Finite-dimensional approximation of Gaussian processes with linear inequality constraints and noisy observations

Hassan Maatouk

ARTICLE TEMPLATE

Finite-dimensional approximation of Gaussian processes with linear inequality constraints and noisy observations

H. Maatouk[a]

[a] Departement of Mathematics, CY Cergy Paris University, CY Tech, France

ARTICLE HISTORY
Compiled June 12, 2021

ABSTRACT
Due to their flexibility, Gaussian processes (GPs) have been widely used in nonparametric function estimation. A prior information about the underlying function is often available. In this paper, the finite-dimensional Gaussian approach (Maatouk and Bay, 2017) which can satisfy linear inequality conditions everywhere (e.g. monotonicity, convexity and boundary) is considered. In a variety of real-world problems, the observed data usually possess noise. In this paper, this approach has been extended to deal with noisy observations. The mean and the maximum of the posterior distribution are well defined. Additionally, to simulate from the posterior distribution two methods have been used: the exact rejection sampling from the Mode and the Hamiltonian Monte Carlo method which is more efficient in high-dimensional cases. The generalization of the Kimeldorf-Wahba correspondence Kimeldorf and Wahba (1970) is proved in noisy observation cases. A comparison shown that the proposed model outperforms all recent models dealing with the same constraints in terms of predictive accuracy and coverage intervals.

KEYWORDS
Gaussian processes; nonparametric estimation; noisy observations; shape constraints; Kimeldorf-Wahba correspondence

1. Introduction

Gaussian processes (GPs) are one of the most famous choices in nonparametric function estimation. This is because of their flexibilities and other nice properties. For example, the conditional GP with linear equality constraints is still a GP (see Cramer and Leadbetter (1967)). Additionally, some linear inequality constraints (such as monotonicity and convexity) of output computer responses are related to partial derivatives. The partial derivatives of the GP remain GPs (see e.g. Cramer and Leadbetter (1967); Parzen (1962)).

In many real problems, a prior information about the underlying function is often available. Including shape constraints, the problem is called Constrained Gaussian Process (CGP) regression (or constrained kriging). It has been studied in the domain of geostatistics (see e.g. Freulon and de Fouquet (1993); Kleijnen and Van Beers (2013)) and used to improve the quality of predictions and to provide more realistic coverage intervals.

CONTACT H. Maatouk. Email: maatouk_hassan@hotmail.fr

In the literature, there are a variety of way to incorporate some linear constraints such as monotonicity and boundedness into a GP regression. Therefore, there exists a variety of methodologies to deal with them. For instance, GP regression with inequality constraints has been considered in Abrahamsen and Benth (2001); Da Veiga and Marrel (2012); Freulon and de Fouquet (1993); Koyejo et al. (2013); Rudovic and Pantic (2011); Salzmann and Urtasun (2010). For example, in Abrahamsen and Benth (2001) the idea is based on a discrete location approximation. In that case, the inequality constraints are satisfied in a finite number of input locations. Recently, a new methodology based on a modification of the covariance function in Gaussian processes to correctly account for known linear constraints is developed in Jidling et al. (2017). Recently, two new methodologies have been developed to incorporate boundedness constraints into a GP regression (see Agrell (2019); Jensen et al. (2013)).

For monotonicity and isotonicity constraints into a GP regression, a variety of methods have been developed. To the best of our knowledge, incorporating monotonicity information into a GP regression has been firstly introduced in Morris et al. (1993). Their idea is based on the fact that the order partial derivatives of a GP remain GPs (see Parzen (1962)). In fact, suppose that the paths of the GP $(y(\boldsymbol{x}))_{\boldsymbol{x}\in\mathbb{R}^d}$ are of class $\mathcal{C}^p$ (i.e., the space of functions that admit derivatives up to order p). This can be guaranteed if the covariance function k is smooth enough, and in particular if k is of class $\mathcal{C}^\infty$ (see Cramer and Leadbetter (1967)). Since differentiation is a linear operator, the order partial derivatives of a GP remain GPs (see Cramer and Leadbetter (1967); Parzen (1962)). Recently, this idea has been reconsidered in the machine learning framework for qualitative information (e.g., increasing or decreasing), see for example Riihimäki and Vehtari (2010). Their method is based on the knowledge of derivatives of the GP at some input locations. They incorporate the monotonicity information by placing virtual derivatives at specified input locations to force the derivative process to be positive at these points. In Golchi et al. (2015), this methodology has been extended to the interpolation case (i.e., free noisy observations). As well as, the simulation from the exact joint posterior distribution rather than relying on an approximation has been proposed. The authors provide complete Bayesian inference for all parameters of the emulator and the predicted function at unsampled inputs. The disadvantage of this methodology is the monotonicity constraints are not guaranteed in the entire domain. However, as mentioned in Wang and Berger (2016), 'only a modest number of virtual derivative points seems to be needed to effectively impose the desired shape constraint'. Another way to incorporate monotonicity information into a GP regression is the so called 'GP projection' (see Lin and Dunson (2014)) which is a competitive method (see Section 5 for a comparison with the proposed GP approximation in terms of prediction accuracy and coverage intervals). A comparison with spline-based models is included.

Recently, this approach has been extended to boundedness constraints (see Zhang and Lin (2018)).

For monotonic function estimations, using B-splines was firstly introduced by Ramsay (1988, 1998). The idea is based on the integration of B-splines defined on a properly set of knots with positive coefficients to ensure monotonicity constraints. Xuming and Peide (1996) take the same approach and suggest the calculation of the coefficients by solving a finite linear minimization problem. In Kong and Eubank (2006), regression spline has been used to estimate monotone smoothing function with application to dose-response curve. In Delecroix et al. (1996), nonparametric function estimation in a general cone is studied. Their method is based on a projection into a discretized version of the cone, using the theory of reproducing kernel Hilbert spaces. In Shively

et al. (2009), a Bayesian approach to estimate nonparametric monotone functions using restricted splines is developed. In Saarela and Arjas (2011); T. and L. (2007), the generalization of monotonic regression to multiple dimensions are studied. Bornkamp and Ickstadt (2009) adopt mixture modelling.

Finally, there is a vast frequentest literature on monotone function and isotonic regression estimation, with a common approach minimizing a least squares loss subject to a restriction (see Brunk et al. (1972); Robertson et al. (1988)). For more recent references see also Bhattacharya and Kong (2007); Bhattacharya and Lin (2010, 2011); Wahl and Espinasse (2018). For restricted kernel methods, refer to Dette et al. (2005); Mammen (1991); Müller and Schmitt (1988).

The methodology developed in the present paper is quite different. It has two key motivations: (1) By the proposed approach, the generalization of the Kimeldorf-Wahba correspondence (see Kimeldorf and Wahba (1970)) for constrained interpolation cases is proved (see Bay et al. (2016)) and for noisy observations can be generalized as we show in the next sections. (2) The linear inequality constraints on the Gaussian processes are respected in the entire domain and it is a competitive approach in terms of prediction and more importantly in terms of coverage intervals. In fact, the proposed approach is based on a finite-dimensional approximation of GPs (or a GP approximation) that converges uniformly pathwise. It can be seen as a linear combination between deterministic basis functions and Gaussian random coefficients, where the coefficients are not independent. The main idea is to choose the basis functions such that the infinite number of linear inequality constraints on the GP approximation are equivalent to a finite number of constraints on the coefficients. Therefore, the simulation of the conditional GP approximation is reduced to the simulation of a Gaussian vector (random coefficients) restricted to convex sets which is a well-known problem with existing algorithms (see e.g. Botts (2013); Chopin (2011); Maatouk and Bay (2016); Philippe and Robert (2003); Robert (1995)). Recently, the model developed in Maatouk and Bay (2017) has been used to estimate the proton radius from electron scattering data based on a non-parametric Gaussian process (see Zhou et al. (2019)).

