Finite-Dimensional Gaussian Approximation with Linear Inequality Constraints

Andrés F. López-Lopera†, François Bachoc‡, Nicolas Durrande†§, and Olivier Roustant†

Abstract. Introducing inequality constraints in Gaussian processes can lead to more realistic uncertainties in learning a great variety of real-world problems. We consider the finite-dimensional Gaussian model from Maatouk and Bay [Math. Geosci., 49 (2017), pp. 557–582] which can satisfy inequality conditions everywhere (either boundedness, monotonicity, or convexity). Our contributions are threefold. First, we extend their approach in order to deal with sets of linear inequalities. Second, we explore different Markov chain Monte Carlo (MCMC) methods to approximate the posterior distribution. Third, we investigate theoretical and numerical properties of a constrained likelihood for covariance parameter estimation. According to experiments on both artificial and real data, our framework together with a Hamiltonian Monte Carlo sampler provides efficient results on both data fitting and uncertainty quantification.

Key words. asymptotic analysis, Gaussian processes regression, inference under constraints, MCMC

AMS subject classifications. 60G15, 62F12, 62F30, 62P30

DOI. 10.1137/17M1153157

1. Introduction. Gaussian processes (GPs) are one of the most famous nonparametric Bayesian frameworks for modeling stochastic processes. In principle, GP models place prior distributions over function spaces, and prior assumptions (e.g., smoothness, stationarity, sparsity) are encoded in covariance functions [1, 2, 3]. Because GPs provide a well-founded approach to learning, their properties have been explored in many decision tasks in regression (Kriging) and classification problems [1, 4]. Computer science, engineering, physics, biology, and neuroscience are some fields where GP models have been applied successfully [1, 5].

Despite the reliable performance of GPs, they provide less realistic uncertainties when physical systems satisfy inequality constraints [6, 7, 8]. Quantifying properly the uncertainties is crucial for understanding real-world phenomena. For example, in nuclear safety criticality assessment, experimental settings typically demand expensive and risky procedures to evaluate neutron productions. Hence, emulators are required to infer these production rates and should assume a priori that the output is positive and usually monotonic with respect to a given set of input parameters. In this sense, to obtain more accurate predictions, both conditions have to be considered in the uncertainty quantification. Other test cases where data exhibit specific inequality constraints are given in computer networking (monotonicity) [8], social system analysis (monotonicity or positivity) [10], and econometrics (monotonicity or positivity) [10].
Several studies have shown that including inequality constraints in GP frameworks can lead to more realistic uncertainty quantifications in learning from real data [7, 8, 9]. In most of the cases, it is assumed that the inequalities are satisfied on a finite set of input locations. Then, the posterior distribution is approximated given those constrained inputs. In practice, an alternative to deal either with positiveness, monotonicity, or convexity constraints is to use (iterated) integrals of positive processes (e.g., log-GPs [11]). However, those approaches have a density with zero mass in zero and are limited to specific inequality conditions. To the best of our knowledge, the framework from [6] is the only Gaussian approach proposed in the literature which satisfies specific inequalities everywhere in the input space. There, the GP samples are approximated in the finite-dimensional space of functions such as piecewise linear functions. It is shown in [12] that the posterior mode converges to the one provided by thin plate splines. This approach has been applied on several real-data (e.g., econometrics, geostatistics) [6, 10], resulting in more realistic uncertainties than unconstrained Kriging.

The framework proposed in [6] still presents some limitations. First, the focus is on either boundedness, monotonicity, or convexity conditions. Second, the proposed rejection sampling method for estimating the posterior [13] results in costly computations when either the order of the finite approximation increases or the inequality constraints become more complex. Third, the proposed leave-one-out technique for parameter estimation [14] restricts the optimal values to be on a finite grid of possible values and provides the same estimation of correlation parameters as for unconstrained GP parameters. In order to address these limitations, our contributions are threefold. First, we extend the framework to deal with general sets of linear inequality constraints. Second, we evaluate efficient Markov chain Monte Carlo (MCMC) algorithms that can be used to approximate the posterior distribution. Third, we investigate theoretical and numerical properties of the conditional likelihood for covariance parameter estimation. According to experiments on both artificial and real data, the resulting framework provides efficient results on both data fitting and uncertainty quantification.

This paper is organized as follows. In section 2, we briefly describe GP modeling with inequality constraints. In section 3, continuing with the finite-dimensional approach from [6], we propose a general formulation to deal with sets of linear inequalities. In section 4, we apply several MCMC techniques to approximate the posterior distribution, and we compare their performances with respect to exact Monte Carlo (MC) algorithms. In section 5, we write the conditional likelihood for the covariance parameter estimation, providing theoretical and empirical properties. In section 6, we assess our framework in two-dimensional (2D) Kriging tasks. Finally, in section 7, we summarize the conclusions, as well as the potential future works.

2. Gaussian process modeling with inequality constraints.

2.1. Finite-dimensional approximation. Let $Y$ be a zero-mean GP on $\mathbb{R}$ with covariance function $k$. Consider $x \in \mathcal{D}$ with compact input space $\mathcal{D} = [0, 1]$, and a set of knots $t_1, \ldots, t_m \in \mathbb{R}$. For simplicity, we will consider equally spaced knots $t_j = (j-1)\Delta_m$ with $\Delta_m = 1/(m-1)$, but this assumption can be relaxed. Then, define a finite-dimensional GP, denoted by $Y_m$, as the piecewise linear interpolation of $Y$ at knots $t_1, \ldots, t_m$:

\begin{equation}
Y_m(x) = \sum_{j=1}^{m} Y(t_j) \phi_j(x),
\end{equation}
Figure 1. Illustration of the finite-dimensional approximation of (1). (Left) Hat functions $\phi_j$ for $j = 1, \ldots, 6$. (Right) Approximation of the function $y(x) = \Phi\left(\frac{x-0.5}{0.2}\right)$, where $\Phi$ is the standard normal cumulative distribution function. Solid red and dashed black lines are the function $y$, and its finite approximation with six knots given by black crosses, respectively. Horizontal black dashed lines denote the bounds.

where $\phi_1, \ldots, \phi_m$ are hat basis functions given by

$$
\phi_j(x) := \begin{cases} 
1 - \frac{|x-t_j|}{\Delta_m} & \text{if } |x-t_j| \leq 1, \\
0 & \text{otherwise}.
\end{cases}
$$

We illustrate the finite-dimensional approach of (1) in Figure 1 for a deterministic function that satisfies two types of inequality constraints: boundedness and monotonicity (nondecreasing).

Now, let $\xi_j := Y(t_j)$ for $j = 1, \ldots, m$. We aim at computing the distribution of $Y_m$ conditionally on $Y_m \in \mathcal{E}$, where $\mathcal{E}$ is a convex set of functions defined by some inequality constraints. For instance, we may have

$$
\mathcal{E} = \mathcal{E}_\kappa := \begin{cases} 
\{ f \in C(D, \mathbb{R}) \text{ s.t. } \ell \leq f(x) \leq u \forall x \in D \} & \text{if } \kappa = 0, \\
\{ f \in C(D, \mathbb{R}) \text{ s.t. } f \text{ is nondecreasing} \} & \text{if } \kappa = 1, \\
\{ f \in C(D, \mathbb{R}) \text{ s.t. } f \text{ is convex} \} & \text{if } \kappa = 2,
\end{cases}
$$

which corresponds to boundedness, monotonicity, and convexity constraints (respectively). The benefit of using hat functions and the finite-dimensional approximation $Y_m$ is that satisfying the inequality conditions $Y_m \in \mathcal{E}$ is equivalent to satisfying only a finite number of inequality constraints [6]. More precisely, for many natural choices of $\mathcal{E}$, we have

$$
Y_m \in \mathcal{E} \iff \xi \in \mathcal{C},
$$

where $\mathcal{C}$ is a convex set of $\mathbb{R}^m$ and $\xi = [\xi_1, \ldots, \xi_m]$. For instance, for the convex set $\mathcal{E}_\kappa$ of (3), we have

$$
\mathcal{C} = \mathcal{C}_\kappa := \begin{cases} 
\{ c \in \mathbb{R}^m; \forall j = 1, \ldots, m : \ell \leq c_j \leq u \} & \text{if } \kappa = 0, \\
\{ c \in \mathbb{R}^m; \forall j = 2, \ldots, m : c_j \geq c_{j-1} \} & \text{if } \kappa = 1, \\
\{ c \in \mathbb{R}^m; \forall j = 3, \ldots, m : c_j - c_{j-1} \geq c_{j-1} - c_{j-2} \} & \text{if } \kappa = 2.
\end{cases}
$$
2.2. Conditioning with interpolation and inequality constraints. Consider the finite representation of GPs as in (1), given the interpolation and inequality constraints

\begin{equation}
Y_m(x) = \sum_{j=1}^{m} \xi_j \phi_j(x), \quad \text{s.t.} \quad \begin{cases} Y_m(x_i) = y_i \quad \text{(interpolation conditions)}, \\ Y_m \in \mathcal{E} \quad \text{(inequality conditions)}, \end{cases}
\end{equation}

where \( x_i \in \mathcal{D} \) and \( y_i \in \mathbb{R} \) for \( i = 1, \ldots, n \). One can note from (6) that noise-free observations are considered but a noise effect can be included assuming \( Y_m(x_i) + \varepsilon_i = y_i \) with Gaussian noise \( \varepsilon_i \sim \mathcal{N}(0, \sigma^2) \). This would lead to very similar developments to the ones below and the noise variance \( \sigma^2 \) can be estimated. We refer to [15] for further details on how to incorporate a noise effect in a constrained GP model. Given a design of experiment (DoE) \( \mathbf{x} = [x_1, \ldots, x_n]^T \), we have matricially

\[ Y_m = [Y_m(x_1), \ldots, Y_m(x_n)]^T = \Phi \xi, \]

where \( \Phi \) is the \( n \times m \) matrix defined by \( \Phi_{i,j} = \phi_j(x_i) \). Let \( \mathbf{y} = [y_1, \ldots, y_m]^T \) be a realization of \( Y \) at points \( x_1, \ldots, x_n \). From (4), the conditional distribution of \( Y_m \), under the inequality constraints \( Y_m \in \mathcal{E} \) and interpolation conditions \( Y_m(x_i) = y_i \) for \( i = 1, \ldots, n \), can be obtained from the conditional distribution of \( \xi \) given \( \xi \in \mathcal{C} \) and \( \Phi \xi = \mathbf{y} \).

Observe that the vector \( \xi \) of the values at the knots is a zero-mean Gaussian vector with covariance matrix \( \Gamma = (k(t_i, t_j))_{1 \leq i,j \leq m} \). Then, the distribution of \( \xi \) given both interpolation and inequality conditions is truncated multivariate normal:

\begin{equation}
\xi \sim \mathcal{N}(0, \Gamma) \quad \text{s.t.} \quad \begin{cases} \Phi \xi = \mathbf{y} \quad \text{(interpolation conditions)}, \\ \xi \in \mathcal{C} \quad \text{(inequality conditions)} \end{cases}
\end{equation}

with \( \mathcal{C} \) as in (4). For sampling purposes (see Algorithm 1), we need to compute the posterior mode which is given by the maximum of the probability density function of the posterior, i.e., \( \mu^*_\xi = \min \{ \xi^T \Gamma^{-1} \xi \mid \Phi \xi = \mathbf{y}, \xi \in \mathcal{C} \} \) (maximum a posteriori). Notice that \( \mu^*_\xi \) converges uniformly to the solution provided by thin plate splines when \( m \to \infty \) [12]. More details and theoretical properties are provided in [6, 12].

Figure 2 shows different Gaussian models for the example of Figure 1. We used a squared exponential (SE) covariance function with parameters \( (\sigma^2 = 1, \alpha = 0.2)^1 \) and we fixed \( m = 100 \). The posterior distribution was approximated via Hamiltonian Monte Carlo (HMC) [16]. From Figures 2(b) and 2(c), we observe that including the inequality constraints in the conditional distribution provides smaller confidence intervals compared to the ones given by the unconstrained GP. However, they do not satisfy both the boundedness and monotonicity conditions exhibited by the function \( y \). On the other hand, from Figure 2(d), imposing both conditions leads to a more accurate prediction and more realistic confidence intervals. Later, in section 3, we will detail how to obtain the results of Figure 2(d).

