis coefficients $\{\xi_j\}$ increases. This confirms the result from [45], which demonstrated that the *mass-shifting* phenomenon is more pronounced in high-dimensional truncated vectors.

4.5 Performance of the MAP estimate

The aim of this section is to show the performance of the MAP estimate in terms of prediction accuracy in different situations. A variety of functions is considered:

$$f_{m_1}(x) = (5x-3)^3 \mathbf{1}_{[0.6,1]}(x), \qquad f_{m_2}(x) = \frac{3}{1+\exp\{-10x+2.1\}},$$
$$f_{m_3}(x) = \sqrt{2}\sum_{j=1}^{100} j^{-1.7}\sin(j) \times \cos(\pi(j-0.5)(1-x)), \qquad f_{m_4}(x) = 5x^2, \tag{27}$$

for $x \in [0,1]$. The function f_{m_1} is monotone (nondecreasing) and flat on $[0, 0.6]$. However, f_{m_2} and f_{m_3} are approximately flat on $[0.7, 1]$. The last function f_{m_4} is an increasing function on the whole domain $[0,1]$. Let us mention that only the function f_{m_3} is decreasing in certain regions which allows us to evaluate the performance of the proposed approach under slight model misspecification.

functions	methods	MSPE (total)	MSPE (flat)	MSPE (increasing)
f_{m_1}	MAP	13.55 (4.97)	**4.39** (1.87)	19.27 (7.59)
	DGL	**11.36** (2.62)	8.13 (1.95)	**14.71** (4.79)
	IGL	13.44 (2.62)	9.86 (1.70)	17.32 (5.23)
	TMVN	65.63 (7.21)	14.53(2.59)	102.6 (11.16)
f_{m_2}	MAP	**7.46** (1.65)	**3.94** (2.31)	8.63 (1.99)
	DGL	8.29 (1.78)	7.13 (2.64)	8.56 (2.32)
	IGL	9.55 (1.92)	8.40 (2.61)	9.84 (2.54)
	TMVN	8.32 (2.11)	8.61 (2.91)	**7.94** (2.75)
f_{m_3}	MAP	7.84 (1.47)	**5.23** (2.23)	8.78 (1.79)
	DGL	7.76 (1.74)	9.16 (2.9)	**6.87** (1.87)
	IGL	**7.72** (1.74)	8.57 (2.45)	7.18 (1.74)
	TMVN	11.36 (1.33)	15.27 (2.85)	8.97 (1.76)
f_{m_4}	MAP	9.44 (1.89)	-	9.44 (1.89)
	DGL	8.67 (2.15)	-	8.67 (2.15)
	IGL	9.34 (2.16)	-	9.34 (2.16)
	tMVN	**5.68** (1.61)	-	**5.68** (1.61)

Table 2: The average of the MSPE $\times 10^2$ (standard deviation $\times 10^2$) over one thousand replicates for different functions and methods. Model (M_h) has been used to compute the MAP estimate.

The simulation studies are based on a dataset of size $n = 500$ generated from (1) using the true functions (27) and a true $\sigma_{\text{noise}} = 0.5$. The dataset is randomly split into training set of size 300 and testing set of size 200. Table 2 summarizes the average of the MSPE $\times 10^2$ (standard deviation $\times 10^2$) over one thousand replicates for the four true functions (27) using different approaches. To evaluate the performance between the flat and increasing regions separately, we additionally report the average partial MSPEs for each region: MSPE (flat) for the flat portion and MSPE (increasing) for the increasing portion, in addition to the overall average MSPE. To avoid overfitting, we set $N = \lfloor n_{tr}/8 \rfloor$ as in [45], where n_{tr} is the number of training samples fixed at 300. In that case, the MAP estimate provides the same MSPE results when using different models (see Section 4.4).