The article is structured as follows. In Section 2, Gaussian processes for computer experiments are briefly reviewed. In Section 3, a finite-dimensional approximation of GPs capable of incorporating linear inequality constraints and noisy observations is developed. Section 4 shows the generalization of the Kimeldorf-Wahba correspondence (see Bay et al. (2017)) in the noisy observations case. Section 5 investigates the performance of the proposed model in terms of predictive accuracy and uncertainty quantification.

2. Gaussian processes for computer experiments

The following model is considered

$$y = f(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{R}^d,$$

where the simulator response y is assumed to be a deterministic real-valued function of the d-dimensional variable $\boldsymbol{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$. The true function is supposed to be continuous and evaluated at data of size n (design of experiments) given by the rows of the $n \times d$ matrix $\boldsymbol{X} = \left(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}\right)^\top$, where $\boldsymbol{x}^{(i)} \in \mathbb{R}^d,\ i = 1 \ldots, n$. In many practical situations, it is not possible to get exact evaluations of y at the design of experiments, but rather pointwise noisy measurements. In such case, an approximate

response $y(\boldsymbol{X}) + \epsilon$ is available, where $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2_{\text{noise}}\mathbf{I})$ with σ^2_{noise} the noise variance and $\mathbf{I}$ the identity matrix. To simplify notations, we denote

$$\tilde{y}_i = y(\boldsymbol{x}^{(i)}) + \epsilon_i, \quad i = 1, \ldots, n.$$

and $\tilde{\boldsymbol{y}} = y(\boldsymbol{X}) + \boldsymbol{\epsilon}$. In the statistical framework, y is viewed as a realization of a continuous GP

$$Y(\boldsymbol{x}) = \eta(\boldsymbol{x}) + z(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathcal{D} \subset \mathbb{R}^d,$$

where $\mathcal{D}$ is a compact subset of $\mathbb{R}^d$ and the deterministic continuous function $\eta : \boldsymbol{x} \in \mathbb{R}^d \longrightarrow \eta(\boldsymbol{x}) \in \mathbb{R}$ is the mean and z is a zero-mean GP with covariance function

$$k : (\boldsymbol{x}, \boldsymbol{x}') \in \mathcal{D} \times \mathcal{D} \longrightarrow k(\boldsymbol{x}, \boldsymbol{x}') \in \mathbb{R}.$$

In that case, the GP can be written as $Y \sim \mathcal{GP}\left(\eta(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}')\right)$. Conditionally to noisy observations $\tilde{\boldsymbol{y}} = (\tilde{y}_1, \ldots, \tilde{y}_n)^\top$, the process remains a GP

$$Y(.) \mid \{y(\boldsymbol{X}) + \boldsymbol{\epsilon}\} \sim \mathcal{GP}\left(\zeta(\boldsymbol{x}), C(\boldsymbol{x}, \boldsymbol{x}')\right),$$

where

$$\zeta(\boldsymbol{x}) = \mathrm{E}\left[Y(\boldsymbol{x}) \mid \tilde{\boldsymbol{y}}\right] = \eta(\boldsymbol{x}) + \boldsymbol{k}(\boldsymbol{x})^\top (\mathbb{K} + \sigma^2_{\text{noise}}\mathbf{I})^{-1} (\tilde{\boldsymbol{y}} - \boldsymbol{\mu}); \tag{1}$$

and $\boldsymbol{\mu} = \eta(\boldsymbol{X})$ is the vector of trend values at the design of experiments, $\mathbb{K}_{i,j} = k(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$, $i, j = 1, \ldots, n$ is the covariance matrix of $Y(\boldsymbol{X})$ and $\boldsymbol{k}(\boldsymbol{x}) = k(\boldsymbol{x}, \boldsymbol{X})$ is the vector of covariance between $Y(\boldsymbol{x})$ and $Y(\boldsymbol{X})$. Additionally, the covariance function between any two inputs is that

$$C(\boldsymbol{x}, \boldsymbol{x}') = \mathrm{Cov}(Y(\boldsymbol{x}), Y(\boldsymbol{x}') \mid y(\boldsymbol{X}) = \tilde{\boldsymbol{y}}) = k(\boldsymbol{x}, \boldsymbol{x}') - \boldsymbol{k}(\boldsymbol{x})^\top (\mathbb{K} + \sigma^2_{\text{noise}}\mathbf{I})^{-1} \boldsymbol{k}(\boldsymbol{x}'),$$

where C is the covariance function of the conditional GP. The mean $\zeta(\boldsymbol{x})$ is called kriging mean prediction of $y(\boldsymbol{x})$ based on the computer model outputs $y(\boldsymbol{X}) = \tilde{\boldsymbol{y}}$ (see Rasmussen and Williams (2006)). We refer to Lyu et al. (2018) for more details on the evaluation of the GP metamodels for noisy level set estimation. In next sections, the generalization of the model developed in Maatouk and Bay (2017) to the case of observations with Gaussian noise is investigated.

3. Finite-dimensional approximation of GPs

The input set is supposed the unit hypercube $\mathcal{D} = [0, 1]^d$ (without loss of generality). The input set $\mathcal{D}$ is discretized uniformly to $(m + 1)^d$ knots. For example, in one dimension where $\mathcal{D} = [0, 1]$, the discretization can be summarized as follow: $0 = t_{m,0}, \ldots, t_{m,m} = 1$. Let $Y \sim \mathcal{GP}(0, k(x, x'))$ be a zero-mean GP with covariance function k. The finite-dimensional approximation of Gaussian processes developed in

Maatouk and Bay (2017) is then considered

$$y^m(x) \quad = \quad \sum_{l=0}^{m} Y(t_{m,l})\phi_l(x) = \sum_{l=0}^{m} \beta_l \phi_l(x), \quad x \in \mathcal{D} \tag{2}$$

where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_m)^\top = (Y(t_{m,0}), \ldots, Y(t_{m,m}))$ is a zero-mean Gaussian vector with covariance matrix Γ^m, with $\Gamma^m_{j,l} = k(t_{m,j}, t_{m,l})$, $0 \leq j, l \leq m$ and ϕ_l is the hat function associated to the knot $t_{m,l}$ (see Maatouk and Bay (2017) for more details). We first discretize the input set as $0 = t_{m,0} < t_{m,1} < \ldots < t_{m,m} = 1$, and on each knot we build a function. For the sake of simplicity, we use a uniform subdivision of the input set, but the methodology can be adapted for any subdivision. For example at the j^{th} knot $t_{m,j} = j\Delta_m = j/m$, the associated function is

$$\phi_j(x) = h\left(\frac{x - t_{m,j}}{\Delta_m}\right), \qquad j = 0, \ldots, m, \tag{3}$$

where $\Delta_m = 1/m$ and $h(x) = (1 - |x|)\, \mathbb{1}_{(|x| \leq 1)},\ x \in \mathbb{R}$. The value of any basis function at any knot is equal to Kronecker's Delta function ($\phi_l(t_{m,j}) = \delta_{j,l},\ j, l = 0, \ldots, m$), where $\delta_{j,l}$ is equal to one if $j = l$ and zero otherwise. The covariance function $k_m(x, x')$ of the Gaussian process approximation y^m is

$$k_m(x, x') = \phi(x)^\top \Gamma^m \phi(x'), \tag{4}$$

where $\phi(x) = (\phi_0(x), \ldots, \phi_m(x)) \in \mathbb{R}^{m+1}$. By this special choice of the basis functions, the finite-dimensional approximation of GPs y^m can be viewed as the piecewise linear interpolation of Y at the knots $t_{m,0}, \ldots, t_{m,m}$. By this approach (2), simulate the GP approximation with linear inequality constraints (i.e., $y^m \in \mathcal{C}$, where $\mathcal{C}$ is the space of functions verifying linear inequality constraints such as monotonicity, boundedness,...) and noisy observations is equivalent to simulate the Gaussian vector $\beta = (\beta_0, \ldots, \beta_m)^\top$ restricted to

$$\begin{aligned} &\beta \times \phi(x^{(i)}) = y_i + \epsilon_i = \tilde{y}_i, \quad i = 1, \ldots, n, \\ &\beta = (\beta_0, \ldots, \beta_m)^\top \in \mathcal{C}_{\text{coef}}, \end{aligned}$$

where $\phi(x^{(i)}) = (\phi_0(x^{(i)}), \ldots, \phi_m(x^{(i)})) \in \mathbb{R}^{m+1}$, $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2_{\text{noise}})$ and $\mathcal{C}_{\text{coef}}$ is the space of coefficients which verify some linear constraints. In the next sections, we show how $\mathcal{C}_{\text{coef}}$ can be computed in the case where $\mathcal{C}$ is the space of functions verifying convexity constraints for example. In this paper, boundedness, monotonicity and convexity constraints are considered but the methodology can be easily adapted to any linear inequality constraints. Additionally, incorporating multiple linear inequality constraints can be easily adapted as well (see López-Lopera et al. (2018)). Let us recall that other basis functions have been used in Maatouk and Bay (2017) to incorporate some linear inequality constraints which are related to derivative information such as monotonicity and convexity (see Remark 2 below).