---

1SE covariance function: \( k_\theta(x-x') = \sigma^2 \exp \left\{ -\frac{(x-x')^2}{2\sigma^2} \right\} \) with \( \theta = (\sigma^2, \alpha) \).
3. Finite-dimensional Gaussian approximation with linear inequality constraints. Now, we consider the case where \( \mathcal{C} \) is composed by a set of \( q \) linear inequalities of the form

\[
\mathcal{C} = \left\{ \mathbf{c} \in \mathbb{R}^m; \forall k = 1, \ldots, q : \ell_k \leq \sum_{j=1}^m \lambda_{k,j} c_j \leq u_k \right\},
\]

where the \( \lambda_{k,j} \)'s encode the linear operations, and the \( \ell_k \)'s and \( u_k \)'s represent the lower and upper bounds. Notice that the convex sets \( \mathcal{C}_k \) of (5) are particular cases of \( \mathcal{C} \). Denote \( \mathbf{\Lambda} = (\lambda_{k,j})_{1 \leq k \leq q, 1 \leq j \leq m}, \mathbf{l} = (\ell_k)_{1 \leq k \leq q}, \) and \( \mathbf{u} = (u_k)_{1 \leq k \leq q} \). Hence, (7) is written

\[
\xi \sim \mathcal{N}(\mathbf{0}, \Gamma) \quad \text{s.t.} \quad \begin{cases} 
\Phi \xi = \mathbf{y} & \text{(interpolation conditions)}, \\
l \leq \mathbf{\Lambda} \xi \leq \mathbf{u} & \text{(inequality conditions)}.
\end{cases}
\]

We further assume that \( q \geq m \) and that \( \mathbf{\Lambda} \) has rank \( m \). By the rank-nullity theorem (see, e.g., [17]), it implies that \( \mathbf{\Lambda} \) is injective. In particular a linear system of the form \( \mathbf{\Lambda} \xi = \eta \) admits a unique solution \( \xi \) when \( \eta \) is in the image space of \( \mathbf{\Lambda} \). This assumption is verified in many practical situations, up to adding inactive constraints. For instance, the monotonicity...
Algorithm 1 Sampling from the finite-dimensional GP with linear inequality constraints.

1: Procedure: Sampling from $\xi | \{ \Phi \xi = y, l \leq \Lambda \xi \leq u \}$, where $\xi \sim \mathcal{N}(0, \Gamma)$
2: Input: $\gamma, \Gamma \in \mathbb{R}^{m \times m}, \Phi \in \mathbb{R}^{n \times m}, \Lambda, l, u.$
3: Compute the conditional mean and covariance of $\xi | \{ \Phi \xi = y \}$
4: $\mu = \Gamma \Phi^\top (\Phi \Gamma \Phi^\top)^{-1}y$, and
5: $\Sigma = \Gamma - \Gamma \Phi^\top (\Phi \Gamma \Phi^\top)^{-1} \Phi \Gamma$.
6: Solve the quadratic problem in $\mathbb{R}^m$: $\mu^* = \min_{\xi \in \mathbb{R}^m} \{ \xi^\top \Gamma^{-1} \xi | \Phi \xi = y, l \leq \Lambda \xi \leq u \}$.
7: Sample from the truncated multinormal distribution
8: $\Lambda \xi | \{ \Phi \xi = y, l \leq \Lambda \xi \leq u \} \sim T \mathcal{N}(\Lambda \mu, \Lambda \Sigma \Lambda^\top, l, u)$.
9: Define $\eta = \Lambda \xi$, and solve the linear system to obtain the sample $\xi$.
10: Remark: use the posterior mode $\nu^*_\xi = \Lambda \mu^*_\xi$ as a starting state for an MCMC sampler (see section 4).

condition $\xi_1 \leq \cdots \leq \xi_m$, which involves only $q = m - 1$ (linearly independent) conditions, can be made compatible by adding the condition $-\infty \leq \xi_1$ (and/or $\xi_m \leq \infty$).

We now explain how to sample $\xi$ from (8). First, we compute the conditional distribution given the interpolation constraints $\xi | \{ \Phi \xi = y \}$. Since $\xi \sim \mathcal{N}(0, \Gamma)$, then $\Phi \xi \sim \mathcal{N}(0, \Phi \Gamma \Phi^\top)$ and the conditional distribution $\xi | \{ \Phi \xi = y \}$ is also Gaussian $\mathcal{N}(\mu, \Sigma)$ [1] with

$$
\mu = \Gamma \Phi^\top (\Phi \Gamma \Phi^\top)^{-1}y, \quad \text{and} \quad \Sigma = \Gamma - \Gamma \Phi^\top (\Phi \Gamma \Phi^\top)^{-1} \Phi \Gamma.
$$

Therefore, we have $\Lambda \xi | \{ \Phi \xi = y \} \sim \mathcal{N}(\Lambda \mu, \Lambda \Sigma \Lambda^\top)$. Let $T \mathcal{N}(m, C, a, b)$ be the truncated multinormal distribution with mean vector $m$, covariance matrix $C$, and bound vectors $(a, b)$ such that $a \leq b$. Thus, the posterior distribution of (8) is obtained from

$$
\Lambda \xi | \{ \Phi \xi = y, l \leq \Lambda \xi \leq u \} \sim T \mathcal{N}(\Lambda \mu, \Lambda \Sigma \Lambda^\top, l, u).
$$

Notice that the inequality conditions are encoded in the posterior mean $\Lambda \mu$, the posterior covariance $\Lambda \Sigma \Lambda^\top$, and bounds $(l, u)$. Finally, the posterior mode is given by $\nu^*_\xi = \Lambda \mu^*_\xi$, where $\mu^*_\xi$ is the solution provided in subsection 2.2. The truncated multinormal of (10) can be approximated using MCMC algorithms. Denoting $\eta = \Lambda \xi$, notice that the samples for $\xi$ can be obtained by using the ones obtained for $\eta$ if the linear system is solved. Indeed, as mentioned above, we assumed that $\Lambda$ has rank $m$, which implies that the solution of $\Lambda \xi = \eta$ exists and is unique. The whole sampling scheme is summarized in Algorithm 1.

Now, we illustrate some examples where the proposed framework enables us to address different types of inequality conditions. The posterior is approximated via HMC. (More details about HMC are given in section 4.)

Example 1. We continue with the example of Figure 1. As we can fix the structure of the linear inequalities $(\Lambda, l, u)$, we can impose both boundedness and monotonicity conditions in the constrained GP. One way to do this is to encode them individually. Let $l_1 \leq \Lambda_1 \xi \leq u_1$ and $l_2 \leq \Lambda_2 \xi \leq u_2$ be the sets of conditions to satisfy boundedness and monotonicity constraints, respectively. Then, we can build an extended set of inequalities $l \leq \Lambda \xi \leq u$ by stacking the constraints (i.e., $\Lambda = [\Lambda_1, \Lambda_2]^\top, l = [l_1^\top, l_2^\top]^\top, u = [u_1^\top, u_2^\top]^\top$), so that Algorithm 1 can be used. Notice that one can encode the same information in a reduced set...
of linear inequalities. Instead of encoding independently the boundedness and monotonicity constraints, which requires \( q = 2m - 1 \) inequalities, one can impose boundedness conditions only for the first and last knot, and monotonicity conditions for all the knots except the first one. Due to monotonicity, the intermediate knots will also satisfy the boundaries. In this way, we only need \( q = m + 1 \) conditions. In many other cases, the size of specific sets of linear constraints can be reduced. However for general discussions, we will use the full extended set and we will apply efficient samplers to approximate the posterior.

**Example 2.** Notice from the previous example that the extension to more than two sets of inequalities is straightforward. Consider, for instance, \( Q \) different sets of conditions. We can build the posterior from (10) with \( \mathbf{A} = [\mathbf{A}_1^\top, \ldots, \mathbf{A}_Q^\top]^\top, \mathbf{l} = [l_1^\top, \ldots, l_Q^\top]^\top, \) and \( \mathbf{u} = [u_1^\top, \ldots, u_Q^\top]^\top, \) and apply Algorithm 1. Figure 3 shows an example with the target function \( y(x) = x^2, \) satisfying three types of inequality constraints: boundedness, monotonicity, and convexity. We proposed different models satisfying one or more inequality constraints. We used an SE covariance function with parameters \((\sigma^2 = 1.0, \alpha = 0.2).\) By imposing the three conditions, we obtain samples that also satisfy the three types of constraints.

**Example 3.** Since the bounds \((\mathbf{l}, \mathbf{u})\) are not forced to be the same everywhere, it is possible to fix specific constraints over nonoverlapping intervals. For instance, if the interval is partitioned into \( G \) subintervals, we consider the corresponding partition \( \mathbf{x} = [x_1, \ldots, x_G]^\top. \) Then, we can impose different types of inequality conditions in each group by considering the same structure used in Example 2. Figure 4 shows an example where the function \( y \) satisfies different behaviors in two nonoverlapping intervals. The output increases monotonically and peaks at \( y(0.4) = 1.0.\) This kind of profile is met in different applications (e.g., step responses in control theory, protein profiles in molecular biology) [5, 18]. We trained three models satisfying different conditions. For the case of multiple constraints, we imposed boundedness and monotonicity. For the case of sequential conditions, we divided the profile in two nonoverlapping intervals satisfying different types of constraints. We used a SE covariance function with parameters \((\sigma^2 = 1.0, \alpha = 0.2).\) By imposing sequentially the constraints, we obtain less restricted uncertainties and more accurate models for data fitting.
4. Simulating from the posterior distribution. As shown in (10), the posterior distribution \( \Lambda \bfitxi \mid \{ \Phi \bfitxi = y, l \leq \Lambda \bfitxi \leq u \} \) is truncated multinormal. It is supported on \( \mathbb{R}^q \), where \( q \geq m \) is defined in section 3. Notice that \( m \) should be chosen large enough for better approximations. An MC algorithm based on rejection sampling was proposed in [13] using the posterior mode. This method, called rejection sampling from the mode (RSM), is an exact sampler that provides independent and identically distributed sample paths. However, the acceptance rate from RSM decreases when \( m \) gets larger, providing a poor performance for high-dimensional spaces. Another MC-based exact sampler using the separation-of-variables technique from [19] was introduced by [20] to deal with truncated multinormals in higher dimensions. As in [19], [20] can both simulate multinormals under linear constraints and estimate the probabilities that these constraints are satisfied, via minimax exponential tilting (ET). Since RSM and ET are exact methods, we will use them as gold standards to evaluate the performance of the MCMC techniques that we describe now.

4.1. MCMC for truncated multinormal distributions. MCMC approaches use a Markov chain to sample the posterior distribution, providing correlated samples but with a higher acceptance rate. Recently, efficient algorithms have been proposed for truncated multinormal distributions such as Gibbs sampling [21], Metropolis–Hastings (MH) [5], and HMC [16]. In this section, we apply them to simulate from the distribution of (10).

Gibbs sampling. Algorithms based on Gibbs sampling are widely used to sample from truncated multinormals due to their easy implementation and their reliable performances [5, 22]. They sample each variable in turn conditional on the values of the other ones [5]. Therefore, sampling from a truncated multinormal is reduced to sampling sequentially from conditional truncated (univariate) normals. Unlike RSM, there is no rejection step. However, the “single site updating” property may produce strong correlations, requiring discarding intermediate samples (thinning effect). Several studies have proposed efficient algorithms to obtain less correlated sample paths (e.g., collapsed Gibbs sampling, blocked Gibbs sampling) [5]. In this paper, we will use the fast Gibbs sampler proposed in [21].
**Metropolis–Hastings.** MH-based algorithms propose to move all the coordinates at a time in each step to obtain less correlated simulations. Given a proposed state $x'$, we either accept or reject the new state according to a given acceptance rule [5]. If the proposal is accepted, the new state is $x'$; otherwise the new state remains at the previous state $x$. For multilinear distributions, a symmetric Gaussian proposal is commonly used, i.e., $q(x'|x) = N(x, \eta \Sigma x'|x)$, where $\eta$ is a scale factor. This approach is known as the random walk Metropolis algorithm [5]. One can increase the acceptance rate by tuning properly the value of $\eta$. In this paper, we assume that $\Sigma x'|x$ is given by the covariance function of the posterior distribution we want to approximate, i.e., $\Sigma x'|x = \Lambda \Sigma \Lambda^\top$ with $\Sigma$ defined as in (9).