The Matérn family of covariance functions is used with smoothness parameter $\nu \sim \mathcal{U}(0.5, 1)$ and correlation length parameter $\theta \sim \mathcal{U}(0.1, 1)$ generated at each replicate as in [45]. For flat regions, the MAP estimate outperformed the shrinkage approaches of [45] (IGL,DGL and TMVN). This confirms the robustness of the MAP estimate for capturing flat regions. According to the MSPE criterion, the MAP estimate is twice (resp. three times) more efficient than IGL and DGL (resp. TMVN)

when using f_{m_2} over the flat region. This was also seen when calculating the total MSPE for f_{m_2}, where the proposed approach had a slightly lower standard deviation than the shrinkage approaches (IGL, DGL, and TMVN). This again confirms the stability of the MAP estimate provided by the proposed approach. Let us recall that the MAP estimate is computed from a quadratic optimization problem given by equation (17). Finally, it should be mentioned that the simulation studies are conducted without any additional constraints.

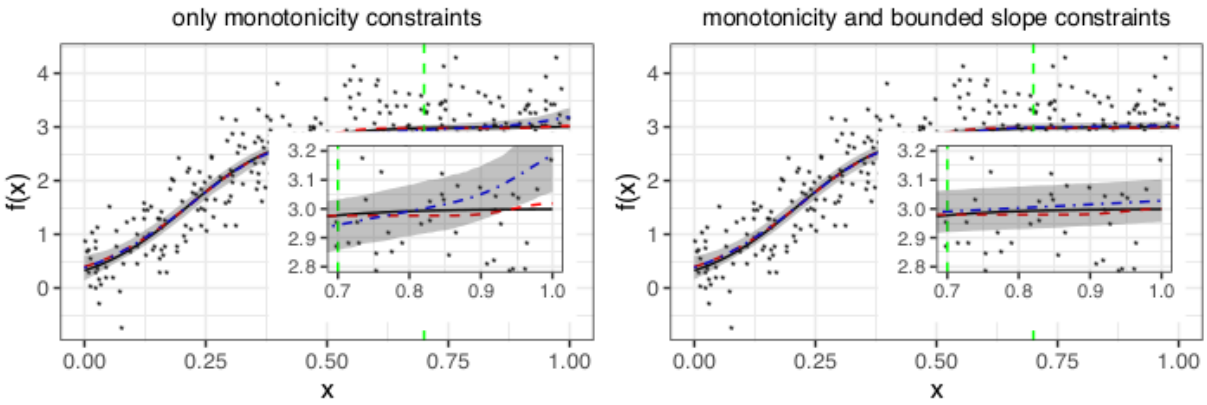


Figure 16: GP approximation from Model (M_ϕ) satisfying monotonicity constraints (left panel), monotonicity and bounded slope constraints (right panel). The panel description is the same as in Figure 3, with zoomed-in inset plots where $x \in [0.7, 1]$. The green vertical dashed line in the right panel represents the starting location of bounded slope constraints.

Figure 16 shows the GP approximation from Model (M_ϕ) with monotonicity constraints only (left panel) and both monotonicity and bounded slope constraints (right panel). The function $f_{m_2}(x) = \frac{3}{1+\exp\{-10x+2.1\}}$ is considered. This function is interesting because it is monotonically increasing and almost flat over the interval $[0.7, 1]$. We used the Matérn covariance function with $\nu = 5/2$ and the efficient HMC technique [30] to sample from the posterior distribution of the basis coefficients $\{\xi_j\}$. The black solid curve represents the function f_{m_2}, while the red dashed (resp. blue dashed-dotted) curve represents the MAP estimate (resp. mAP estimate). The gray shaded area represents the 95% pointwise confidence interval. The black stars are the 300 noisy observations generated from (1) using f_{m_2} and a true noise standard deviation of $\sigma_{\text{noise}} = 0.5$. The green vertical dashed line in the right panel corresponds to the starting point where the bounded slope constraints are imposed. In the right panel, we impose an upper bound slope constraints on the proposed approach in the flat region $[0.7, 1]$. In the left panel, we observe that the pointwise 95% confidence intervals fail to capture the true nondecreasing function f_{m_2} for a substantial part of the input domain. This is due to the *mass-shifting* phenomenon highlighted in [45]. Including bounded slope constraints (right panel) provides smaller and more realistic credible intervals as compared to those without such constraints (left panel). Let us mention that the MAP estimate is robust in both scenarios, with or without bounded slope constraints. This is because the MAP estimate converges to the constrained optimal smoothing function (as proved in [15]). Additionally, the mAP estimate tends towards the MAP estimate when bounded slope constraints are added, as seen in the right panel. Finally, the average 95% posterior coverage over 25 replicates is equal to 78% when using the proposed approach with only monotonicity (nondecreasing) constraints. However, it increases to 89% when adding the bounded slope constraints. Thus, by adding multiple constraints, the prediction accuracy is improved as the posterior coverage is closer to 95%.