3.1. GP approximation with convexity constraints

In this section, $\mathcal{C}$ is the space of functions verifying convexity constraints. For the sake of simplicity, the convexity constraint in two dimensions is considered. The multidimensional case is a simple extension of two dimensional one. The finite-dimensional approximation of GPs defined as

$$y^m(x_1,x_2)=\sum_{j,l=0}^{m} Y(t_{m,j},t_{m,l})\phi_j(x_1)\phi_l(x_2)=\sum_{j,l=0}^{m}\beta_{j,l}\phi_j(x_1)\phi_l(x_2),\quad (x_1,x_2)\in\mathcal{D}^2 \tag{5}$$

is convex with respect to the two inputs *if and only if* the random coefficients verify

(1) $\frac{\beta_{j,l}-\beta_{j-1,l}}{t_{m,j}-t_{m,j-1}}\leq\frac{\beta_{j+1,l}-\beta_{j,l}}{t_{m,j+1}-t_{m,j}}$ and $\frac{\beta_{j,l}-\beta_{j,l-1}}{t_{m,l}-t_{m,l-1}}\leq\frac{\beta_{j,l+1}-\beta_{j,l}}{t_{m,l+1}-t_{m,l}},\ j,l=1,\dots,m-1;$
(2) $\frac{\beta_{j,0}-\beta_{j-1,0}}{t_{m,j}-t_{m,j-1}}\leq\frac{\beta_{j+1,0}-\beta_{j,0}}{t_{m,j+1}-t_{m,j}},\ j=1,\dots,m-1;$
(3) $\frac{\beta_{0,l}-\beta_{0,l-1}}{t_{m,l}-t_{m,l-1}}\leq\frac{\beta_{0,l+1}-\beta_{0,l}}{t_{m,l+1}-t_{m,l}},\ l=1,\dots,m-1.$

Remark 1. From Model (5), the convexity constraints can be obtained easily with respect to one of the two input variables. For example, the convexity constraint of y^m with respect to the first input variable is equivalent to the following constraints

$$\frac{\beta_{j,l}-\beta_{j-1,l}}{t_{m,j}-t_{m,j-1}}\leq\frac{\beta_{j+1,l}-\beta_{j,l}}{t_{m,j+1}-t_{m,j}},\quad j=1,\dots,m-1,\ l=0,\dots,m. \tag{6}$$

In that case, $\mathcal{C}_{\text{coef}}=\left\{(\beta_{j,l})_{j,l}\in\mathbb{R}^{(m+1)^2}\ \middle|\ \beta_{j,l}\text{ verify }(6)\right\}$.

Remark 2. In one-dimensional cases, the convexity constraints can be also obtained by replacing the basis functions in (2) by the two times primitive functions of $\phi_l(x)$

$$\psi_l(x)=\int_0^x\left(\int_0^v\phi_l(u)du\right)dv.$$

In that case, the finite-dimensional approximation of GPs defined as

$$y^m(x)=Y(0)+Y'(0)x+\sum_{l=0}^{m}Y''(t_{m,l})\psi_l(x),\quad x\in\mathcal{D}$$

is convex if and only if the $(m+1)$ random coefficients $(Y''(t_{m,l}))$ are all non-negative. Thus, $\mathcal{C}_{\text{coef}}=\{(Y(0),Y'(0),(Y''(t_{m,l}))_l)\in\mathbb{R}^{(m+3)^2}\mid Y''(t_{m,l})\geq 0,\ \forall\ l=0,\dots,m\}$.

3.2. Simulated paths

This subsection is devoted to the sampling scheme of the proposed model conditionally to linear inequality constraints and noisy observations. To simplify notations, the finite-dimensional approximation of GPs in one dimension is considered

$$y^m(x)=\sum_{l=0}^{m}Y(t_{m,l})\phi_l(x)=\sum_{l=0}^{m}\beta_l\phi_l(x),\quad x\in\mathcal{D}.$$

In this paper, the GP is observed with error. The space of coefficients is defined

$$\begin{aligned} I_{\text{coef}} &= \{\boldsymbol{\beta} \in \mathbb{R}^{m+1} \ : \sum_{l=0}^{m} \beta_l \phi_l(x^{(i)}) = \tilde{y}_i, \ i = 1, \ldots, n\} \\ &= \left\{\boldsymbol{\beta} \in \mathbb{R}^{m+1} \ : A\boldsymbol{\beta} = \tilde{\boldsymbol{y}}\right\}, \end{aligned}$$

where $\tilde{y}_i = y_i + \epsilon_i, \ i = 1, \ldots, n, \ \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2_{\text{noise}})$ and $A_{i,l} = \phi_l(x^{(i)})$. The sampling scheme can be summarized in two steps: first, the conditional Gaussian vector $\boldsymbol{\beta}$ with only noisy observations is simulated

$$\boldsymbol{\beta} \mid A\boldsymbol{\beta} = \tilde{\boldsymbol{y}} \sim \mathcal{N}\left((A\Gamma^m)^\top (A\Gamma^m A^\top + \sigma^2_{\text{noise}}\mathbf{I})^{-1}\tilde{\boldsymbol{y}}, \Gamma^m - (A\Gamma^m)^\top (A\Gamma^m A^\top + \sigma^2_{\text{noise}}\mathbf{I})^{-1} A\Gamma^m\right).$$

Second, by an improved rejection sampling (see Maatouk and Bay (2016)), only the random coefficients in the convex set $\mathcal{C}_{\text{coef}}$ are selected.

Definition 3.1. The so-called unconstrained mean which is the mean of the Gaussian process approximation conditionally only to noisy observations is defined as

$$m^m(x) = \mathrm{E}\left[y^m(x) \ \middle| \ y^m(x^{(i)}) = \tilde{y}_i, \ i = 1, \ldots, n\right] = \phi(x)^\top \boldsymbol{\beta}_I,$$

where $\boldsymbol{\beta}_I = \mathrm{E}\left[\boldsymbol{\beta} \mid \boldsymbol{\beta} \in I_{coef}\right] = \Gamma^m A^\top \left(A\Gamma^m A^\top + \sigma^2_{noise}\boldsymbol{I}\right)^{-1} \tilde{\boldsymbol{y}}$.

Similarly to the kriging mean of the original GP y (equation (1), when η is the null function), the kriging mean m^m of the finite-dimensional approximation of GPs y^m can be written as

$$m^m(x) = \boldsymbol{k}_m(x)^\top \left(\mathbb{K}_m + \sigma^2_{\text{noise}}\mathbf{I}\right)^{-1} \tilde{\boldsymbol{y}},$$

where $\boldsymbol{k}_m(x) = k_m(x, \boldsymbol{X}) = (A\Gamma^m \phi(x))$ is the vector of covariance between $y^m(x)$ and $y^m(\boldsymbol{X})$ and $(\mathbb{K}_m)_{i,j} = k_m(x^{(i)}, x^{(j)}) = \left(A\Gamma^m A^\top\right)_{i,j}$, $i, j = 1, \ldots, n$ is the covariance matrix of $y^m(\boldsymbol{X})$.