**Hamiltonian Monte Carlo.** Today hybrid methods have been subject to great attention from the statistical community due to the inclusion of physical interpretation that may provide useful intuition [22, 23]. In [24], an efficient hybrid approach was introduced using the properties of Hamiltonian dynamics. Later in [23], the hybrid approach from [24] was extended to statistical applications and was introduced formally as HMC. The Hamiltonian dynamics provide distant proposal distributions producing less correlated sample paths without diminishing the acceptance rate. In this paper, we use the HMC-based approach for truncated multnormals introduced in [16].

### 4.2. Results.

In Table 1, we evaluate the efficiency of the MC and MCMC approaches described in subsection 4.1 on the examples from Figure 2. In order to reduce the simulation cost, we used $m = 30$ hat basis functions. Hence, the problem is to sample a vector of length 30 from a truncated multnormal distribution. Note from Algorithm 1 that the posterior mode is used as the starting state for the MCMC samplers. As a result, the chains are initialized into high probability regions, and we have to “burn” only a small amount of simulations in order to obtain samples that are appear to be independent of the initialization location. Therefore, we only burned the first 100 simulations from all the MCMC samplers. We set the tuning hyperparameters such that the effective sample size (ESS) is within the ranges

<table>
<thead>
<tr>
<th>Toy example</th>
<th>Method</th>
<th>CPU time ([s])</th>
<th>ESS ([\times 10^4]) ([90%: 50%: 10%])</th>
<th>mvESS ([\times 10^4\pi^{-1}])</th>
<th>TN-ESS ([\times 10^4\pi^{-1}])</th>
<th>Hyperparameter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toy</td>
<td>RSM</td>
<td>99.06</td>
<td>(0.93, 0.98, 1.00)</td>
<td>1.22</td>
<td>0.01</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>ET</td>
<td>0.44</td>
<td>(0.94, 0.97, 1.00)</td>
<td>1.17</td>
<td>2.14</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Gibbs</td>
<td>3.54</td>
<td>(0.95, 0.98, 1.00)</td>
<td>1.16</td>
<td>0.27</td>
<td>thining = 200</td>
</tr>
<tr>
<td></td>
<td>MH</td>
<td>52.21</td>
<td>(0.98, 1.00, 1.00)</td>
<td>1.21</td>
<td>0.02</td>
<td>$\eta = 1$</td>
</tr>
<tr>
<td></td>
<td>HMC</td>
<td>0.44</td>
<td>(0.90, 0.93, 1.00)</td>
<td>1.26</td>
<td>2.04</td>
<td>-</td>
</tr>
<tr>
<td>Toy</td>
<td>RSM</td>
<td>190.62</td>
<td>(1.00, 1.00, 1.00)</td>
<td>1.21</td>
<td>0.01</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>ET</td>
<td>0.77</td>
<td>(0.95, 0.98, 1.00)</td>
<td>1.18</td>
<td>1.23</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Gibbs</td>
<td>3.04</td>
<td>(1.00, 1.00, 1.00)</td>
<td>1.15</td>
<td>0.33</td>
<td>thining = 200</td>
</tr>
<tr>
<td></td>
<td>MH</td>
<td>96.64</td>
<td>(0.98, 1.00, 1.00)</td>
<td>1.23</td>
<td>0.01</td>
<td>$\eta = 1$</td>
</tr>
<tr>
<td></td>
<td>HMC</td>
<td>0.33</td>
<td>(0.89, 0.94, 0.98)</td>
<td>1.28</td>
<td>2.70</td>
<td>-</td>
</tr>
<tr>
<td>Toy</td>
<td>RSM</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>ET</td>
<td>41.16</td>
<td>(0.99, 1.00, 1.00)</td>
<td>1.23</td>
<td>0.02</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Gibbs</td>
<td>40.28</td>
<td>(0.37, 0.6, 0.91)</td>
<td>1.09</td>
<td>0.01</td>
<td>thining = 1000</td>
</tr>
<tr>
<td></td>
<td>MH</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>HMC</td>
<td>12.92</td>
<td>(0.85, 0.93, 1.00)</td>
<td>1.26</td>
<td>0.07</td>
<td>-</td>
</tr>
</tbody>
</table>

**Table 1**

Efficiency of MC/MCMC samplers (by rows) in terms of ESS-based indicators (by columns). Samplers: RSM [13], ET [20], Gibbs sampling (Gibbs) [21], MH [5], HMC [16]. Indicators: ESS: ESS = $n_s/(1+2\sum_{k=1}^{s} \hat{\rho}_k)$ [26], mvESS [27], TN-ESS [28].
produced by both RSM and ET (grey columns). The ESS is a heuristic used commonly to evaluate the quality of correlated sample paths, and it gives an intuition on how many samples from the path can be considered independent [25]. A standard ESS is given by
\[ \text{ESS} = n_s / (1 + 2 \sum_{k=1}^{n_s} \rho_k), \]
where \( n_s \) is the size of the sample path and \( \rho_k \) is the sample autocorrelation with lag \( k \). However, the drawback of this indicator is that it accepts negative correlations to evaluate the quality of mean estimators (e.g., for variance reduction). Thus, we suggest to use the initial convex sequence estimator proposed in [26] in order to obtain correlations to evaluate the quality of mean estimators (e.g., for variance reduction). Thus, we suggest to use the initial convex sequence estimator proposed in [26] in order to obtain correlations.

\[ (12) \]

\[ \text{ESS (initial convex sequence)} = \text{ESS}^{\text{conv}} = \text{ESS}^{\text{conv}}(n), \]

with \( K \) the initial convex sequence estimator proposed in [26] in order to obtain correlations.

Finally, using the procedure proposed in [28], we test the efficiency of each method by computing the time normalized ESS (TN-ESS) at \( q_{10\%} \) (worst case) using the CPU time in seconds, i.e., TN-ESS = \( q_{10\%}(\text{ESS})/\text{(CPU time)} \).

Table 1 shows the efficiency of MC/MCMC algorithms in terms of ESS indicators. Notice that for the two examples of Figures 2(b) and 2(c), the MC/MCMC techniques tend to produce similar ESS intervals, but RSM and MH are the most expensive procedures due to their high rejection rates. Although the Gibbs sampler requires discarding a large amount of simulations in order to be within reasonable ESS ranges, it also yields accurate results in both efficiency and CPU time. In general, both ET and HMC methods yield more efficient results than the other samplers in the first two examples. For more complex constraints as in the example of Figure 2(d), the efficiency is reduced dramatically for all the methods. For example, the acceptance rates of both RSM and MH are so small that sampling was not feasible in a reasonable time. For the other methods, the TN-ESS rates are smaller but HMC still gives a reasonable value (three times larger than for ET), which leads us to conclude that HMC is an efficient sampler for the proposed framework.

5. Covariance parameter estimation with inequality constraints.

5.1. Conditional maximum likelihood. Let \( \{k_\theta; \theta \in \Theta\} \) with \( \Theta \subset \mathbb{R}^p \) be a parametric family of covariance functions. We assume in this section that the zero-mean GP \( Y \) has covariance function \( k_\theta \) for an unknown \( \theta^* \in \Theta \). We consider the problem of estimating \( \theta^* \). Commonly, \( \theta^* \) is estimated by maximizing the unconstrained Gaussian likelihood \( p_\theta(Y_m) \) with respect to \( \theta \in \Theta \) (maximum likelihood (ML)) with \( Y_m = [Y_m(x_1), ..., Y_m(x_n)]^T \). Let \( \mathcal{L}_m(\theta) \) be the log likelihood of \( \theta \)

\[ (11) \]

\[ \mathcal{L}_m(\theta) = \log p_\theta(Y_m) = -\frac{1}{2} \log(\det(K_\theta)) - \frac{1}{2} Y_m^T K_\theta^{-1} Y_m - \frac{n}{2} \log 2\pi \]

with \( K_\theta = \Phi \Gamma_\theta \Phi^T \) and \( \Gamma_\theta = (k_\theta(t_i, t_j))_{1 \leq i, j \leq m} \). Then, the ML estimator (MLE) is

\[ \hat{\theta}_\text{MLE} = \arg \max_{\theta \in \Theta} \mathcal{L}_m(\theta). \]
When we maximize the likelihood in (12), we are looking for a parameter $\theta$ that improves the ability of our model to explain the data [1]. However, because the unconstrained ML itself does not take into account the constraints $\xi \in C$, the estimated $\hat{\theta}_{\text{MLE}}$ may produce less realistic models. Here, we suggest to use the constrained likelihood. Let $p_\theta(Y_m|\xi \in C)$ be the conditional probability density function of $Y_m$ given $\xi \in C$, when $Y$ has covariance function $k_\theta$. By using Bayes’s theorem, the constrained log likelihood $L_{C,m}(\theta) = \log p_\theta(Y_m|\xi \in C)$ is

$$L_{C,m}(\theta) = \log p_\theta(Y_m) + \log P_\theta(\xi \in C|\Phi_\xi = Y_m) - \log P_\theta(\xi \in C),$$

(13)

where the first term is the unconstrained log-likelihood, and the last two terms depend on the inequality constraints. Then, the constrained ML (CML) estimator is given by

$$\hat{\theta}_{\text{CML}} = \arg \max_{\theta \in \Theta} L_{C,m}(\theta).$$

(14)

Notice that $P_\theta(\xi \in C|\Phi_\xi = Y_m)$ and $P_\theta(\xi \in C)$ are Gaussian orthant probabilities. As they have no explicit expressions, numerical procedures have been investigated [20, 19]. The estimator proposed in [19] is based on a separation-of-variables transformation which reduces the problem to standard numerical multiple integration algorithms. On the other hand, as briefly described in section 4, the estimator from [20] efficiently deals with hitherto intractable Gaussian integrals in high dimensions via minimax exponential tilting. In this paper, we used the estimator from [20] in further experiments. Hence, the likelihood evaluation and optimization of (13) and (14) have to be done numerically.

5.2. Simulation study. To assess the performance of the estimator of (14), we simulated sample paths from a zero-mean constrained GP $Y$ using a Matérn 5/2 covariance function with $\theta^* = (1, 0.2)$.

\(^2\) We sampled 100 realizations of $Y$ on $D = [0, 1]$ such that $Y \in [-1, 1]$. Then, for each realization, we trained a constrained model (CM) assuming boundedness conditions with bounds $[-1, 1]$. We used 10 training points regularly spaced in $D$ and $m = 50$ hat basis functions.\(^3\) For ML and CML optimizations, we used multistart with ten initial vectors of covariance parameters located on a maximin Latin hypercube DoE with $\rho^2 \in [0, 2]$ and $\alpha \in [0.04, 0.40]$.\(^4\) We used the Nlopt optimization tools from [30], and we tested the different gradient-based optimisers. After some tests, we concluded that the globally convergent method of moving asymptotes (MMA) [31] yielded more stable results for estimating the covariance parameters. As the parameters of the Matérn 5/2 covariance function are nonmicroergodic for 1D input spaces, they cannot be estimated consistently [32]. Therefore, we evaluated the quality of the likelihood estimators using the consistently estimable ratio $\rho = \sigma^2/\alpha^5$. In Figure 5(a), we show the boxplots of the estimated ratios obtained with the 100 simulations drawn from the GP. Notice that the estimated logged ratios $\log \hat{\rho}_{\text{MLE}}$ and $\log \hat{\rho}_{\text{CML}}$ are reasonably close to the true value $\log \rho^* = \log(1^2/0.2^5)$, but the one using cMLE is slightly better in terms of variance and bias.

\(^2\)Matérn 5/2 kernel function: $k_\theta(x - x') = \sigma^2(1 + \sqrt{5}(x - x') + \frac{5}{3}(x - x')^2)\exp(-\sqrt{5}|x - x'|)$ with $\theta = (\sigma^2, \alpha)$.

\(^3\)In this experiment, we manually tuned the number of basis function $m$. We used different values of $m = 25, 50, 100, 150, 200$, and we observed that results of Figure 5 remained stable after $m = 50$.