5 Real-world data applications

5.1 Light detection and ranging (LiDAR)

In this section, the proposed approach developed in this paper was applied on the light detection and ranging (LiDAR) real-world data that consist of 221 observations from a LiDAR experiment and it contain information on range and logratio. The predictor range represents the distance travelled before the light is reflected back to its source, however, the response variable logratio represents the logarithm of the ratio of received light from two laser sources. This real-world data is available from the R package HRW. The data suggest that the underlying function is nonpositive and monotone nonincreasing with a flat region when the range is less than 550.

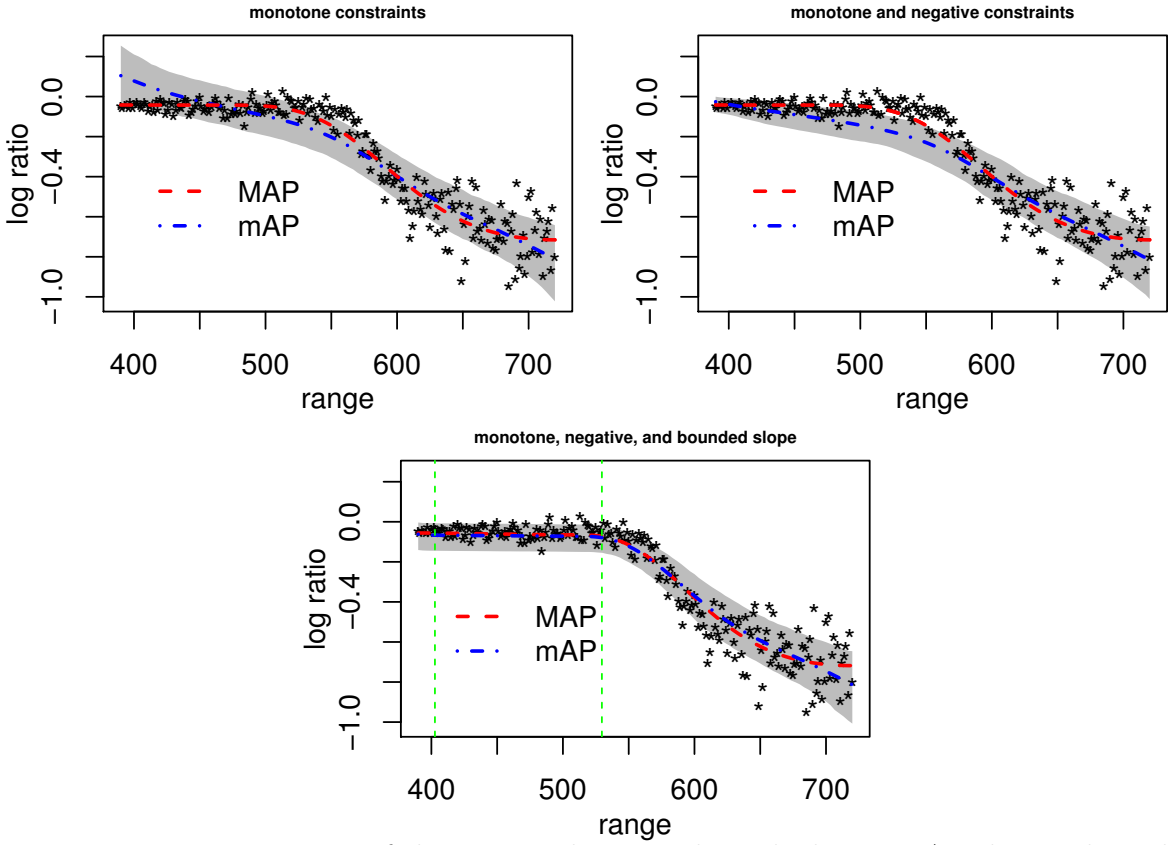


Figure 17: Estimation accuracy of the proposed approach applied on LiDAR data. The red dashed curve corresponds to the MAP estimate, while the blue dashed-dotted corresponds to the mAP estimate. The black stars represent real-world data, and the gray shaded area represents the 95% pointwise confidence interval.