Remark 3. The unconstrained mean $m^m(x)$ respects linear inequality constraints in the entire domain *if and only* if the conditional Gaussian vector to only noisy observations $\boldsymbol{\beta}_I$ lies inside the convex set $\mathcal{C}_{coef}$.

Definition 3.2. The mean of the posterior distribution of y^m conditionally to linear inequality constraints and noisy observations is defined as

$$m^m_{pos}(x) = \mathrm{E}\left[y^m(x) \ \middle| \ y^m(x^{(i)}) = \tilde{y}_i, \ \boldsymbol{\beta} \in \mathcal{C}_{coef}\right] = \phi(x)^\top \boldsymbol{\beta}_{pos},$$

where $\boldsymbol{\beta}_{pos} = \mathrm{E}\left[\boldsymbol{\beta} \mid \tilde{y}, \boldsymbol{\beta} \in \mathcal{C}_{coef}\right]$ is the posterior mean which is computed from simulations.

Finally, let μ be the maximum of the probability density function (pdf) of $\boldsymbol{\beta}$ restricted to $I_{\text{coef}} \cap \mathcal{C}_{\text{coef}}$. It is the solution of the following convex optimization problem

$$\mu = \arg \min_{x \in I_{\text{coef}} \cap \mathcal{C}_{\text{coef}}} \left(\frac{1}{2} x^\top (\Gamma^m)^{-1} x\right), \tag{7}$$

where Γ^m is the covariance matrix of the Gaussian vector $\boldsymbol{\beta}$. The quadratic optimization problem (7) is equivalent to

$$\begin{aligned} \mu &= \arg\min_{x\in\mathcal{C}_{\text{coef}}} \left(\left(\frac{1}{2}x^\top - \boldsymbol{\beta}_{\text{I}}^\top \right) (\Gamma^m_{\text{cond}})^{-1} x \right), \\ &= \arg\min_{x\in\mathcal{C}_{\text{coef}}} \left(-\boldsymbol{\beta}_{\text{I}}^\top (\Gamma^m_{\text{cond}})^{-1} x + \frac{1}{2}x^\top (\Gamma^m_{\text{cond}})^{-1} x \right), \end{aligned} \tag{8}$$

where Γ^m_{cond} is the covariance matrix of the conditional Gaussian vector $\boldsymbol{\beta} \mid A\boldsymbol{\beta} = \tilde{\boldsymbol{y}}$. Equation (8) done by the log-likelihood of $\boldsymbol{\beta} \mid A\boldsymbol{\beta} = \tilde{\boldsymbol{y}}$. In fact,

$$\begin{aligned} \frac{1}{2}(x-\boldsymbol{\beta}_{\text{I}})^\top (\Gamma^m_{\text{cond}})^{-1}(x-\boldsymbol{\beta}_{\text{I}}) &= \frac{1}{2}x^\top (\Gamma^m_{\text{cond}})^{-1} x - \frac{1}{2}x^\top (\Gamma^m_{\text{cond}})^{-1}\boldsymbol{\beta}_{\text{I}} \\ &- \frac{1}{2}\boldsymbol{\beta}_{\text{I}}^\top (\Gamma^m_{\text{cond}})^{-1} x + \frac{1}{2}\boldsymbol{\beta}_{\text{I}}^\top (\Gamma^m_{\text{cond}})^{-1}\boldsymbol{\beta}_{\text{I}} \\ &= \frac{1}{2}x^\top (\Gamma^m_{\text{cond}})^{-1} x - \boldsymbol{\beta}_{\text{I}}^\top (\Gamma^m_{\text{cond}})^{-1} x + \frac{1}{2}\boldsymbol{\beta}_{\text{I}}^\top (\Gamma^m_{\text{cond}})^{-1}\boldsymbol{\beta}_{\text{I}}. \end{aligned}$$

Thus,

$$\begin{aligned} &\arg\min_{x\in\mathcal{C}_{\text{coef}}} \left(\frac{1}{2}x^\top (\Gamma^m_{\text{cond}})^{-1} x - \boldsymbol{\beta}_{\text{I}}^\top (\Gamma^m_{\text{cond}})^{-1} x + \frac{1}{2}\boldsymbol{\beta}_{\text{I}}^\top (\Gamma^m_{\text{cond}})^{-1}\boldsymbol{\beta}_{\text{I}} \right) \\ &= \arg\min_{x\in\mathcal{C}_{\text{coef}}} \left(-\boldsymbol{\beta}_{\text{I}}^\top (\Gamma^m_{\text{cond}})^{-1} x + \frac{1}{2}x^\top (\Gamma^m_{\text{cond}})^{-1} x \right). \end{aligned}$$

The vector μ represents the maximum of the pdf of the Gaussian vector $\boldsymbol{\beta}$ restricted to $I_{\text{coef}} \cap \mathcal{C}_{\text{coef}}$ and its numerical calculation is a standard problem in the minimization of positive quadratic forms subject to convex constraints (see Boyd and Vandenberghe (2004); Goldfarb and Idnani (1983)). Let us mention that in all simulated examples illustrated in this paper, the R-package 'solve.QP' described in Goldfarb and Idnani (1983) is used to solve the quadratic convex optimization problems (7)-(8).

Definition 3.3. The maximum of the posterior distribution of y^m conditionally to linear inequality constraints and noisy observations is defined as

$$M^m_{pos}(x) = \sum_{l=0}^{m} \mu_l \phi_l(x), \quad x \in \mathcal{D},$$

where $\mu = (\mu_0, \ldots, \mu_m)^\top$ is computed by (8).

Remark 4. The maximum *a posteriori* estimate M^m_{pos} does not depend on the variance hyper-parameter σ of the covariance function k as well as on the simulations but depends on the length hyper-parameters of the covariance function $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$.

Remark 5. In the case where the GP is observed without error (i.e., with noise-free data), the maximum *a posteriori* estimate M^m_{pos} is computed as follow

$$M^m_{pos}(x) = \sum_{l=0}^{m} \tilde{\mu}_l \phi_l(x), \quad x \in \mathbb{R}^d,$$

Algorithm 1: Sampling scheme

Initialization:
$\boldsymbol{\beta} \notin \mathcal{C}_{\text{coef}}$; $\boldsymbol{\beta} = \boldsymbol{\beta}_{\text{current}}$
unif $= 1$; $t = 0$
while unif $> t$ **do**
  $\boldsymbol{\beta} = \boldsymbol{\beta}_{\text{current}}$
  while $\boldsymbol{\beta}_{\text{current}} \notin \mathcal{C}_{\text{coef}}$ **do**
    $\boldsymbol{\beta}_{\text{current}} \sim \mathcal{N}(\mu, \Gamma^m_{\text{cond}})$
  end
  $t = \exp\left(\mu^\top (\Gamma^m_{\text{cond}})^{-1}(\mu - \boldsymbol{\beta}_{\text{I}} - \boldsymbol{\beta}_{\text{current}}) + \boldsymbol{\beta}^\top_{\text{current}} (\Gamma^m_{\text{cond}})^{-1} \boldsymbol{\beta}_{\text{I}}\right)$
  unif $\sim \mathcal{U}(0, 1)$
end

where $\tilde{\mu} = (\tilde{\mu}_0, \ldots, \tilde{\mu}_m)^\top$ is computed by the following new quadratic optimization problem

$$\tilde{\mu} = \arg \min_{x \in \mathcal{C}_{\text{coef}}} \left(\left(\frac{1}{2} x^\top + \tilde{\boldsymbol{\beta}}_{\text{I}}^\top \right) \left(\tilde{\Gamma}^m_{\text{cond}} \right)^{-1} x \right), \tag{9}$$

where $\tilde{\Gamma}^m_{\text{cond}}$ is the covariance matrix of the conditional Gaussian vector $\boldsymbol{\beta} \mid A\boldsymbol{\beta} = \boldsymbol{y}$, with $\boldsymbol{y} = (y(\boldsymbol{x}^{(1)}), \ldots, y(\boldsymbol{x}^{(n)}))$ and $\tilde{\boldsymbol{\beta}}_I = \mathrm{E}\left[\boldsymbol{\beta} \mid A\boldsymbol{\beta} = \boldsymbol{y}\right] = \Gamma^m A^\top (A\Gamma^m A^\top)^{-1} \boldsymbol{y}$. In that case, the authors in Bay et al. (2016, 2017) show that the maximum *a posteriori* estimate M^m_{pos} converges uniformly to the constrained interpolation function solution of the following convex optimization problem

$$\arg \min_{h \in H \cap I \cap \mathcal{C}} \|h\|^2_H,$$

where H is the reproducing kernel Hilbert space (RKHS) associated to the positive type kernel K (see Aronszajn (1950); Berlinet and Thomas-Agnan (2004)), I is the set of functions verify interpolation conditions (i.e., $h(\boldsymbol{x}^i) = y_i$) and the convex set $\mathcal{C}$ is the space of functions verifying linear inequality constraints.