\(^4\)A maximin Latin hypercube DoE is a space-filling design consisting in the iterative maximization of the distance between two closest design points from a random Latin hypercube design. In this paper, we used the simulated annealing routine maximinSA_LHS from the R package DiceDesign [29].
We also evaluate the efficiency of the two estimators in terms of prediction accuracy. For each realization, we estimated the covariance parameters $\theta^*$ by MLE and cMLE. We then simulated the posterior at 50 new regularly spaced locations using the estimated covariance parameter $\hat{\theta}$. The conditional sample paths were simulated via HMC. We used the $Q^2$ and coverage accuracy (CA) criteria to assess the quality of predictions over the 50 new values. Denoting by $n_t$ the number of test points, $z_1, \ldots, z_{n_t}$ and $\hat{z}_1, \ldots, \hat{z}_{n_t}$, the sets of test and predicted observations (respectively), then $Q^2 = 1 - \sum_{i=1}^{n_t} (\hat{z}_i - z_i)^2 / \sum_{i=1}^{n_t} (\bar{z} - z_i)^2$, where $\bar{z}$ is the average of the test data. Notice that for noise-free observations, the $Q^2$ indicator is equal to one if the predictors $\hat{z}_1, \ldots, \hat{z}_{n_t}$ are exactly equal to the test data (ideal case), zero if they are equal to the constant prediction $\bar{z}$, and negative if they perform worse than $\bar{z}$. On the other hand, CA assesses the quality of predictive variances $\hat{\sigma}_i^2$ for $i = 1, \ldots, n_t$. In this paper, we use one standard deviation intervals $(\hat{z}_i \pm \hat{\sigma}_i)$ which should provide a pointwise coverage of the test data around 68%. Departure from $\text{CA}_{\pm \sigma} = 0.68$ may indicate that the confidence level is overestimated (resp., underestimated) for coverage values of $\text{CA}_{\pm \sigma} > 0.68$ (resp., $\text{CA}_{\pm \sigma} < 0.68$) of the predictive variances.

Figure 5 shows the inferred sample paths for one realization using 5(d) $\hat{\theta}_{\text{MLE}}$, 5(e) $\hat{\theta}_{\text{cMLE}}$, and 5(f) $\theta^*$. We observe that, in the three cases, the models tend to fit properly the test

---

**Figure 5.** Assessment of the likelihood (ML) and conditional likelihood (CML) estimators for 100 samples drawn from a GP with true parameters $\theta^* = (1, 0.2)$, and satisfying the bounds $[-1, 1]$. (a) Estimated values of the log-ratio $\log \rho^* = \log(1^2/0.2^2)$ (dashed green line) using MLE and cMLE. Predictive accuracies are evaluated using the (b) $Q^2$ and (c) $\text{CA}_{\pm \sigma}$ criteria. In (c), the horizontal dashed green line represents the 68% pointwise coverage. Predictions are shown for one sample using (d) $\hat{\theta}_{\text{MLE}}$, (e) $\hat{\theta}_{\text{cMLE}}$, and (f) $\theta^*$. For the predictions, the panel description is the same as Figure 2.
data with accurate confidence intervals. According to Figures 5(b) and 5(c), we see that they provide $Q^2$ and CA$_{\pm\sigma}$ median values close to the ones obtained when the true $\theta^*$ is used. Although the predictive accuracies obtained using cMLE are better than for MLE in terms of bias, we observe larger variances in the CA$_{\pm\sigma}$ criterion for cMLE. We also compute the lengths of the one standard deviation intervals and we observed that cMLE provides smaller intervals than the ones by MLE. This is consistent with the fact that the confidence level is overestimated for MLE (Figure 5(c)). Since the Gaussian orthant terms from the conditional likelihood of (13) have to be approximated, we believe that this affects the effectiveness of cMLE. Furthermore, existing estimators of Gaussian orthant probabilities present some numerical instabilities limiting the CML optimization routine and providing suboptimal results. Finally, notice that MLE also provides reliable predictions. This suggests that, if we properly take into account the inequality constraints in the posterior distribution, the unconstrained ML optimization can be used for practical implementation.

5.3. Asymptotic properties. Now, we study the asymptotic properties of likelihood-based estimators for constrained GPs. We consider the fixed-domain asymptotic setting [33], with a dense sequence of observation points in a bounded domain. It should be noted that, when the GP is not constrained, significant contributions have been provided to study the consistency or asymptotic normality of the ML estimator [32, 34, 35, 36, 37]. In this paper, we show that, loosely speaking, any consistency result for ML with unconstrained GPs is preserved when adding either boundedness, monotonicity, or convexity constraints. Furthermore, this consistency occurs for both the unconditional and conditional likelihood functions.

For $\kappa \in \{0, 1, 2\}$, let $Y$ be a GP with $C^\kappa$ trajectories on a bounded set $\mathbb{X} \subset \mathbb{R}^d$. Let $\mathcal{E}_\kappa$ be one of the following convex sets of functions:

$$(15)$$

$$\mathcal{E}_\kappa = \begin{cases} 
 f : \mathbb{X} \rightarrow \mathbb{R}, f \text{ is } C^0 \text{ and } \forall x \in \mathbb{X}, \ell \leq f(x) \leq u & \text{if } \kappa = 0, \\
 f : \mathbb{X} \rightarrow \mathbb{R}, f \text{ is } C^1 \text{ and } \forall x \in \mathbb{X}, \forall i = 1, \ldots, d, \frac{\partial}{\partial x_i} f(x) \geq 0 & \text{if } \kappa = 1, \\
 f : \mathbb{X} \rightarrow \mathbb{R}, f \text{ is } C^2 \text{ and } \forall x \in \mathbb{X}, \frac{\partial^2}{\partial x_i \partial x_j} f(x) \text{ is a nonnegative definite matrix} & \text{if } \kappa = 2.
\end{cases}$$

For the purpose of asymptotic analysis, we do not consider the hat basis functions anymore, and we focus on the GP $Y$ and the observation vector $Y_n = [Y(x_1), \ldots, Y(x_n)]^\top$. We study the (unconstrained) likelihood function based on $p_\theta(Y_n)$ and the constrained likelihood function based on $p_\theta(Y_n | Y \in \mathcal{E}_\kappa)$. Notice that these quantities are more challenging to evaluate in practice than for subsections 5.1 and 5.2, but the purpose is a theoretical analysis.

In Proposition 5.1, we prove that if ML is consistent, when considering the (unconditional) distribution of $Y$, then it remains consistent when conditioning to $Y \in \mathcal{E}_\kappa$. In Proposition 5.2, we prove that, under mild conditions (such as $Y$ being $\kappa$ times continuously differentiable), implying the consistency of the ML estimator with the (unconditional) distribution of $Y$, the CML remains consistent when adding the constraint $Y \in \mathcal{E}_\kappa$. The proofs of Propositions 5.1 and 5.2 require supplementary conditions and lemmas, which are given in Appendix A.

Proposition 5.1. Let $Y$ be a zero-mean GP on a bounded set $\mathbb{X} \subset \mathbb{R}^d$ with covariance function $k$ satisfying Condition A.1. Let $\Theta$ be a compact set on $(0, \infty)^{d+1}$. Let $k_\theta$ be the covariance function of $x \rightarrow \sigma^2 Y(\alpha_1 x_1, \ldots, \alpha_d x_d)$ for $\theta = (\sigma^2, \alpha_1, \ldots, \alpha_d) \in \Theta$. Let $\theta^* = (1, \ldots, 1)$. Remark that $k = k_{\theta^*}$ and assume that $\theta^* \in \Theta$. Let $(x_i)_{i \in \mathbb{N}}$ be a dense sequence in
X. Let $Y_n = [Y(x_1), ..., Y(x_n)]^\top$. Let

$$\mathcal{L}_n(\theta) = -\frac{1}{2} \log(\det(R_\theta)) - \frac{1}{2} Y_n^\top R_\theta^{-1} Y_n - \frac{n}{2} \log 2\pi$$

with $R_\theta = (k_\theta(x_i,x_j))_{1 \leq i,j \leq n}$. Let $\hat{\theta} \in \arg\max_{\theta \in \Theta} \mathcal{L}_n(\theta)$. Assume that for all $\varepsilon > 0$,

$$P(\|\hat{\theta} - \theta^*\| \geq \varepsilon) \xrightarrow{n \to \infty} 0.$$ 

Let $\kappa \in \{0, 1, 2\}$. Let $\mathcal{E}_\kappa$ be as in (15). Then, we have $P(Y \in \mathcal{E}_\kappa) > 0$ from Lemmas A.3 to A.5, and thus

$$P(\|\hat{\theta} - \theta^*\| \geq \varepsilon \mid Y \in \mathcal{E}_\kappa) \xrightarrow{n \to \infty} 0.$$ 

Proof. We have

$$P(\|\hat{\theta} - \theta^*\| \geq \varepsilon \mid Y \in \mathcal{E}_\kappa) = \frac{P(\|\hat{\theta} - \theta^*\| \geq \varepsilon, Y \in \mathcal{E}_\kappa)}{P(Y \in \mathcal{E}_\kappa)} \leq \frac{P(\|\hat{\theta} - \theta^*\| \geq \varepsilon)}{P(Y \in \mathcal{E}_\kappa)}.$$ 

Since $P(Y \in \mathcal{E}_\kappa) > 0$ is fixed, and $P(\|\hat{\theta} - \theta^*\| \geq \varepsilon) \xrightarrow{n \to \infty} 0$, the result follows.

Proposition 5.2. We use the same notation and assumptions as in Proposition 5.1. Let $\kappa \in \{0, 1, 2\}$ be fixed. Let $P_\theta$ be the distribution of $Y$ with covariance function $k_\theta$. Let

$$\mathcal{L}_{\kappa,n}(\theta) = \mathcal{L}_n(\theta) + \log P_\theta(Y \in \mathcal{E}_\kappa \mid Y_n) - \log P_\theta(Y \in \mathcal{E}_\kappa).$$

Assume that for all $\varepsilon > 0$ and for all $M < \infty$,

$$P\left(\sup_{\|\theta - \theta^*\| \geq \varepsilon} (\mathcal{L}_n(\theta) - \mathcal{L}_n(\theta^*)) \geq -M\right) \xrightarrow{n \to \infty} 0.$$ 

Then,

$$P\left(\sup_{\|\theta - \theta^*\| \geq \varepsilon} (\mathcal{L}_{\kappa,n}(\theta) - \mathcal{L}_{\kappa,n}(\theta^*)) \geq -M \mid Y \in \mathcal{E}_\kappa\right) \xrightarrow{n \to \infty} 0.$$ 

Consequently

$$\arg\max_{\theta \in \Theta} \mathcal{L}_n(\theta) \xrightarrow{P \text{ } n \to \infty} \theta^*, \quad \text{and} \quad \arg\max_{\theta \in \Theta} \mathcal{L}_{\kappa,n}(\theta) \xrightarrow{P \mid Y \in \mathcal{E}_\kappa \text{ } n \to \infty} \theta^*,$$

where $\xrightarrow{P \text{ } n \to \infty}$ denotes the convergence in probability under the distribution of $Y$, and $\xrightarrow{P \mid Y \in \mathcal{E}_\kappa \text{ } n \to \infty}$ denotes the convergence in probability under the distribution of $Y$ given $Y \in \mathcal{E}_\kappa$.

Proof. For any fixed $\delta > 0$, since $\log(P_\theta(Y \in \mathcal{E}_\kappa \mid Y_n)) \leq 0$ for all $\theta \in \Theta$, the quantity $\mathcal{P} = P\{\sup_{\|\theta - \theta^*\| \geq \varepsilon} \log(P_\theta(Y \in \mathcal{E}_\kappa \mid Y_n)) \leq \log(P_\theta^*(Y \in \mathcal{E}_\kappa \mid Y_n)) \geq \delta \mid Y \in \mathcal{E}_\kappa\}$ satisfies

$$\mathcal{P} \leq P\left\{ -\log(P_\theta^*(Y \in \mathcal{E}_\kappa \mid Y_n)) \geq \delta \mid Y \in \mathcal{E}_\kappa\right\}$$

$$= P\left\{ P_\theta^*(Y \in \mathcal{E}_\kappa \mid Y_n) \leq \exp(-\delta) \mid Y \in \mathcal{E}_\kappa\right\} \xrightarrow{n \to \infty} 0.$$
from Lemmas A.1 and A.2. Also, from Lemma A.6, there exists \( \Delta > 0 \) so that we have

\[
\inf_{\|\theta - \theta^*\| \geq \varepsilon} P_{\theta}(Y \in \mathcal{E}_\kappa) \geq \Delta > 0,
\]

so that

\[
\sup_{\|\theta - \theta^*\| \geq \varepsilon} - \log(P_{\theta}(Y \in \mathcal{E}_\kappa)) + \log(P_{\theta^*}(Y \in \mathcal{E}_n)) \leq - \log(\Delta) < \infty.
\]

Hence, the proposition follows.