In Figure 17, the proposed approach using Model (M_ϕ) has been applied to the $n = 221$ LiDAR data (shown as black stars). We fix $N = \lfloor n/8 \rfloor$ to avoid overfitting (this choice is justified later in this section), and we use the Matérn covariance functions with $\nu = 5/2$. The red dashed curve represents the MAP estimate, while the blue dashed-dotted curve represents the mAP estimate. The gray shaded area corresponds to the 95% pointwise credible interval. Top left: Model (M_ϕ) satisfying monotonicity constraints only. We observe that, unlike the MAP estimate, the credible interval and mAP estimate fail to follow the behavior of the data in the flat region (range less than 550). To rigorously compare the MAP and mAP estimates in terms of prediction accuracy, we propose to place a uniform prior distribution on the correlation length parameter $\theta \sim \mathcal{U}(50, 300)$ as well as on the noise standard deviation $\sigma_{\text{noise}} \sim \mathcal{U}(0.1, 0.5)$. By randomly splitting the total dataset of size 221 into 80% training and 20% testing datasets, we obtain an average MSPE over

one thousand replicates of 8.23×10^{-2} when using the MAP estimate and 9.76×10^{-2} when using the mAP estimate. Top right: we added nonnegativity constraints (as per Proposition 4) and found that, once again, credible intervals as well as the mAP estimate fail to capture the flat region. Bottom: we added bounded slope constraints in the flat region between the green vertical dashed lines (see Proposition 3). In that case, we observe that both the mAP estimate and the 95% pointwise credible interval follow the observations and capture the flat region. The proposed model with triple inequality constraints seems to align with the data better, specifically over the *flat* region and when logratio starts to decreases. We also observe that the mAP estimate tends towards the MAP estimate, which behaves well in all three situations. As expected, adding multiple constraints improves the prediction accuracy of the proposed approach (MAP and mAP estimates) as well as the behavior of the posterior distribution, which provides more realistic pointwise credible intervals.

For the estimation of the correlation length parameter θ, and the noise standard deviation σ_{noise}, we propose an adjustment of the 5-fold CV technique based on minimizing the MSPE using the MAP as an estimator. The procedure is as follows:

$$(\widehat{\theta}, \widehat{\sigma}_{\text{noise}})_{\text{CV}} = \arg\min \left\{ \sqrt{\frac{1}{n_t} \sum_{i=1}^{n_t} (y_i - \widehat{y}_i)^2} \right\}, \tag{28}$$

where $\widehat{y}_i$ is the value of the MAP estimate M^N at the test point x_i, which depends on σ_{noise} and θ, and n_t is the number of test samples, representing 20% of the total dataset.

Now, we analyze the value of the number of grid points N as a function of the number of samples n. Its value influences the prediction accuracy of the proposed approach. In order to avoid overfitting, it is more reasonable to choose the number of grid points N to be smaller than the number of training samples n. We consider the case where $N \in \{n, \lfloor n/2 \rfloor, \lfloor n/4 \rfloor, \lfloor n/8 \rfloor\}$ to conduct a thorough analysis on our real-world dataset.