This generalizes to the case of interpolation conditions and linear inequality constraints the well known correspondence established by Kimeldorf and Wahba (1970) between Bayesian estimation on stochastic process and smoothing by splines. By this result, the constrained interpolation function has a nice probabilistic interpretation as a Bayesian estimator. It can be seen as the most likely function which verifies the interpolation condition and linear inequality constraints. By this correspondence, one can compute the constrained interpolation function in a deterministic setting and sample from the posterior distribution of a GP to quantify the uncertainty in a Bayesian framework. Additionally, this new correspondence shows that the *maximum a posteriori* estimate when m is large enough has the same smoothness of the given reproducing kernel.

Based on the result developed in the present paper (see section 4), the well known correspondence established by Kimeldorf and Wahba (1970) is generalized to the linearly constrained cases with noisy observations.

In Algorithm 1, the sampling scheme of the proposed model is described. It is based on the rejection sampling from the Mode (RSM) algorithm to simulate the Gaussian

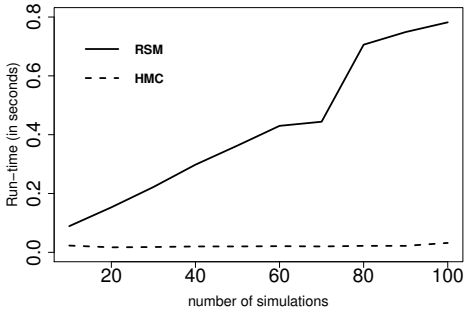


Figure 1. Run-time (in seconds) of the simulations using Hamiltonian Monte Carlo (HMC) and Rejection Sampling from the Mode (RSM).

vector $\boldsymbol{\beta}$ restricted to the convex set $I_{\text{coef}} \cap \mathcal{C}_{\text{coef}}$ (see, Maatouk and Bay (2016) for more details). This is an exact method to simulate from the posterior distribution in multidimensional cases. Additionally, it is simple to implement. However, in high-dimensional cases, the acceptance rate becomes low. In the literature, many different algorithms have been developed to simulate from the truncated Gaussian vector restricted to linear constraints. Recently, three different algorithms have been proposed: Gibbs sampling (see Souris et al. (2018); Taylor and Benjamini (2017)), Metropolis-Hastings (see Murphy (2012)), and Hamiltonian Monte Carlo (HMC) (see Pakman and Paninski (2014)). In this paper, the HMC proposed in Pakman and Paninski (2014) has been used which is an efficient sampler for the proposed framework (see López-Lopera et al. (2018)).

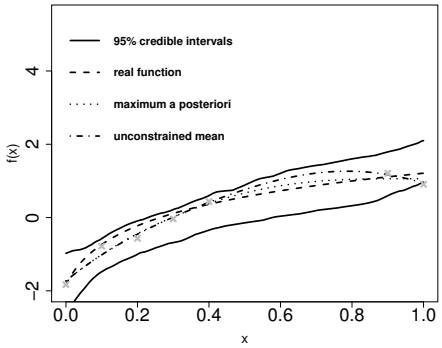
In Figure 1, the boundedness constraints example used in Section 4 to show the new generalization of the Kimeldorf-Wahba correspondence is considered. The run-time (in seconds) of the simulations has been computed using the exact Rejection Sampling from the Mode (RSM) and the Hamiltonian Monte Carlo (HMC).

3.3. Predictive accuracy and uncertainty quantification

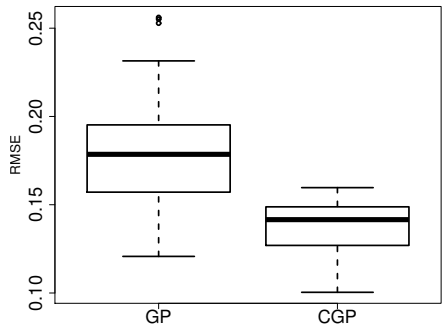
The aim of this section is to show the performance of the proposed model in terms of prediction and more importantly in terms of coverage intervals. The monotonicity constraint is considered

$$\mathcal{C} = \left\{ f \in \mathcal{C}^0([0,1]) \ : \ f(u) \leq f(v), \ \forall \ u \leq v \right\}.$$

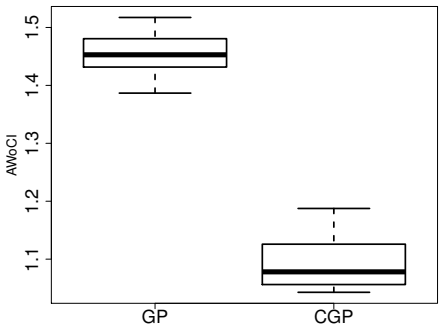
To do this, we take the real increasing function $f(x) = log(20x + 1)$ used in Golchi et al. (2015) (black dashed lines in Figure 2a). Suppose that f is evaluated at $X = (0, 0.1, 0.2, 0.3, 0.4, 0.9, 1)$. As mentioned in Golchi et al. (2015), this is a challenge situation for unconstrained GP since we have a large gap between the fifth and sixth design points (i.e. $0.4 < x < 0.9$). In Figure 2a, prediction intervals and maximum *a posteriori* estimate together with the real function are shown using GP approximation with monotonicity constraints. As in Golchi et al. (2015), the Matérn 5/2 covariance function is used. Applying a suited cross validation method to estimate covariance hyper-parameters Cousin et al. (2016); Maatouk et al. (2015), we get $\sigma = 335.5$ and $\theta = 4.7$. Let us mention that the maximum likelihood estimation under inequality constraints proposed in Bachoc et al. (2019) can be used.



(a)



(b)



(c)

Figure 2. GP approximation together with the true function (a). Boxplot of the RMSEs (b) and average width of the 95% credible intervals (c).

In Figure 2b and Figure 2c, boxplot of the root-mean-square error (RMSEs) and average width of the 95% credible intervals (AWoCI) for unconstrained and constrained GP are shown. Let us recall that the AWoCI criteria used in Figure 2c is defined as

$$\text{AWoCI} = \frac{1}{n_t}\sum_{i=1}^{n_t}(Q_{0.975}^{(i)} - Q_{0.025}^{(i)}),$$

where $Q_p^{(i)}$ is the pth posterior sample quantile and $n_t = 7$ the number of tested points specified by a Latin hypercube design. Because the prediction accuracy depends on the design locations, we repeated the procedure with twenty different random Latin hypercube design. Figures 2b and 2c show that the constrained GP approximations often outperform the unconstrained ones. According to the AWoCI criterion, the constrained GP approximations provide more realistic coverage intervals. In Figure 3, the predictive uncertainties for four design points chosen equidistant between 0.4 and 0.9 are given using unconstrained and constrained GP respectively. The reduction in prediction uncertainty for the constrained GP approximation is evident from the boxplot of the AWoCI criteria.