Remark 5.1. Propositions 5.1 and 5.2 can be extended to the case of noisy observations of GPs (or to the case of a nugget effect). More precisely, these propositions would still hold if \( Y_n \) was replaced by \( Y_n + z_n \), where \( z_n = (z_1, \ldots, z_n)^T \), where \( z_1, \ldots, z_n \) are independent, independent of \( Y \), and follow the \( N(0, \tau) \) distribution with \( \tau > 0 \) fixed, known, and not depending on \( n \). Naturally, \( Y_n \) should also be \( Y_n + z_n \) in the definition of the ML and constrained maximum likelihood functions. The proofs of these adaptations of Propositions 5.1 and 5.2 would still hold if \( Y_n \) is replaced by \( Y_n + z_n \). Similarly, we believe that Propositions 5.1 and 5.2 can also be adapted if \( \tau \) is estimated by ML or conditional ML.

Remark 5.2. We refer to Appendix B for additional results that account for the hat basis functions in (1).


6.1. 2D Case. The finite-dimensional Gaussian representation of section 3 can be extended to \( d \)-dimensional input spaces by tensorization. For readability, we focus on the case \( d = 2 \) with \( D = [0,1]^2 \) and \( m_1 \times m_2 \) knots located on a regular grid. Then, the finite approximation is given by

\[
Y_{m_1,m_2}(x_1, x_2) := \sum_{j_1=1}^{m_1} \sum_{j_2=1}^{m_2} \xi_{j_1,j_2}\phi_{j_1}(x_1)\phi_{j_2}(x_2), \quad \text{s.t.} \quad \begin{cases} Y_{m_1,m_2}(x_1, x_2) = y_i, & (i = 1, \ldots, n), \\ \xi_{j_1,j_2} \in \mathcal{C}, \end{cases}
\]

where \( \xi_{j_1,j_2} = Y(t_{j_1}, t_{j_2}) \), \( \phi_{j_1}(x_1) \) and \( \phi_{j_2}(x_2) \) are hat basis functions given by (2), and \( (x_1, x_2), \ldots, (x_1, x_2) \) constitute a DoE. If we follow a similar procedure as in section 3, we observe that \( \xi = [\xi_{1,1}, \ldots, \xi_{1,m_2}, \ldots, \xi_{m_1,1}, \ldots, \xi_{m_1,m_2}]^T \) is a zero-mean Gaussian vector with covariance matrix \( \Gamma \) as in (8). Notice that each row of the matrix \( \Phi \) is given by

\[
\Phi_{i,:} = \left[ \phi_{1}(x_1)^2 \phi_{1}(x_2)^2 \cdots \phi_{m_1}(x_1)^2 \phi_{m_2}(x_2)^2 \cdots \phi_{m_1}(x_1)^2 \phi_{m_2}(x_2)^2 \right]
\]

for \( i = 1, \ldots, n \). Finally, the posterior distribution of (10) can be computed, and the routine follows Algorithm 1. The tensor form of (16) was also used in [6] for monotonicity constraints in multidimensional cases.

Figure 6 shows two examples where boundedness or monotonicity is studied. We used a 2D SE covariance function and we estimated the covariance parameters using cMLE.\(^5\)

\(^5\)2D SE covariance function: \( k_\theta(x - x') = \sigma^2 \exp\left(-\frac{(x_1 - x'_1)^2}{2\alpha_1^2} - \frac{(x_2 - x'_2)^2}{2\alpha_2^2}\right) \) with \( \theta = (\sigma^2, \alpha_1, \alpha_2) \).
Figure 6. Examples of 2D Gaussian models with boundedness or monotonicity constraints for interpolating the toy examples from subsection 6.1. Boundedness and monotonicity results are shown in the first and second row, respectively. Each row shows training points (black dots), the conditional mean function (on the left), and some conditional realizations (the last two columns on the right).

used multistart for the constrained ML optimization with ten initial vectors of parameters \( \theta = (\sigma^2, \alpha_1, \alpha_2) \) with \( \sigma^2 \in [0.2, 2] \) and \( \alpha_1, \alpha_2 \in [0.1, 0.9] \) in both examples. The training points were generated with a maximin Latin hypercube DoE over \([0, 1]^2\). The functions are 

\[
6(a) \quad y(x_1, x_2) = -\frac{1}{2}[\sin(9x_1) - \cos(9x_2)], \quad \text{and} \quad 6(d) \quad y(x_1, x_2) = \arctan(5x_1) + \arctan(x_2).
\]

We note that for the case of monotonicity in two dimensions, the constraints to be satisfied are given by \( \xi_{i+1,j} \geq \xi_{i,j} \) and \( \xi_{i,j+1} \geq \xi_{i,j} \), for \( i = 1, \ldots, m_1 - 1 \) and \( j = 1, \ldots, m_2 - 1 \). This means that the function is nondecreasing with respect to its two input variables.

6.2. 2D application: Nuclear criticality safety. For assessing the stability of neutron production in nuclear reactors, safety criteria based on the effective neutron multiplication factor \( k_{\text{eff}} \) are commonly used [40, 41]. This factor is defined as the ratio of the total number of neutrons produced by a fission chain reaction to the total number of neutrons lost by absorption and leakage. Besides the geometry and composition of fissile materials (e.g., mass, density), \( k_{\text{eff}} \) is sensitive to other types of parameters like the structure materials
characteristics (e.g., concrete) and the presence of specific materials (e.g., moderators). Since the optimal control of an individual parameter or a combination of them can lead to safe conditions, the understanding of their influence in criticality safety assessment is crucial.

In this section, we applied the proposed framework to a dataset provided by the Institut de Radioprotection et de Sûreté Nucléaire (IRSN), France. The $k_{\text{eff}}$ factor was obtained from a nuclear reactor called the “Lady Godiva device” originally situated at the Los Alamos National Laboratory, New Mexico, where uranium materials were managed. Two parameters of the uranium spheres are considered: the radius $r$ and density $d$. The dataset contains 121 observations in a $11 \times 11$ grid (see Figure 7). Notice that, on the domain considered for the input variables, $k_{\text{eff}}$ increases as the radius and density of the uranium sphere increase.

We trained different Gaussian models whether the inequality constraints are considered or not. For all the models, we normalized the input space to be in $[0,1]^2$. We used the same 2D SE covariance functions as for the example from Figure 6. For the unconstrained model (UM), we used multistart for the ML optimization with six initial vectors of covariance parameters $\theta = (\sigma^2, \alpha_1, \alpha_2)$ with $\sigma^2 \in [0.2, 1]$ and $\alpha_1, \alpha_2 \in [0.1, 0.9]$. For the CMs, since the $k_{\text{eff}}$ factor indicates the production rate of neutron population, the output of the constrained processes has to be positive. Taking also into account nondecreasing behaviors, we also consider the monotonicity constraints. We estimated the covariance parameters by MLE or cMLE using the MMA optimiser from NLopt [30]. We trained both unconstrained and CMs with a fixed maximin Latin hypercube DoE at eight locations extracted from the unit grid. We used the remaining data to assess the quality of prediction tasks.

Figure 8 shows the performance of the proposed models using four or eight points from the proposed fixed DoE. For the UMs, we observe that the quality of the predictions depends strongly on both the amount of training data and their distribution in the input space. Notice from Figure 8(a) that if only a few training points are available, predictions are poor and they do not satisfy positive and nondecreasing behaviors. In Figure 8(d), we observe that if
there are enough training data that cover the input space, the UM behaves well and provides reliable predictions. On the other hand, we observe that the CMs produce accurate prediction results also when the training set is small.

Because the prediction accuracy depends on the training set, we repeated the procedure with twenty different random Latin hypercube DoEs using several values of \( n \). We used the \( Q^2 \) and \( CA_{\pm \sigma} \) criteria to evaluate the quality of the predictions (see subsection 5.2). Figure 9 shows that the CMs often outperform the unconstrained ones. Notice that although the \( Q^2 \) results obtained by the UM are comparable with the constrained ones when the number of training points is large enough, we observe, according to the \( CA_{\pm \sigma} \) criterion, that the CM using cMLE provides more reliable confidence intervals. This means that, if we consider both positivity and monotonicity conditions to take into account the physics of the \( k_{\text{eff}} \) factor, we can obtain more informative and robust models. Furthermore, we have to note that
Figure 9. Assessment of the Gaussian models for interpolating the dataset from Figure 7 using various numbers of training points \( n \) and using twenty different random Latin hypercube designs. Predictive accuracy is evaluated using the (left) \( Q^2 \) and (right) \( CA_{\pm\sigma} \) criteria. Results are shown for the UM using MLE (red) and the CM using either MLE (blue) or cMLE (green).

7. Conclusions. Continuing the approach proposed in [6], we have introduced a full Gaussian-based framework to satisfy linear sets of inequality constraints. The proposed finite-dimensional approach takes into account the inequalities for both data interpolation and covariance parameter estimation. Because the posterior distribution is expressed as a truncated multinormal distribution, we compared different MCMC methods as well as exact sampling methods. According to experiments, we concluded that the HMC-based sampler adapts to our needs. For parameter estimation, we suggested the constrained likelihood which takes into account the inequality constraints. We showed that, loosely speaking, any consistency result for ML with unconstrained GPs is preserved when adding boundedness, monotonicity, and convexity constraints. Furthermore, this consistency occurs for both the unconditional and conditional likelihood functions. We remark that, under the fixed domain asymptotic framework we consider, some covariance parameters (for instance, \( \sigma^2 \) and \( \alpha \) in subsection 5.2) cannot be estimated consistently and do not have an asymptotic impact on prediction and conditional distributions. These results appear in the literature for unconstrained GPs [32, 33]. We believe that it can be shown that, roughly speaking, these results also hold for constrained GPs. Hence, from an asymptotic point of view, covariance parameters that cannot be estimated consistently have an asymptotically negligible impact on prediction, also for constrained GPs.

We tested our model in both synthetic and real-world data in one or two dimensions. According to the experimental results under different types of inequalities, the proposed framework fits properly the observations and provides realistic confidence intervals. Our approach is also flexible enough to satisfy multiple inequality conditions and to deal with specific types of constraints which are sequentially activated. Finally, as we showed in the
2D nuclear criticality safety assessment, the proposed framework provides reliable predictions on both data prediction and uncertainty quantification satisfying the inequality constraints exhibited by the neutron population (positivity and monotonicity conditions). We also observe that the unconstrained MLE achieves a good tradeoff between prediction accuracy/reliability and computational cost.

The framework presented in this paper can be improved in different ways. First, the precision of the results depends on the number of knots \( m \) used in the finite approximation. For higher values of \( m \), the interpolation is better but more expensive. In this sense, we could consider the optimal location of the knots over the input space instead of using regular grids. This potentially allows us to reduce the computational cost of the full framework. Second, as we discussed for 2D input spaces, the model can be generalized to \( d \)-dimensional problems. However, due to its tensor structure, its practical application could be time-consuming. Hence, there is a need to find an extension of the model to higher dimensions that can be applied in real-world problems (e.g., a sparse representation). Third, the estimation of the Gaussian orthant probabilities can be improved in order to exploit the advantages of the constrained likelihood. The use of alternative MC/MCMC methods, such as path sampling approaches [42], could be further investigated in future work. Finally, Propositions B.1 and B.2 could be extended to larger values of \( d \) and for \( \kappa = 0, 1, 2 \) using additional assumptions on the triangular array of observation points.

Appendix A. Conditional maximum likelihood: Asymptotic properties. We detail in this appendix the conditions and lemmas we used in Propositions 5.1 and 5.2 from subsection 5.3. We use the same notation and assumptions as in subsection 5.3.