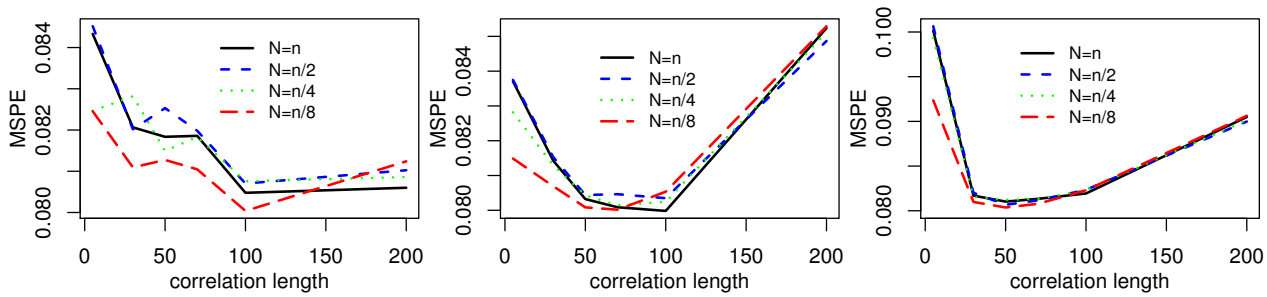


Figure 18: MSPEs as a function of the correlation length parameter θ using 5-fold CV technique for different values of $N \in \{n, \lfloor n/2 \rfloor, \lfloor n/4 \rfloor, \lfloor n/8 \rfloor\}$ and standard deviation $\sigma_{\text{noise}} = 0.1$ (resp., $\sigma_{\text{noise}} = 0.5$ and $\sigma_{\text{noise}} = 1$) in the left (resp., middle and right) panel.

Figure 18 shows the 5-fold CV MSPEs repeated fifty times as a function of the length parameter $\theta \in \{5, 50, 100, 200, 300\}$, with $N \in \{n, \lfloor n/2 \rfloor, \lfloor n/4 \rfloor, \lfloor n/8 \rfloor\}$. We fix $\sigma_{\text{noise}} = 0.1$ (resp., $\sigma_{\text{noise}} = 0.5$ and $\sigma_{\text{noise}} = 1$) in the left (resp., middle and right) panel. First, we observe that in all three situations, the MSPE drops rapidly for small values of the correlation length parameter θ, and then increases for larger values of θ. Second, we also observe that using a smaller number of discretization points, $N = \lfloor n/8 \rfloor$, results in lower MSPEs. Third, we observe that the optimal value of the correlation length parameter θ depends on the noise standard deviation parameter σ_{noise}. A small value of σ_{noise} leads to an optimal MSPE for a large value of θ, and vice versa. These numerical experiments guided us to choose $N = \lfloor n/8 \rfloor$ for the remainder of this study.

For hyperparameter estimation, we use the NLopt optimization tools from [16], specifically the Constrained Optimization BY Linear Approximations (COBYLA) [32] optimizer. This choice

was justified by numerical comparison tests with the optim function in R, which showed that the COBYLA optimizer method provided more accurate results for estimating the correlation length and noise standard deviation parameters θ and σ_{noise}, respectively. Let us mention that a multistart optimization was conducted using ten initial vectors of the correlation length parameter $\theta \in [40, 200]$ and the noise standard deviation parameter $\sigma_{\text{noise}} \in [0.05, 1]$.

5.2 Nuclear safety

In this section, we investigate the performance of the proposed model using real-world data provided by the Institut de Radioprotection et de Sret Nuclaire (IRSN) in France. We studied the nuclear reactor of the uranium sphere known as the Lady Godiva device, located at Los Alamos National Laboratory (LANL) in New Mexico, U.S. The reactor's output increases with respect to two input parameters: its radius, which ranges between 0 and 20 cm, and its density, which ranges between 10 and 20 g/cm^3.