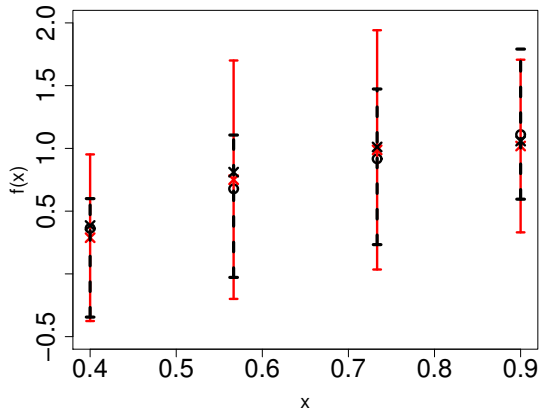


Figure 3. Posterior mean and 95% confidence intervals obtained by unconstrained GP (red lines) and Posterior maximum and 95% confidence intervals obtained by constrained GP (black dashed-lines); the circles show the true function values.

4. Generalization of the Kimeldorf-Wahba correspondence for linear inequality constraints and noisy observations

The goal of this section is twofold: first, to illustrate the condition simulation of the GP approximation developed in the present paper with boundedness constraints and noisy observations. Second, to describe the two different cases in the simulation which can help the readers to better understand the new generalization of the Kimeldorf-Wahba correspondence in the case of linear inequality constraints and noisy observations.

- The unconstrained mean respects the constraints and then coincides with the maximum of the posterior distribution.
- The unconstrained mean does not respect the constraints, then the unconstrained mean and the maximum of the posterior distribution are different.

In the following, the generalization of the Kimeldorf-Wahba correspondence for linear inequality constraints and noisy observations is considered.

Proposition 4.1. *Let* $\{y^m(x)\}_{x\in\mathcal{D}}$ *be a finite-dimensional GP of the form* (2) *and* $H_m := \mathrm{Vect}\{\phi_l,\ 1 \le l \le m\} = \left\{y^m \in \mathbb{R}^{\mathcal{D}} \ :\ (c_1,\dots,c_m) \in \mathbb{R}^m,\ y^m = \sum_{l=1}^m c_l\phi_l\right\}$ *be the linear space spanned by the basis functions* ϕ_j*. According to Bay et al. (2016),* H_m *is the RKHS associated with the kernel function* k_m *given in* (4)*. Let* I *be the space of functions verifying noisy observations. Then, the MAP estimator* $\hat{y}_m$ *defined as the mode of the posterior distribution density of* $\{y^m \mid y^m \in \mathcal{C}\cap I\}$ *is equal to the constrained function* $y_{opt,m}$ *solution of*

$$\arg\min_{y^m\in H_m\cap\mathcal{C}} \|y^m\|_m^2 + \sum_{i=1}^n\sum_{k=1}^n [y^m(x^{(i)}) - y_i](b_{ik})^{-1}[y^m(x^{(k)}) - y_k],$$

where $B = (b_{ik})_{i,k} = \sigma^2_{noise}\mathbf{I}$ *is a positive definite matrix and* $\|y^m\|_m^2 = \langle y^m, y^m\rangle_m = c^\top(\Gamma^m)^{-1}c$ *(see Bay et al. (2016) for more details). In fact,* $b_{ik} = \mathrm{E}[\epsilon_i\epsilon_k]$*, for all* $i,k = 1,\dots,n$.

As mentioned in Kimeldorf and Wahba (1970), if $\sum_{i=1}^{n}\sum_{k=1}^{n}[y^m(x^{(i)}) - y_i](b_{ik})^{-1}[y^m(x^{(k)}) - y_k]$ *is interpreted as a measure of the disparity of* y^m *with the data, then* $y_{opt,m}$*, is a compromise between smoothness and fidelity to the data.*

Proof. According to Bay et al. (2016), the (unconstrained) density function of y^m is

$$y^m \in H_m \longmapsto \frac{1}{\sqrt{2\pi}^m |\Gamma^m|^{1/2}} \exp\left(-\frac{1}{2}\|y^m\|_m^2\right),$$

where for all $y^m = \sum_{l=0}^{m} c_l\phi_l \in H_m$, $\|y^m\|_m^2 = (y^m, y^m)_m = c^\top(\Gamma^m)^{-1}c$. Let us now introduce the space of functions which verify noisy observations. According to Kimeldorf and Wahba (1970) (see Theorem 3.1, where the positive definite matrix B is $\sigma^2_{\text{noise}}\mathbf{I}$ in our case), The posterior likelihood L_{pos} defined as the pdf of y^m conditionally to noisy observations is

$$y^m \in H_m \longmapsto k^{-1}\mathbb{1}_{(y^m\in H_m)} \exp\left(-\frac{1}{2}\|y\|_m^2 - \frac{1}{2}\sigma^{-2}_{\text{noise}}\sum_{i=1}^{n}\sum_{k=1}^{n}[y^m(x^{(i)}) - y_i][y^m(x^{(k)}) - y_k]\right),$$

where $k \neq 0$ is a normalizing constant. In the Bayesian framework, the *prior* including inequality constraints is the following truncated pdf :

$$y^m \in H_m \longmapsto k^{-1}\mathbb{1}_{(y^m\in H_m\cap\mathcal{C})} \exp\left(-\frac{1}{2}\|y^m\|_m^2\right),$$

where k is a different normalizing constant. The posterior likelihood L_{pos} of y^m conditionally to inequality constraints and noisy observations is

$$L_{pos}(y^m) = k^{-1}\mathbb{1}_{(y^m\in H_m\cap\mathcal{C})} \exp\left(-\frac{1}{2}\|y^m\|_m^2 - \frac{1}{2}\sigma^{-2}_{\text{noise}}\sum_{i=1}^{n}\sum_{k=1}^{n}[y^m(x^{(i)}) - y_i][y^m(x^{(k)}) - y_k]\right). \tag{10}$$

By definition, the MAP estimator $\hat{y}_m$ is the solution of the following optimization problem

$$\arg\max L_{pos}(y^m) = \arg\min\left(-2\log L_{pos}(y^m)\right).$$

From expression (10), the MAP estimator $\hat{y}_m$ is the *constrained* function $y_{opt,m}$ which minimizes

$$\|y^m\|_m^2 + \sum_{i=1}^{n}\sum_{k=1}^{n}[y^m(x^{(i)}) - y_i](b_{ik})^{-1}[y^m(x^{(k)}) - y_k].$$

In the case of noise-free evaluations (i.e., $y^m(x^{(i)}) = y_i,\ \forall i = 1,\ldots,n$), we found the result of Proposition 1 in Bay et al. (2016). □

4.1. Numerical illustration

In this section, the numerical illustration of the Kimeldorf-Wahba correspondence for inequality constraints and noisy observations is now considered.

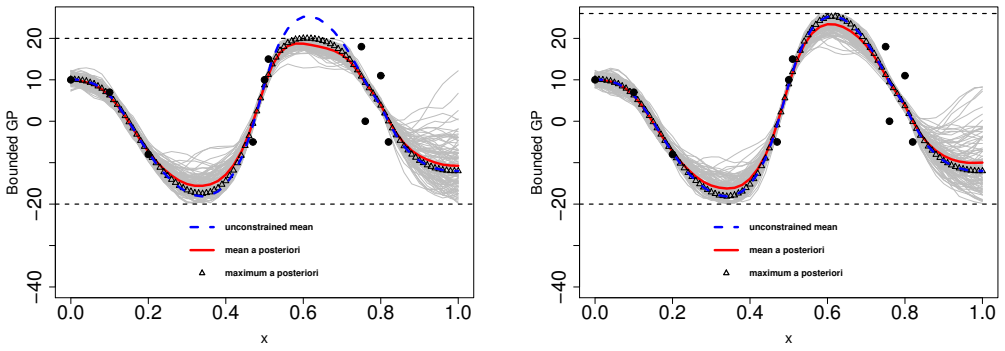
The real function is supposed to respect boundedness constraints

$$\mathcal{C} = \left\{ f \in \mathcal{C}^0([0,1]) \ : \ -\infty \leq a \leq f(x) \leq b \leq +\infty, \ x \in [0,1] \right\}. \tag{11}$$

The constrained data of size $n = 10$ (black points in Figure 4) are not taken from constrained functions. The noise variance is fixed to $\sigma^2_{\text{noise}} = 1.1^2$. Additionally, the Matérn 3/2 covariance function

$$k_{\theta,\sigma}(x, x') = \sigma^2 \left(1 + \frac{\sqrt{3} \mid x - x' \mid}{\theta}\right) \exp\left(-\frac{\sqrt{3}|x - x'|}{\theta}\right),$$

is used with the hyper-parameters fixed to $(\theta, \sigma) = (0.3, 10)$.