**Condition A.1.** Let \( x, x' \in \mathbb{X} \). For a fixed \( \kappa \in \{0, 1, 2\} \), assume one of the following conditions:

- If \( \kappa = 0 \). Assume that \( Y \) has continuous trajectories. Let \( k \) be the covariance function of \( Y \). Let 
  \[
  d_k(x, x') = \sqrt{k(x, x) + k(x', x') - 2k(x, x')}. 
  \]

Let \( N(\mathbb{X}, d_k, \rho) \) be the minimum number of balls with radius \( \rho \) (with respect to the distance \( d_k \)), required to cover \( \mathbb{X} \). Assume that

\[
\int_0^\infty \sqrt{\log(N(\mathbb{X}, d_k, \rho))} \, d\rho < \infty. \tag{17}
\]

Assume also that the Fourier transform \( \hat{k} \) of \( k \) satisfies

\[
\exists P < \infty \quad \text{so that as } \|w\| \to \infty, \quad \|\hat{k}(w)\|_P^P \to \infty. \tag{18}
\]

- If \( \kappa = 1 \). Assume that \( Y \) has \( C^1 \) trajectories. Let \( k_i^{[1]} \) be the covariance function of \( \frac{\partial}{\partial x_i} Y \). Let \( d_i^{[1]} \) and \( N(\mathbb{X}, d_i^{[1]}, \rho) \) be defined as \( d_k \) and \( N(\mathbb{X}, d_k, \rho) \) for \( \kappa = 0 \). Assume that

\[
\int_0^\infty \sqrt{\log(N(\mathbb{X}, d_i^{[1]}, \rho))} \, d\rho < \infty \quad \forall i = 1, \ldots, d. \tag{19}
\]

Assume also that the Fourier transform \( \hat{k_i^{[1]}} \) of \( k_i^{[1]} \) satisfies the same conditions as for \( \kappa = 0 \).
• If $\kappa = 2$. Assume that $Y$ has $C^2$ trajectories. Let $\phi_{k,ij}^{(1)}$ be the covariance function of $\frac{\partial^2}{\partial x_i \partial x_j} Y$. Let $d_{k,ij}^{(1)}$ and $N(X, d_{k,ij}^{(1)}, \rho)$ be defined as $d_k$ and $N(X, d_k, \rho)$ for $\kappa = 0$. Assume that

$$
\int_0^\infty \sqrt{\log(N(X, d_{k,ij}^{(1)}, \rho))} \, d\rho < \infty \quad \forall i, j = 1, \ldots, d.
$$

Assume also that the Fourier transform $\widehat{k}_{k,ij}^{(1)}$ of $k_{k,ij}^{(1)}$ satisfies the same conditions as for $\kappa = 0$.

Let us discuss Condition A.1. For $\kappa = 0$, it is assumed that $Y$ has continuous trajectories, which implies that the covariance function $k$ of $Y$ is continuous (see, e.g., [43, Theorem 1.6] or [44, Lemma 1]). Hence, Condition A.1 implies that $Y$ is mean square continuous [33]. Mean square continuity is perhaps a more commonly used notion than trajectory continuity in the statistical literature [1, 45]. Nevertheless, in the context of this paper, trajectory continuity is needed to define the event $E_\kappa$. Furthermore, (17) is here needed, in particular for the proof of Lemma A.3, where Dudley’s inequality (see, e.g., [46, Theorem 2.10]) is used. We also remark that (17), for $\kappa = 0$, is not significantly stronger than assuming that $k$ is continuous. In particular this condition holds if $k$ is $\alpha$-Hölder continuous with $\alpha > 0$ (since then one can show that in this case $N(X, d_k, \rho)$ is a $O(\rho^{-2d/\alpha})$ as $\rho \to 0$), which is the case for the Matérn covariance function with any smoothness parameter $\nu > 0$ [33]. Hence, Condition A.1 holds for the Matérn covariance function with $\nu > 0$, since then also (18) holds. (See, e.g., [33] for the expression of the Fourier transform of the Matérn covariance function.) We remark however that (18) does not hold for the squared exponential covariance function whose Fourier transform vanishes too fast as $w \to \infty$ [33].

The discussion is similar for $\kappa = 1, 2$. In these cases, Condition A.1 implies that $Y$ is $\kappa$ times mean square differentiable and that $k$ has partial derivatives of order $2\kappa$. Having derivatives of order $2\kappa$ is arguably a minimal condition for mean square differentiability of order $\kappa$ [1, 33, 45]. Furthermore, if the derivatives of order $2\kappa$ of $k$ are Hölder continuous, then (19) or (20) hold. Hence Condition A.1 is not significantly stronger than mean square differentiability and holds for the Matérn covariance function with $\nu > \kappa$.

**Lemma A.1.** Let $0 \leq \ell < u \leq \infty$. Let

$$
P_{n,\ell,u}(Y_n) = P_{\theta^*}(Y \in \mathcal{E}_0 \mid Y_n).
$$

Then, for all $\varepsilon \geq 0$, we have

$$
P(P_{n,\ell,u}(Y_n) \leq 1 - \varepsilon \mid Y \in \mathcal{E}_0) \xrightarrow{n \to \infty} 0.
$$

**Proof.** From Lemma A.3 we have $P(Y \in \mathcal{E}_0) > 0$. Hence, it is sufficient to show

$$
P(P_{n,\ell,u}(Y_n) \leq 1 - \varepsilon, Y \in \mathcal{E}_0) \xrightarrow{n \to \infty} 0.
$$

The term $P_{n,\ell,u}(Y_n)$, being a conditional expectation, is a martingale with respect to the $\sigma$-algebra generated by $Y(x_1), \ldots, Y(x_n)$. Furthermore, $0 \leq P_{n,\ell,u}(Y_n) \leq 1$. Hence

$$
P_{n,\ell,u}(Y_n) \xrightarrow{n \to \infty} P(Y \in \mathcal{E}_0 \mid \mathcal{F}_\infty).
$$
where $\mathcal{F}_\infty$ is the $\sigma$-algebra generated by $[Y(x_i)]_{i \in \mathbb{N}}$ using Theorem 6.2.3 from [38]. Let $\mu_n$ and $k_n$ be the mean and the covariance function (respectively) of $Y$ given $Y_n$. From proposition 2.8 in [39], the conditional distribution of $Y$ given $\mathcal{F}_\infty$ is the distribution of a GP with mean function $\mu_\infty$ and covariance function $k_\infty$. Furthermore, a.s., $\mu_n$ and $k_n$ converge uniformly to $\mu_\infty$ and $k_\infty$, respectively. Hence we can show simply that, because $(x_i)_{i \in \mathbb{N}}$ is dense in $\mathbb{X}$, we have a.s. $\mu_\infty = Y$ and $k_\infty$ is the zero function. Hence a.s. if $Y \in \mathcal{E}_0$ holds, then

$$P(Y \in \mathcal{E}_0 \mid \mathcal{F}_\infty) = 1,$$

so that $P_{n,\ell,u}(Y_n) \xrightarrow{n \to \infty} 1$.

Hence by the dominated convergence theorem

$$P(P_{n,\ell,u}(Y_n) \leq 1 - \varepsilon, Y \in \mathcal{E}_0) \xrightarrow{n \to \infty} 0.$$

**Lemma A.2.** Let $\kappa = \{1, 2\}$. Let

$$P_n(Y_n) = P_{\theta^n}(Y \in \mathcal{E}_n \mid Y_n).$$

Then, for all $\varepsilon > 0$, we have

$$P(P_n(Y_n) \leq 1 - \varepsilon \mid Y \in \mathcal{E}_n) \xrightarrow{n \to \infty} 0.$$

**Proof.** The proof is the same as that of Lemma A.1. In particular, we remark that $\mathbb{1}_{Y \in \mathcal{E}_n}$ is a measurable random variable, as $Y$ has $C^\infty$ trajectories.

**Lemma A.3.** Let $\kappa = 0$. Assume that Condition A.1 is satisfied. Then

$$P(Y \in \mathcal{E}_0) > 0 \quad \text{for} \quad -\infty \leq \ell < u \leq \infty.$$

**Proof.** We first prove that for any $\delta > 0$

$$P(\forall x \in \mathbb{X} : |Y(x)| \leq \delta) > 0.$$

This result is true and appears implicitly in the literature about small ball estimates for GP [47]. We nevertheless provide a proof of it for self-consistency. Let $(v_i)_{i \in \mathbb{N}}$ be a dense sequence in $\mathbb{X}$. Let $Y_v = [Y(v_1), \ldots, Y(v_n)]^\top$. Let $\mu_n$ and $k_n$ be the mean and the covariance function of $Y$ given $Y_n$. Then we let

$$d^2_{k_n}(x, x') = \text{var} \{(Y(x) - Y(x')) | \mathcal{F}_n\},$$

where $\mathcal{F}_n = \sigma(Y(v_1), \ldots, Y(v_n))$. Note that, for a Gaussian vector, the conditional variance is deterministic, i.e., $\text{var} \{(Y(x) - Y(x')) | \mathcal{F}_n\} = \mathbb{E} \{\text{var} \{(Y(x) - Y(x')) | \mathcal{F}_n\}\}$. Thus

$$d^2_{k_n}(x, x') = \mathbb{E} \{\text{var} \{(Y(x) - Y(x')) | \mathcal{F}_n\}\} \leq \text{var} \{(Y(x) - Y(x'))\} = d^2_{k}(x, x'),$$

from the law of total variance. Hence $N(\mathbb{X}, d_{k_n}, \rho) \leq N(\mathbb{X}, d_{k}, \rho)$ for all $\rho$. Also, from Theorem 2.10 in [46] (together with a union bound and using that $\max_{x \in \mathbb{X}} Y(x)$ and $\max_{x \in \mathbb{X}} [-Y(x)]$
have the same law) we have, with $C$ a universal constant,
\[
\mathbb{E} \left\{ \max_{x \in \mathcal{X}} |Y(x) - \mu_n(x)| \right\} \leq C \int_0^\infty \sqrt{\log(N(\mathcal{X}, d_{k_n}, \rho))} \, d\rho
\]
\[
= C \int_0^2 \sup_{x \in \mathcal{X}} k_n(x, x) \sqrt{\log(N(\mathcal{X}, d_{k_n}, \rho))} \, d\rho
\]
\[
\leq C \int_0^2 \sup_{x \in \mathcal{X}} k_n(x, x) \sqrt{\log(N(\mathcal{X}, d_k, \rho))} \, d\rho.
\]
This last integral goes to 0 as $n \to \infty$ because $\sup_{x \in \mathcal{X}} k_n(x, x) \to 0$ (see the proof of Lemma A.1), and because of Condition A.1. Hence $\max_{x \in \mathcal{X}} |Y(x) - \mu_n(x)|$ goes to 0 in probability. Furthermore, $\mathcal{P} = P(\text{for all } x \in \mathcal{X}, \ -\delta \leq Y(x) \leq \delta)$ satisfies
\[
\mathcal{P} \geq P \left( \forall x \in \mathcal{X}, \ -\frac{\delta}{2} \leq \mu_n(x) \leq \frac{\delta}{2}, \ -\frac{\delta}{2} \leq Y(x) - \mu_n(x) \leq \frac{\delta}{2} \right)
\]
\[
= P \left( \forall x \in \mathcal{X}, \ -\frac{\delta}{2} \leq \mu_n(x) \leq \frac{\delta}{2} \right) P \left( \forall x \in \mathcal{X}, \ -\frac{\delta}{2} \leq Y(x) - \mu_n(x) \leq \frac{\delta}{2} \right),
\]
since the distribution of $Y - \mu_n$ does not depend on $Y_v$. We now fix $n \in \mathbb{N}$ for which the second probability is nonzero. (The existence is guaranteed from above.) Then, the first probability is nonzero by continuity since, when $Y_v = 0$, then $\mu_n$ is the zero function. Hence we have
\[
P(\forall x \in \mathcal{X} : |Y(x)| \leq \delta) > 0.
\]
Let $f$ be a $C^\infty$ function on $\mathbb{R}^d$, square integrable, satisfying
\[
\forall x \in \mathcal{X}, \ \ell + \delta \leq f(x) \leq u - \delta,
\]
for $\delta > 0$. ($f$ exists for $\delta > 0$ small enough and can be taken, for instance, as $f(x) = \exp\{-\tau \|x - x_0\|^2\} \left[ \frac{u + \ell}{2} \right]$ with $\tau > 0$ small enough, and for any $x_0 \in \mathcal{X}$.) Let $Z$ be a GP with covariance function $k$ and mean function $f$. Then, from what we have shown before, we have
\[
P(\forall x \in \mathcal{X} : |Z(x) - f(x)| \leq \delta) > 0,
\]
so that
\[
P(\forall x \in \mathcal{X} : \ell \leq Z(x) \leq u) > 0.
\]
From [48, p. 138], as discussed by [33, p. 121], the Gaussian measures of $Y$ and $Z$ are equivalent. Thus
\[
P(Y \in \mathcal{E}_0) = P(\forall x \in \mathcal{X} : \ell \leq Y(x) \leq u) > 0.
\]

**Lemma A.4.** Let $\kappa = 1$. Assume that Condition A.1 is satisfied. Then
\[
P(Y \in \mathcal{E}_1) > 0.
\]
Proof. We first prove that for any $\delta > 0$

$$P \left( \forall i = 1, \ldots, d, \forall x \in \mathbb{X}: \left| \frac{\partial}{\partial x_i} Y(x) \right| \leq \delta \right) > 0.$$ 