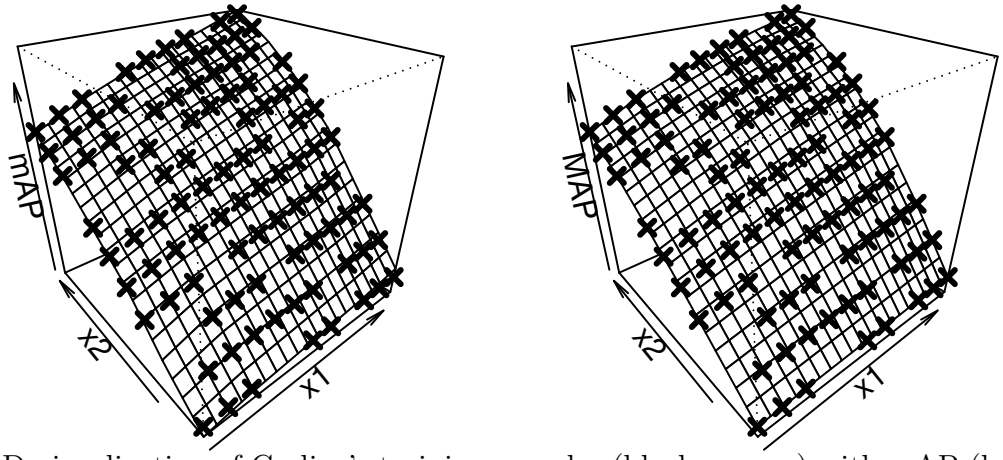


Figure 19: 3D visualization of Godiva's training samples (black crosses) with mAP (left panel) and MAP (right panel) estimates. The noise standard deviation is fixed at $\sigma_{\text{noise}} = 0.1$, and the average MSPE over 100 replicates is 6.99×10^{-2} for mAP and 3.87×10^{-2} for MAP.

We used 121 observations defined on the interval $[0, 20] \times [10, 20]$ to demonstrate the effectiveness of the proposed model in terms of prediction accuracy and to compare the mAP and MAP estimates. Figure 19 shows the application of the GP approximation from Model (18). To avoid overfitting, we set $N_1 = N_2 = \lfloor n/16 \rfloor = 6$, which results in 36 knots and basis functions. Later in this section, we will provide justification for this choice of grid discretization. The two-dimensional SE covariance function (19) is used, with a nugget effect of order 10^{-12}. We randomly split the real-world dataset, which has a size of 121, into 80% training set and 20% testing set. We fix the noise standard deviation at $\sigma_{\text{noise}} = 0.1$. The left panel displays the mAP estimate based on 5,000 sample paths generated using the efficient HMC sampler along with the training samples (represented by black crosses), while the right panel shows the MAP estimate, also with the training samples represented by black crosses. To rigorously compare the mAP and MAP estimates, we propose to place a prior on the correlation length parameters: $\theta_1 \sim \mathcal{U}(1, 20)$ and $\theta_2 \sim \mathcal{U}(1, 20)$, as well as on the noise standard deviation: $\sigma_{\text{noise}} \sim \mathcal{U}(0.1, 1)$. The numerical experiment is conducted within 1,000 replicates. The given real-world dataset of size 121 is split randomly into 80% training and 20% testing datasets. In that case, the average MSPE over one hundred replicates is 3.87×10^{-2} when using the MAP estimate, and 6.99×10^{-2} when using the mAP estimate.

Now we investigate the choice of the number of discretization points N_1 and N_2 in (18). To conduct a thorough analysis on our real-world dataset, we consider the case where $N = N_1 = N_2$ are selected from the grid $\{\lfloor n/4 \rfloor, \lfloor n/8 \rfloor, \lfloor n/16 \rfloor\}$.