(a) (b)

Figure 4. Unconstrained and mean a posteriori together with the maximum a posteriori (MAP) estimator using the constrained model. The lower and upper bounds are equal to -20 and 20 (a) and equal to -20 and 26 (b). The black points represent the observations.

In Figure 4a, we generate one hundred sample paths taken from Model (2) with $d = 1$ and $m = 50$ conditionally to boundedness constraints (i.e., $a = -20$ and $b = 20$ in (11)). The simulated trajectories (gray lines) respect boundedness constraints in the entire domain as well as the mean and the maximum of the posterior distribution, contrarily to the unconstrained mean. This is the case where $\boldsymbol{\beta}_{\text{I}}$ lies outside the acceptance region $\mathcal{C}_{\text{coef}}$ (Remark 3). In Figure 4b, we just relax the boundedness constraints such that the unconstrained mean respects it. In that case, the unconstrained mean coincides with the MAP estimator but not with the mean of the posterior distribution. It corresponds to the situation where the conditional Gaussian vector $\boldsymbol{\beta}_{\text{I}}$ lies inside the acceptance region $\mathcal{C}_{\text{coef}}$ (Remark 3). Hence, in the constrained case the correspondence established by Kimeldorf-Wahba (see Kimeldorf and Wahba (1970)) is achieved with the MAP estimator and not the mean of the posterior distribution.

5. Simulation study

In this section, a comparison between the finite-dimensional approximation of GPs developed in the present paper and models deal with monotonicity and isotonicity constraints is shown. The real non-decreasing functions proposed by Holmes and Heard

Table 1. Length hyper-parameter estimates using a suited cross-validation method.

	Flat	Step	Linear	Exponential	Logistic	Sinusoidal
$\hat{\theta}$	100.0	0.1	3	2.5	3.4	3.5

Table 2. Root-mean-square error ($\times$ 100) for data of size $n = 100$. The results are obtained by repeating the simulation 5000 times.

	Flat	Step	Linear	Exponential	Logistic	Sinusoidal
Gaussian process	15.1	27.1	16.7	19.7	25.5	21.9
Gaussian process projection	11.3	25.3	16.3	19.1	22.4	21.1
Regression spline	9.7	28.5	24.0	21.3	19.4	22.9
Gaussian process approximation	0.5	25.3	16.3	18.6	19.5	20.4

(2003); Neelon and Dunson (2004) and used in a comparative study by Lin and Dunson (2014); Shively et al. (2009) are considered

- flat function $f_1(x) = 3,\ x \in (0, 10]$;
- step function $f_2(x) = 3$ if $x \in (0, 8]$ and $f_3(x) = 8$ if $x \in (8, 10]$;
- linear function $f_3(x) = 0.3x,\ x \in (0, 10]$;
- exponential function $f_4(x) = 0.15 \exp(0.6x - 3),\ x \in (0, 10]$;
- logistic function $f_5(x) = 3/\{1 + \exp(-2x + 10)\},\ x \in (0, 10]$;
- sinusoidal function $f_6(x) = 0.32\{x + sin(x)\},\ x \in (0, 10]$.

These functions are supposed to be evaluated at data of size $n = 100$ with standard deviation $\sigma_{\text{noise}} = 1$. The RMSE of the estimates is computed at the one hundred x values taken uniformly (equidistant) in the interval $(0, 10]$:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\Big(f(x_i) - \hat{f}(x_i)\Big)^2}, \tag{12}$$

where $\hat{f}(x)$ is the estimate of $f(x)$ and x_i are the n equally-spaced x-values. For the GP approximation developed in this paper, the maximum *a posteriori* estimate (Definition 3.3) is used as an estimate of $f(x)$, where N is fixed to fifty. Let us recall that this estimate depends only on the length hyper-parameter θ. The Matérn 5/2 covariance function

$$k_{\theta,\sigma}(x, x') = \sigma^2\left(1 + \frac{\sqrt{5}\mid x - x'\mid}{\theta} + \frac{5(x - x')^2}{3\theta^2}\right)\exp\left(-\frac{\sqrt{5}|x - x'|}{\theta}\right) \tag{13}$$

is used in the simulation, with σ fixed to 1 and θ estimated using the suited cross-validation method (see Cousin et al. (2016); Maatouk et al. (2015)). Table 1 shows the values of the parameter estimation $\hat{\theta}$.

In Table 2, the RMSE of the estimates is calculated for the finite-dimensional approximation of GPs, and it is compared with results of Gaussian process with and without projection given in Lin and Dunson (2014) and results of the regression spline method given in Shively et al. (2009). To ensure stability of results, the simulations have been repeated 5000 times. Table 2 shows that the finite-dimensional approximation of GPs outperforms regression splines and Gaussian process with and without projection.

To compare the proposed approach with the methodology based on the knowledge

of the derivatives of the GP at some input locations, the logistic artificial function $f(x) = 2/(1 + \exp(-8x + 4)),\ x \in [0, 1]$ defined in Riihimäki and Vehtari (2010) is considered. This function is supposed to be evaluated at data of size n with standard deviation $\sigma_{\text{noise}} = 0.5$. The squared exponential covariance function (13) is used. In

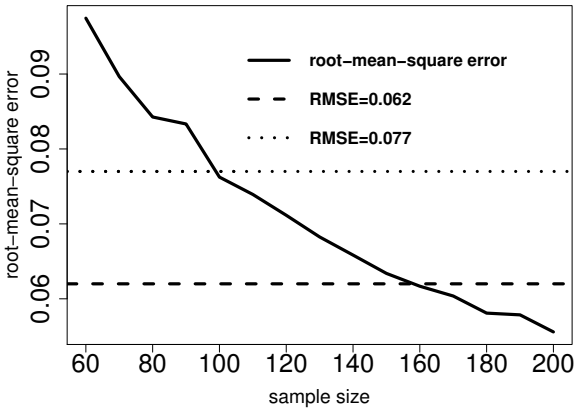


Figure 5. The root-mean-square error at different sample sizes together with the optimal values obtained in Riihimäki and Vehtari (2010).

Riihimäki and Vehtari (2010), the RMSE is equal to 0.077 (resp. 0.062) for $n = 100$ (resp. $n = 200$). In Figure 5, the root-mean-square error using the GP approximation is illustrated at different sample sizes together with the optimal values obtained by Riihimäki and Vehtari (2010). Notice that, we just need data at size $n = 160$ to reach the optimal value 0.062 obtained by Riihimäki and Vehtari (2010). The results are based on 1000 simulation replicates.

6. Conclusion

In this paper, a finite-dimensional approximation of Gaussian processes which satisfy linear inequality constraints everywhere (such as boundedness, monotonicity and convexity) and noisy observations is developed. We show that the generalization of the Kimeldorf-Wahba correspondence (see Kimeldorf and Wahba (1970)) proved in Bay et al. (2016) is conserved in the observations with Gaussian noise case. The performance of the proposed model in terms of predictive accuracy and uncertainty quantification is shown by a comparison with several recent models dealing with the same constraints.

References

Abrahamsen, P. and Benth, F. E. (2001). Kriging with inequality constraints. *Math. Geo.*, 33(6):719–744.

Agrell, C. (2019). Gaussian processes with linear operator inequality constraints. *arXiv e-prints*, page arXiv:1901.03134.

Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. Am. Math. Soc.*, 68:337–404.

Bachoc, F., Lagnoux, A., and Lpez-Lopera, A. F. (2019). Maximum likelihood estimation for gaussian processes under inequality constraints. *Electron. J. Statist.*, 13(2):2921–2969.

Bay, X., Grammont, L., and Maatouk, H. (2016). Generalization of the Kimeldorf-Wahba correspondence for constrained interpolation. *Electron. J. Statist.*, 10(1):1580–1595.