We let $(v_i)_{i \in \mathbb{N}}$ and $Y_v$ be defined as in the proof of Lemma A.3. Then, as in this proof, we can show that for $i = 1, \ldots, d$

$$\max_{x \in \mathbb{X}} \left| \frac{\partial}{\partial x_i} Y(x) - \mathbb{E} \left\{ \frac{\partial}{\partial x_i} Y(x) \middle| Y_v \right\} \right| \xrightarrow{n \to \infty} 0.$$ 

Furthermore, $\mathcal{P} = P \ (\text{for all } i = 1, \ldots, d, \forall x \in \mathbb{X}: \left| \frac{\partial}{\partial x_i} Y(x) \right| \leq \delta)$ satisfies

$$\mathcal{P} \geq P \left( \forall i = 1, \ldots, d, \forall x \in \mathbb{X}, -\frac{\delta}{2} \leq \mathbb{E} \left\{ \frac{\partial}{\partial x_i} Y(x) \middle| Y_v \right\} \leq \frac{\delta}{2} \right) \forall i = 1, \ldots, d, \forall x \in \mathbb{X}, -\frac{\delta}{2} \leq \mathbb{E} \left\{ \frac{\partial}{\partial x_i} Y(x) \middle| Y_v \right\} \leq \frac{\delta}{2} \right) \times P \left( \forall i = 1, \ldots, d, \forall x \in \mathbb{X}, -\frac{\delta}{2} \leq \frac{\partial}{\partial x_i} Y(x) - \mathbb{E} \left\{ \frac{\partial}{\partial x_i} Y(x) \middle| Y_v \right\} \leq \frac{\delta}{2} \right).$$

Notice that the last equality holds because the distribution of the process $x \rightarrow \frac{\partial}{\partial x_i} Y(x) - \mathbb{E} \left\{ \frac{\partial}{\partial x_i} Y(x) \middle| Y_v \right\}$ does not depend on $Y_v$. We now fix $n \in \mathbb{N}$ so that the second probability is nonzero. (The existence is guaranteed from above.) Then, the first probability is nonzero by continuity since, when $Y_v = 0$, then for $i = 1, \ldots, d$, $\mathbb{E} \left\{ \frac{\partial}{\partial x_i} Y \middle| Y_v \right\}$ is the zero function. Hence, we have obtained

$$P \left( \forall i = 1, \ldots, d, \forall x \in \mathbb{X}: \left| \frac{\partial}{\partial x_i} Y(x) \right| \leq \delta \right).$$

We now conclude the proof in the same way as for Lemma A.3. We consider the mean function

$$f(x) = \left[ \sum_{i=1}^{d} x_i \right] \exp\{-\tau \|x - x_0\|^2\}$$

with $x_0 \in \mathbb{X}$ and $\tau > 0$. For $\tau$ small enough, $f$ is $C^\infty$ and square integrable and satisfies

$$\forall i = 1, \ldots, d, \forall x \in \mathbb{X}, \frac{\partial}{\partial x_i} f(x) \geq \frac{1}{2}.$$ 

Then, we conclude the proof as in the proof of Lemma A.3.

Lemma A.5. Let $\kappa = 2$. Assume that Condition A.1 is satisfied. Then,

$$P \left( Y \in \mathcal{E}_2 \right) > 0.$$
This is done in a similar way as for showing $P \left( \forall i, j = 1, \ldots, d, \forall \mathbf{x} \in \mathbb{X} : \left| \frac{\partial^2}{\partial x_i \partial x_j} Y(\mathbf{x}) \right| \leq \delta \right) > 0$. We then conclude similarly as the rest of the proof this Lemma. In particular, we consider the mean function

$$f(\mathbf{x}) = \left[ \sum_{i=1}^{d} x_i^2 \right] \exp\{-\tau \|\mathbf{x} - \mathbf{x}_0\|^2\}$$

with $\mathbf{x}_0 \in \mathbb{X}$ and $\tau > 0$. Let $\lambda_{\text{inf}}(M)$ be the smallest eigenvalue of a symmetric matrix $M$. Then, for $\tau$ small enough, $f$ is $C^\infty$ and square integrable and satisfies

$$\forall \mathbf{x} \in \mathbb{X}, \quad \lambda_{\text{inf}} \left( \frac{\partial^2}{\partial \mathbf{x}^2} f(\mathbf{x}) \right) \geq 1.$$ 

Lemma A.6. Let $\kappa \in \{0, 1, 2\}$. Assume that Condition A.1 holds. Let $Y_\mathbf{\theta}$ be the GP defined by

$$Y_\mathbf{\theta}(t) = \sigma Y(\alpha_1 t_1, \ldots, \alpha_d t_d).$$

Let $P_\mathbf{\theta}^\kappa = P(Y_\mathbf{\theta} \in \mathcal{E}_\kappa)$ (see (15)). Then,

$$\inf_{\theta \in \Theta} P_\mathbf{\theta}^\kappa > 0.$$ 

Proof. We do the proof for $\kappa = 2$. The proof for $\kappa = 0, 1$ is similar. Let $\varepsilon > 0$ and let

$$P_{\theta, \varepsilon}^\kappa = \mathbb{E} \left\{ \mathbb{1}_{I(Y_\mathbf{\theta}) \geq \varepsilon} + \frac{I(Y_\mathbf{\theta})}{\varepsilon} \mathbb{1}_{0 \leq I(Y_\mathbf{\theta}) \leq \varepsilon} \right\}$$

with $I(Y_\mathbf{\theta}) = \inf_{\mathbf{x} \in \mathbb{X}} \lambda_{\text{inf}} \left( \frac{\partial^2}{\partial \mathbf{x}^2} Y_\mathbf{\theta}(\mathbf{x}) \right)$. We have $P_{\theta, \varepsilon}^\kappa \leq P_{\mathbf{\theta}}^\kappa$ for any $\varepsilon > 0$. With the proof of Lemma A.5, we also obtain for $\varepsilon > 0$ small enough

$$\forall \theta \in \Theta \quad P_{\theta, \varepsilon}^\kappa > 0.$$ 

Hence, the proof is concluded, by compacity, if we show that $\theta \to P_{\theta, \varepsilon}^\kappa$ is a continuous function on $\Theta$. Let us show this. Let $\theta = (\sigma_1^2, \alpha_1, \ldots, \alpha_d) \in (0, \infty)^{d+1}$ and $\theta_n = (\sigma_n^2, \alpha_n, \ldots, \alpha_{nd}) \to \theta$. We have

$$\frac{\partial^2}{\partial x_i \partial x_j} Y_{\theta_n}(\mathbf{x}) = \sigma_n \left( (\alpha_n)_i (\alpha_n)_j \frac{\partial^2}{\partial x_i \partial x_j} Y(\alpha_n x_1, \ldots, \alpha_{nd} x_d) \right).$$

Hence, because $Y$ is $C^2$, we have a.s.

$$\sup_{\mathbf{x} \in \mathbb{X}} \left\| \frac{\partial^2}{\partial \mathbf{x}^2} Y_{\theta_n}(\mathbf{x}) - \frac{\partial^2}{\partial \mathbf{x}^2} Y_\theta(\mathbf{x}) \right\| \xrightarrow{n \to \infty} 0.$$
for any matrix norm $\|\cdot\|$. Hence also since $Y$ is $C^2$, we can show a.s.

$$\left( \inf_{x \in \mathbb{X}} \lambda_{\inf} \left( \frac{\partial^2}{\partial x^2} Y_{\theta_0}(x) \right) - \inf_{x \in \mathbb{X}} \lambda_{\inf} \left( \frac{\partial^2}{\partial x^2} Y_{\theta}(x) \right) \right) \xrightarrow{n \to \infty} 0.$$ 

Hence, we conclude by dominated convergence observing that $t \to (\mathbb{I}_{t \geq \varepsilon} + \frac{t}{\varepsilon} \mathbb{1}_{0 \leq t \leq \varepsilon})$ is a continuous function on $\mathbb{R}$. 

**Appendix B. Asymptotic properties with the finite-dimensional approximation.**  In this section, we give asymptotic results, similar to these of subsection 5.3, but where we additionally take into account the finite-dimensional approximation in (1). We shall assume that $Y$ is defined on $[0,1]$, and focus on boundedness and monotonicity constraints when studying conditional ML. We refer to Remark B.1 for a discussion of potential extensions to higher-dimensional input spaces and to convexity constraints for conditional ML.

We consider a dense triangular array of observation points $(x_i^{(n)})_{n \in \mathbb{N}, i=1,...,n}$ on $[0,1]$. For concision, let $(x_1^{(n)}, ..., x_n^{(n)}) = (x_1, ..., x_n)$. Without loss of generality, we assume that for $n \in \mathbb{N}$, $x_1 \leq \cdots \leq x_n$. Hence we assume that $x_1 \to 0$, $x_n \to 1$ and $\max_{i=2,...,n} x_i - x_{i-1} \to 0$ as $n \to \infty$.

We let $(m_n)_{n \in \mathbb{N}}$ be a sequence of numbers of hat basis functions for the finite-dimensional approximation in (1). We shall only assume that $m_n > n$ for all $n \in \mathbb{N}$. Then, we let $Y_m$, $\phi_j$, $\xi$, $\mathcal{C}_\kappa$ ($\kappa = 0,1,2$) be defined as in subsection 2.1 but with $m$ replaced by $m_n$. We also use the same notation $Y_m$, $\mathcal{L}_m$, $\widehat{\theta}_{MLE}$, $\mathcal{C}_m$, $\widehat{\theta}_{cMLE}$ as in subsection 5.1. Under this setting, we first extend Proposition 5.1 to the finite-dimensional approximation case.

**Proposition B.1.** Let $Y$ be a zero-mean GP on $[0,1]$ with covariance function $k$ satisfying Condition A.1. Let $\Theta$ be a compact set on $(0, \infty)^2$. Let $k_\theta$ be the covariance function of $x \to \sigma Y(\alpha x)$ for $\theta = (\sigma^2, \alpha) \in \Theta$. Let $\theta^* = (1,1)$. Note that $k = k_{\theta^*}$ and assume that $\theta^* \in \Theta$. Assume that for all $\varepsilon > 0$,

$$P(\|\hat{\theta}_{MLE} - \theta^*\| \geq \varepsilon) \xrightarrow{n \to \infty} 0.$$ 

Let $\kappa \in \{0,1,2\}$. Let $\mathcal{E}_\kappa$ be as in (15). Then, we have $P(Y \in \mathcal{E}_\kappa) > 0$ from Lemmas A.3 to A.5, and thus

$$P(\|\hat{\theta}_{MLE} - \theta^*\| \geq \varepsilon \mid Y \in \mathcal{E}_\kappa) \xrightarrow{n \to \infty} 0.$$ 

**Proof.** The proof is the same as that of Proposition 5.1.

We now extend Proposition 5.2 to the finite-dimensional approximation case.

**Proposition B.2.** We use the same notation and assumptions as in Proposition B.1. Let $\kappa \in \{0,1\}$ be fixed. Let $P_\theta$ be as in subsection 5.1. Assume that for all $\varepsilon > 0$ and for all $M < \infty$,

$$P\left( \sup_{\|\theta - \theta^*\| \geq \varepsilon} (\mathcal{L}_m(\theta) - \mathcal{L}_m(\theta^*)) \geq -M \right) \xrightarrow{n \to \infty} 0.$$ 

Then,

$$P\left( \sup_{\|\theta - \theta^*\| \geq \varepsilon} (\mathcal{L}_{c,m}(\theta) - \mathcal{L}_{c,m}(\theta^*)) \geq -M \mid Y \in \mathcal{E}_\kappa \right) \xrightarrow{n \to \infty} 0.$$
Consequently
\[ \hat{\theta}_{MLE} \xrightarrow{P_{n \to \infty}} \theta^*, \quad \text{and} \quad \hat{\theta}_{cMLE} \xrightarrow{P|Y\in\mathcal{E}_n} \theta^*, \]
where \( P_{n \to \infty} \) denotes the convergence in probability under the distribution of \( Y \), and \( P|Y\in\mathcal{E}_n \) denotes the convergence in probability under the distribution of \( Y \) given \( Y \in \mathcal{E}_n \).