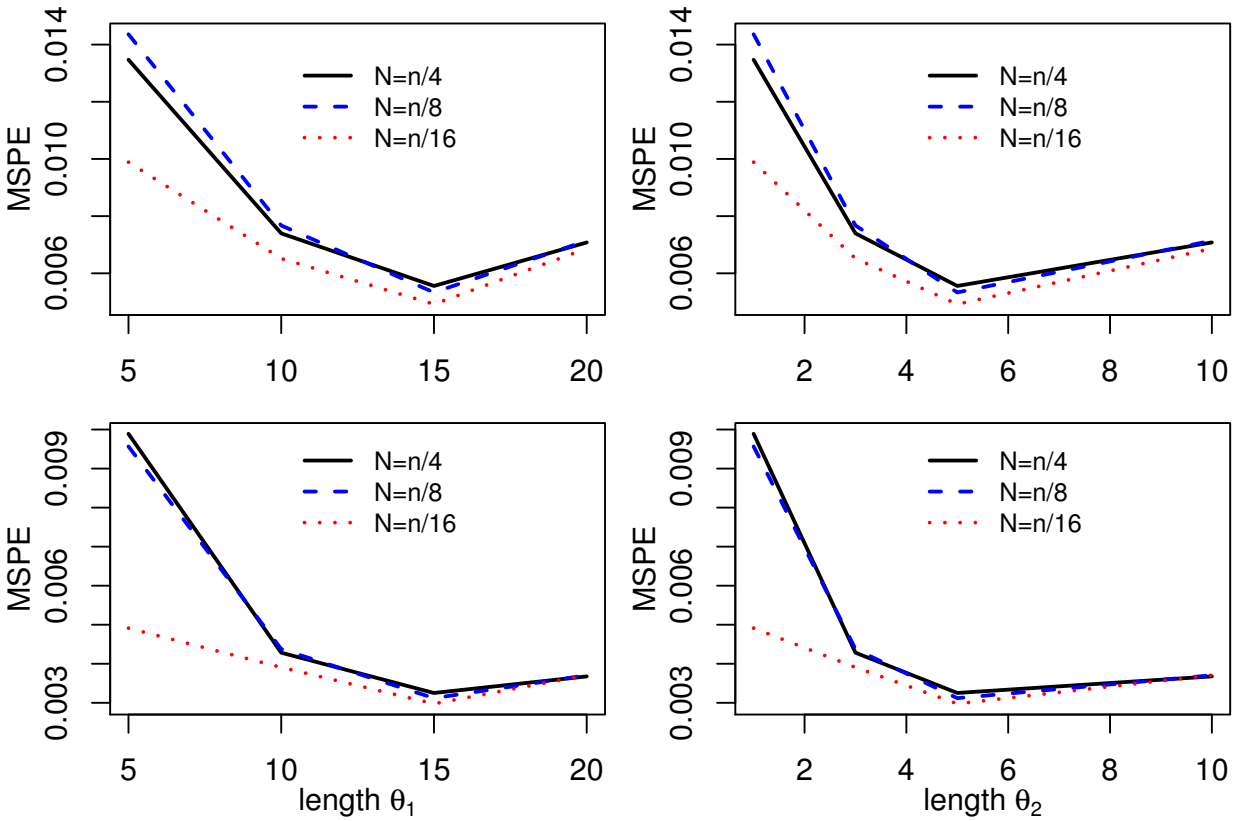


Figure 20: MSPE as a function of the correlation length parameters θ_1 (left) and θ_2 (right), using the 5-fold CV technique for different values of $N \in \{\lfloor n/4 \rfloor, \lfloor n/8 \rfloor, \lfloor n/16 \rfloor\}$. The noise standard deviation is fixed at 0.1 (top) and at 0.04 (bottom).

Figure 20 shows the 5-fold CV MSPEs as a function of the length parameters $\theta_1 \in \{5, 10, 15, 20\}$ in the left panel and $\theta_2 \in \{1, 3, 5, 10\}$ in the right panel, with σ_{noise} fixed at 0.1 in the top plot and at 0.04 in the bottom plot. First, we observe that in all four situations, the MSPE drops quickly for small values of the correlation length parameters θ_1 and θ_2, and then increases for larger values ($\theta_1 \geq 15$ and $\theta_2 \geq 5$). Second, we observe that a smaller number of discretization points, $N = \lfloor n/16 \rfloor$, provides lower MSPEs. These analyses guided us to choose $N_1 = N_2 = \lfloor n/16 \rfloor$. Third, the optimal values of the correlation length parameters (θ_1, θ_2) are around $(15, 5)$ for both small and large noise standard deviations, σ_{noise}. Fourth, a smaller value of $\sigma_{\text{noise}} = 0.04$ provides a much smaller MSPE compared to the case when $\sigma_{\text{noise}} = 0.1$. In fact, σ_{noise} plays the role of a compromise between smoothness and fidelity to data samples.