Bay, X., Grammont, L., and Maatouk, H. (2017). A new method for interpolating in a convex subset of a Hilbert space. *Comput. Optim. Appl.*, 68(1):95–120. doi:10.1007/s10589-017-9906-9.

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic Publishers.

Bhattacharya, R. and Kong, M. (2007). Consistency and asymptotic normality of the estimated effective doses in bioassay. *J. Stat. Plan. Infer.*, 137(3):643–658.

Bhattacharya, R. and Lin, L. (2010). An adaptive nonparametric method in benchmark analysis for bioassay and environmental studies. *Stat. Probab. Lett.*, 80(23-24):1947–1953.

Bhattacharya, R. and Lin, L. (2011). Nonparametric benchmark analysis in risk assessment: a comparative study by simulation and data analysis. *Sankhya B*, 73(1):144.

Bornkamp, B. and Ickstadt, K. (2009). Bayesian nonparametric estimation of continuous monotone functions with applications to dose–response analysis. *Biometrics*, 65(1):198–205.

Botts, C. (2013). An accept-reject algorithm for the positive multivariate normal distribution. *Comput. Statist.*, 28(4):1749–1773.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

Brunk, H., Barlow, R., Bartholomew, D., and Bremner, J. (1972). Statistical inference under order restrictions.(the theory and application of isotonic regression). Technical report, MISSOURI UNIV COLUMBIA DEPT OF STATISTICS.

Chopin, N. (2011). Fast simulation of truncated Gaussian distributions. *Stat. Comput.*, 21(2):275–288.

Cousin, A., Maatouk, H., and Rullière, D. (2016). Kriging of financial term-structures. *Eur. J. Oper. Res.*, 255(2):631 – 648.

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

Da Veiga, S. and Marrel, A. (2012). Gaussian process modeling with inequality constraints. *Annales de la faculté des sciences de Toulouse*, 21(3):529–555.

Delecroix, M., Simioni, M., and Thomas-Agnan, C. (1996). Functional estimation under shape constraints. *J. Nonparametr. Stat.*, 6(1):69–89.

Dette, H., Neumeyer, N., and Pilz, K. F. (2005). A note on nonparametric estimation of the effective dose in quantal bioassay. *J. Amer. Statist. Assoc.*, 100(470):503–510.

Freulon, X. and de Fouquet, C. (1993). Conditioning a Gaussian model with inequalities. In Soares, A., editor, *Geostatistics Tróia '92: Volume 1*, pages 201–212. Springer Netherlands, Dordrecht.

Golchi, S., Bingham, D., Chipman, H., and Campbell, D. (2015). Monotone emulation of computer experiments. *SIAM/ASA J. Uncertain. Quantif.*, 3(1):370–392.

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

Holmes, C. and Heard, N. (2003). Generalized monotonic regression using random change points. *Stat. Med.*, 22(4):623–638.

Jensen, B. S., Nielsen, J. B., and Larsen, J. (2013). Bounded Gaussian process regression. In *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6.

Jidling, C., Wahlström, N., Wills, A., and Schön, T. B. (2017). Linearly constrained Gaussian processes. In *Advances in Neural Information Processing Systems*, pages 1215–1224.

Kimeldorf, G. S. and Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Stat .*, 41(2):495–502.

Kleijnen, J. P. and Van Beers, W. C. (2013). Monotonicity-preserving bootstrapped kriging metamodels for expensive simulations. *J. Oper. Res. Soc.*, 64(5):708–717.

Kong, M. and Eubank, R. L. (2006). Monotone smoothing with application to dose-response curve. *Commun. Stat. Simul. Comput.*, 35(4):991–1004.

Koyejo, O., Lee, C., and Ghosh, J. (2013). Constrained Gaussian process regression for gene-disease association. In *2013 IEEE 13th International Conference on Data Mining Workshops*, pages 72–79.

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

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

Lyu, X., Binois, M., and Ludkovski, M. (2018). Evaluating gaussian process metamodels and sequential designs for noisy level set estimation.

Maatouk, H. and Bay, X. (2016). A new rejection sampling method for truncated multivariate Gaussian random variables restricted to convex sets. In Cools, R. and Nuyens, D., editors, *Monte Carlo and Quasi-Monte Carlo Methods*, pages 521–530. Springer International Publishing, Cham.

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

Maatouk, H., Roustant, O., and Richet, Y. (2015). Cross-validation estimations of hyper-parameters of Gaussian processes with inequality constraints. *Procedia Environmental Sciences*, 27:38 – 44. Spatial Statistics conference 2015.

Mammen, E. (1991). Estimating a smooth monotone regression function. *Ann. Stat.*, pages 724–740.

Morris, M. D., Mitchell, T. J., and Ylvisaker, D. (1993). Bayesian design and analysis of computer experiments: use of derivatives in surface prediction. *Technometrics*, 35(3):243–255.

Müller, H.-G. and Schmitt, T. (1988). Kernel and probit estimates in quantal bioassay. *J. Amer. Statist. Assoc.*, 83(403):750–759.

Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective (Adaptive Computation And Machine Learning Series). *The MIT Press*.

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

Pakman, A. and Paninski, L. (2014). Exact hamiltonian monte carlo for truncated multivariate Gaussians. *J. Comp. Graph. Stat.*, 23(2):518–542.

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

Philippe, A. and Robert, C. P. (2003). Perfect simulation of positive Gaussian distributions. *Stat. Comput.*, 13(2):179–186.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statist. Sci.*, 3(4):425–

441.

Ramsay, J. O. (1998). Estimating smooth monotone functions. *J. R. Stat. Soc. B*, 60(2):365–375.

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian processes for machine learning*. MIT Press, Cambridge.

Riihimäki, J. and Vehtari, A. (2010). Gaussian processes with monotonicity information. *J. Mach. Learn. Res.*, 9:645–652.

Robert, C. P. (1995). Simulation of truncated normal variables. *Stat. Comput.*, 5(2):121–125.

Robertson, T., Wright, F. T., and Dykstra, R. L. (1988). *Order Restricted Statistical Inference*. Wiley, New York.

Rudovic, O. and Pantic, M. (2011). Shape-constrained Gaussian process regression for facial-point-based head-pose normalization. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1495–1502. IEEE.

Saarela, O. and Arjas, E. (2011). A method for Bayesian monotonic multiple regression. *Scand. J. Statist.*, 38(3):499–513.

Salzmann, M. and Urtasun, R. (2010). Implicitly constrained Gaussian process regression for monocular non-rigid pose estimation. In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A., editors, *Advances in Neural Information Processing Systems 23*, pages 2065–2073. Curran Associates, Inc.

Shively, T. S., Sager, T. W., and Walker, S. G. (2009). A Bayesian approach to non-parametric monotone function estimation. *J. R. Stat. Soc. B*, 71(1):159–175.

Souris, A., Bhattacharya, A., and Pati, D. (2018). The soft multivariate truncated normal distribution with applications to bayesian constrained estimation.

T., G. and L., F. (2007). Generalized smooth monotonic regression in additive modeling. *J. Comp. Graph. Stat.*, 16(1):165–188.

Taylor, J. and Benjamini, Y. (2017). RestrictedMVN: multivariate normal restricted by affine constraints. https://cran.r-project.org/web/packages/restrictedMVN/index.html.

Wahl, F. and Espinasse, T. (2018). Simplex Regression: Multivariable parametric regression under shape constraints. working paper or preprint.

Wang, X. and Berger, J. O. (2016). Estimating shape constrained functions using Gaussian processes. *SIAM/ASA J. Uncertain. Quantif.*, 4(1):1–25.

Xuming, H. and Peide, S. (1996). Monotone B-spline smoothing. *J. Amer. Statist. Assoc.*, 93:643–650.

Zhang, J. and Lin, L. (2018). Bounded regression with Gaussian process projection. *arXiv preprint arXiv:1810.11881*.

Zhou, S., Giulani, P., Piekarewicz, J., Bhattacharya, A., and Pati, D. (2019). Reexamining the proton-radius problem using constrained gaussian processes. *Phys. Rev. C*, 99:055202.