**Proof.** Clearly, we have for \( \kappa = 0, 1 \) that \( Y \in \mathcal{E}_\kappa \) implies \( \xi \in \mathcal{C}_\kappa \). Hence, from the proof of Proposition 5.2 we have, for any \( \epsilon > 0 \),
\[ \inf_{\|\theta - \theta^*\| \geq \epsilon} P_\theta(Y \in \mathcal{E}_\kappa) \geq \Delta > 0, \quad \text{and so} \quad \inf_{\|\theta - \theta^*\| \geq \epsilon} P_\theta(\xi \in \mathcal{C}_\kappa) \geq \Delta > 0. \]
Hence, by using the same proof as for Proposition 5.2, in order to conclude the proof, it is sufficient to show that for \( \epsilon > 0 \),
\[ P\{P_{\theta^*}(\xi \in \mathcal{C}_\kappa \mid Y_m) \geq 1 - \epsilon \mid Y \in \mathcal{E}_\kappa\} \to_{n \to \infty} 1. \]
As \( P_{\theta^*}(Y \in \mathcal{E}_\kappa) > 0 \) it is sufficient to show
\[ P\{P_{\theta^*}(\xi \in \mathcal{C}_\kappa \mid Y_m) \leq 1 - \epsilon, Y \in \mathcal{E}_\kappa\} \to_{n \to \infty} 0, \]
that is,
\[ P\{P_{\theta^*}(\xi \notin \mathcal{C}_\kappa \mid Y_m) \geq \epsilon, Y \in \mathcal{E}_\kappa\} \to_{n \to \infty} 0. \]
We shall prove \((21)\) separately in the two cases \( \kappa = 0, 1 \).

**Case \( \kappa = 0 \) (boundedness).** From Tsirelson theorem in [46], we have
\[ R_{\sup} := \sup_{x,y \in \mathbb{R}} \frac{P_{\theta^*}\{\sup_{t \in [0,1]} Y(t) \leq y\} - P_{\theta^*}\{\sup_{t \in [0,1]} Y(t) \leq x\}}{y - x} < \infty. \]
Hence, since \( Y \) has zero mean function, we also obtain
\[ R_{\inf} := \sup_{x,y \in \mathbb{R}} \frac{P_{\theta^*}\{\inf_{t \in [0,1]} Y(t) \leq y\} - P_{\theta^*}\{\inf_{t \in [0,1]} Y(t) \leq x\}}{y - x} < \infty. \]
Let
\[ \mathcal{E}_{0,\delta} = \{ f : [0,1] \to \mathbb{R}, f \text{ is } C^0 \text{ and } \forall x \in [0,1], \ell + \delta \leq f(x) \leq u - \delta \}. \]
We observe that \( Y \notin \mathcal{E}_{0,\delta} \) and \( Y \notin \mathcal{E}_0 \) imply \( \sup_{t \in [0,1]} Y(t) \in [u - \delta, u] \) or \( \inf_{t \in [0,1]} Y(t) \in [\ell, \ell + \delta] \). Thus we have, from the union bound,
\[ P\{Y \notin \mathcal{E}_{0,\delta}, Y \in \mathcal{E}_0\} \leq P \left\{ \sup_{t \in [0,1]} Y(t) \in [u - \delta, u] \right\} + P \left\{ \inf_{t \in [0,1]} Y(t) \in [\ell, \ell + \delta] \right\} \leq R_{\sup} \delta + R_{\inf} \delta. \]
Hence, for \( \nu > 0 \) there exists \( \delta > 0 \) (where \( \delta \) depends on \( \nu \)) so that
\[
P\{ Y \notin \mathcal{E}_{0,\delta}, Y \in \mathcal{E}_0 \} \leq \nu.
\]
Hence, in order to prove (21) it is sufficient to show that for a fixed \( \delta > 0 \)
\[
P\{ P^\theta_r(\xi \notin \mathcal{C}_0 \mid Y_m) \geq \varepsilon, Y \in \mathcal{E}_{0,\delta} \} \to_{n \to \infty} 0.
\]
We now assume that the opposite display does not hold and aim at reaching a contradiction.

Up to taking a subsequence, we hence assume that there exists \( r > 0 \) so that for all \( n \in \mathbb{N} \)
\[
P\{ P^\theta_r(\xi \notin \mathcal{C}_0 \mid Y_m) \geq \varepsilon, Y \in \mathcal{E}_{0,\delta} \} \geq r.
\]
We have that \( Y \in \mathcal{E}_{0,\delta} \) implies \( Y_m \in \mathcal{D}_{0,\delta} \) with \( \mathcal{D}_{0,\delta} = [\ell + \delta, u - \delta]^n \). Hence
\[
P\{ P^\theta_r(\xi \notin \mathcal{C}_0 \mid Y_m) \geq \varepsilon, Y_m \in \mathcal{D}_{0,\delta} \} \geq r.
\]

We have
\[
P\{ \xi \notin \mathcal{C}_0, Y_m \in \mathcal{D}_{0,\delta} \} = \mathbb{E}_{\theta^r} \left\{ 1_{Y_m \in \mathcal{D}_{0,\delta}} P^\theta_r(\xi \notin \mathcal{C}_0 \mid Y_m) \right\} \geq r \varepsilon.
\]

Hence, in order to reach a contradiction, it is sufficient so show that
\[
(22) \quad P\{ \xi \notin \mathcal{C}_0, Y_m \in \mathcal{D}_{0,\delta} \} \to_{n \to \infty} 0.
\]
We first observe that, because of the density of \( x_1, \ldots, x_n \) as \( n \to \infty \), there exists a sequence \( (\Delta_n)_{n \in \mathbb{N}} \), so that \( \Delta_n \to 0 \) as \( n \to \infty \) and so that for \( n \in \mathbb{N} \) and \( x \in [0, 1] \), there exists \( q \in \{1, \ldots, n\} \) so that \( |x - x_q| \leq \Delta_n \), where \( q \) depends on the pair \( (n, x) \) under consideration.

We now assume that the event \( \xi \notin \mathcal{C}_0, Y_m \in \mathcal{D}_{0,\delta} \) holds. Then, there exists \( i \in \{1, \ldots, m_n\} \) so that \( \xi_i = Y(t_i) \notin [\ell, u] \). We assume that \( Y(t_i) \leq \ell \). (The other possibility is treated similarly as below and we omit the details.) Then, there exists \( j \in \{1, \ldots, n\} \) so that \( |t_i - t_j| \leq \Delta_n \). Then, \( \{Y_m\}_j \in [\ell + \delta, u - \delta] \). Furthermore, there exists \( k \in \{1, \ldots, m_n - 1\} \) so that \( \{Y_m\}_k \) is a convex combination of \( Y(t_k) \) and \( Y(t_{k+1}) \) with \( t_k \leq t_j \leq t_{k+1} \). (This follows from the definition of the hat basis functions.) Hence, we have \( Y(t_k) \geq \ell + \delta \) or \( Y(t_{k+1}) \geq \ell + \delta \).

Putting everything together, there exists \( v, w \in \{1, \ldots, m_n\} \) so that \( |t_v - t_w| \leq \Delta_n + 1/(m_n - 1) \) and \( |Y(t_v) - Y(t_w)| \geq \delta \). Hence we have
\[
(23) \quad \sup_{\substack{x, y \in [0, 1] \\ |x - y| \leq \Delta_n + 1/(m_n - 1)}} |Y(x) - Y(y)| \geq \delta.
\]

Since \( Y \) is almost surely continuous (thus absolutely continuous) on \([0, 1]\), the probability of the event in (23) goes to 0 as \( n \to \infty \), by dominated convergence. Hence (22) is proved.

Case \( \kappa = 1 \) (monotonicity). The proof is similar as for the case \( \kappa = 0 \). We let
\[
\mathcal{E}_{1,\delta} = \left\{ f : [0, 1] \to \mathbb{R}, f \text{ is } C^1 \text{ and } \forall x \in [0, 1], f'(x) \geq \delta \right\}.
\]
We first apply the Tsirelson theorem in [46] to the derivative GP \( Y' \). This gives us that for
\( \nu > 0 \) there exists \( \delta > 0 \) so that

\[
P\{Y \notin \mathcal{E}_{1,\delta}, Y \in \mathcal{E}_1\} \leq \nu.
\]

Hence, in order to prove (21) it is sufficient to show that for \( \delta > 0 \)

\[
P\{P_{\theta^*}(\xi \notin \mathcal{C}_1 \mid Y_m) \geq \varepsilon, Y \in \mathcal{E}_{1,\delta}\} \rightarrow_{n \rightarrow \infty} 0.
\]

We introduce

\[
\mathcal{D}_{1,\delta} = \left\{ v \in \mathbb{R}^n; \text{for } i = 1, \ldots, n - 1, \frac{v_{i+1} - v_i}{x_{i+1} - x_i} \geq \delta \right\}
\]

and we observe that \( Y \in \mathcal{E}_{1,\delta} \) implies \( Y_m \in \mathcal{D}_{1,\delta} \). Hence, proceeding as in the case \( \kappa = 0 \), it is sufficient to show, in order to conclude the proof for \( \kappa = 1 \), that

\[
(24) \quad P\{\xi \notin \mathcal{C}_1, Y_m \in \mathcal{D}_{1,\delta}\} \rightarrow_{n \rightarrow \infty} 0.
\]

We now assume that the event \( \xi \notin \mathcal{C}_1, Y_m \in \mathcal{D}_{1,\delta} \) holds. Then, there exists \( i \in \{1, \ldots, m_n - 1\} \) so that \( \xi_{i+1} = Y(t_{i+1}) \leq \xi_i = Y(t_i) \). By density, we can find two indices \( a, b \in \{1, \ldots, n\} \) so that \( x_a + 8/(m_n - 1) \leq x_b \leq t_i - 8/(m_n - 1) \) and \( t_{i+1} - x_a = \delta_n \rightarrow 0 \) as \( n \rightarrow \infty \). Let \( a, b \in \{1, \ldots, m_n - 1\} \) be so that

\[
t_a \leq x_a \leq t_{a+1} \leq t_b \leq x_b \leq t_{b+1}.
\]

We have

\[
(Y_m)_b - (Y_m)_a \geq \delta(x_b - x_a) \geq \delta(t_b - t_{a+1}).
\]

Since \( (Y_m)_a \) is a convex combination of \( Y(t_a) \) and \( Y(t_{a+1}) \) and since \( (Y_m)_b \) is a convex combination of \( Y(t_b) \) and \( Y(t_{b+1}) \), there exists \( s_a \in \{a, a + 1\} \) and \( s_b \in \{b, b + 1\} \) so that

\[
Y(t_{s_b}) - Y(t_{s_a}) \geq \delta(t_b - t_{a+1}) \geq \delta(s_b - s_a - 2/(m_n - 1)) \geq \frac{\delta}{2} (s_b - s_a),
\]

where the last above inequality holds because \( s_b - s_a \geq 6/m_n \). Putting everything together, there exists \( v, w \in \{1, \ldots, m_n\} \) so that \(|t_v - t_w| \leq \delta_n + 1/(m_n - 1)|\) and \(|Y'(t_v) - Y'(t_w)| \geq \delta/2\).

Hence we have

\[
(25) \quad \sup_{x,y \in [0,1], |x-y| \leq \delta_n + 1/(m_n - 1)} |Y'(x) - Y'(y)| \geq \delta/2.
\]

Since \( Y' \) is almost surely continuous (thus absolutely continuous) on \([0,1]\), the probability of the event in (25) goes to 0 as \( n \rightarrow \infty \). Hence (24) is proved.

Remark B.1. We have assumed \( d = 1 \) in Propositions B.1 and B.2 for concision and simplicity of notation. Proposition B.1 can of course be extended to larger values of \( d \), similarly as with Proposition 5.1. It is likely that Proposition B.2 could be extended to larger values of
by using similar proof techniques as for $d = 1$. For the case of boundedness constraints, the extension would in fact be relatively straightforward. For the cases of monotonicity constraints, it is possible that additional assumptions on the triangular array of observation points (beyond density) would be needed.

In the case $d = 1$, we believe that Proposition B.2 can be shown also for $\kappa = 2$ (convexity), using similar proof techniques as for $\kappa = 0, 1$, but at the cost of more technicality.

Acknowledgment. This research was conducted within the frame of the Chair in Applied Mathematics OQUAIDO, gathering partners in technological research (BRGM, CEA, IFPEN, IRSN, Safran, Storengy) and academia (CNRS, Ecole Centrale de Lyon, Mines Saint-Etienne, University of Grenoble, University of Nice, University of Toulouse) around advanced methods for computer experiments. We thank Yann Richet (IRSN) for providing the nuclear criticality safety data. We also thank reviewers and editors-in-chief for their helpful recommendations to improve the quality of the paper.

REFERENCES