Conclusion

The Gaussian process approximation originally proposed in [22] is considered, which verify interpolation conditions and inequality constraints in the entire domain. The flexibility of this approach to incorporate both noisy observations and multiple convex and non-convex constraints is investigated. This leads to significant improvement in prediction accuracy and more realistic credible intervals. We propose an adjustment to the cross-validation technique that uses *Maximum a Posteriori* (MAP) to estimate both covariance and noise variance hyperparameters. Additionally, we propose new basis functions to enhance the smoothness of the sample paths and ensure differentiability of class C^p, for any $p \geq 1$. The behavior of this approach in challenging situations, such as monotonicity with a flat region or boundedness where the underlying function is flat and close to lower and/or upper bounds, is investigated. In that case, we show that, unlike the MAP esti-

mate, the truncated multivariate normal distribution is not suitable for capturing the flat region. To address this issue, we propose adding multiple constraints, such as monotonicity with bounded slope constraints. The superiority of the MAP estimate over the mean a posterior (mAP) estimate is demonstrated in a wide range of settings based on its theoretical convergence. Real-world data studies show that the MAP estimate effectively captures flat regions and that incorporating multiple constraints accurately reflects the behavior of the posterior distribution.

Acknowledgements

The authors would like to thank Yann Richet from the Institut of Radioprotection and Nuclear Safety (IRSN, Paris) for providing the nuclear safety data. 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. *Electron. J. Statist.*, 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] Hugh A. Chipman, Edward I. George, Robert E. McCulloch, and Thomas 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] A. Cousin, A. Deleplace, and A. Misko. Gaussian process regression for swaption cube construction under no-arbitrage constraints. *Risks*, 10(12):232, 2022.

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

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

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

[16] S. G. Johnson. *The NLopt nonlinear-optimization package*.

[17] 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.

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

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

[20] 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.

[21] H. Maatouk. Finite-dimensional approximation of Gaussian processes with linear inequality constraints and noisy observations. *Communications in Statistics-Theory and Methods*, pages 1–20, 2022.

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

[23] H. Maatouk, X. Bay, and D. Rullière. A note on simulating hyperplane-truncated multivariate normal distributions. *Statistics & Probability Letters*, 191:109650, 2022.

[24] H. Maatouk, D. Rullière, and X. Bay. Sampling large hyperplane-truncated multivariate normal distributions. working paper or preprint, August 2022.

[25] H. Maatouk, D. Rullière, and X. Bay. Large scale Gaussian processes with Matheron's update rule and Karhunen-Loève expansion. In *To appear in: A. Hinrichs, P. Kritzer, F. Pillichshammer (eds.). Monte Carlo and Quasi-Monte Carlo Methods 2022*. Springer Verlag, 2023.

[26] A. Maradesa, B. Py, E. Quattrocchi, and F. Ciucci. The probabilistic deconvolution of the distribution of relaxation times with finite Gaussian processes. *Electrochimica Acta*, 413:140119, 2022.

[27] 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.

[28] K. P. Murphy. Machine learning: A probabilistic perspective (adaptive computation and machine learning series), 2018.

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

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

[31] E. Parzen. *Stochastic processes*. Holden-Day series in probability and statistics. Holden-Day, San Francisco, London, Amsterdam, 1962.

[32] M. JD Powell. Direct search algorithms for optimization calculations. *Acta numerica*, 7:287–336, 1998.

[33] C. E. Rasmussen and C. K.I. Williams. *Gaussian processes for machine learning*. MIT Press, Cambridge, 2006.

[34] 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.

[35] J. Riihimäki and A. Vehtari. Gaussian processes with monotonicity information. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 645–652. JMLR Workshop and Conference Proceedings, 2010.

[36] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn. Design and analysis of computer experiments. *Statistical Science*, 4(4):409–423, 1989.

[37] 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.

[38] 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.

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

[40] 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.

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

[42] N. J. Williams, C. Osborne, I. D. Seymour, M. Z. Bazant, and S. J. Skinner. Application of finite Gaussian process distribution of relaxation times on SOFC electrodes. *Electrochemistry Communications*, page 107458, 2023.

[43] Y. Yang, A. Bhattacharya, and D. Pati. Frequentist coverage and sup-norm convergence rate in Gaussian process regression. *arXiv preprint arXiv:1708.04753*, 2017.

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

[45] 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.