Gaussian process regression models under linear inequality conditions

A. F. López-Lopera[1], F. Bachoc[2], N. Durrande[1,3], and O. Roustant[1].
[1]Mines Saint-Étienne, France. [2]Institut de Mathmatiques de Toulouse, France. [3]PROWLER.io, UK.
Second workshop on Gaussian Processes at Saint-Étienne. 9 – 11 October 2018, Saint-Étienne, France

Introduction

- Gaussian processes (GPs) have become one of the most attractive Bayesian frameworks in different decision tasks [1].
- It is shown that considering inequality constraints in GPs (e.g. positiveness, monotonicity) can lead to more accurate regression models [2].
- We build on the framework proposed in [2] and our contributions are threefold:
  1. We extend their framework for general sets of linear inequality constraints.
  2. We suggest an efficient MCMC sampler to approximate the posterior.
  3. We investigate theoretical/numerical properties of a constrained likelihood.

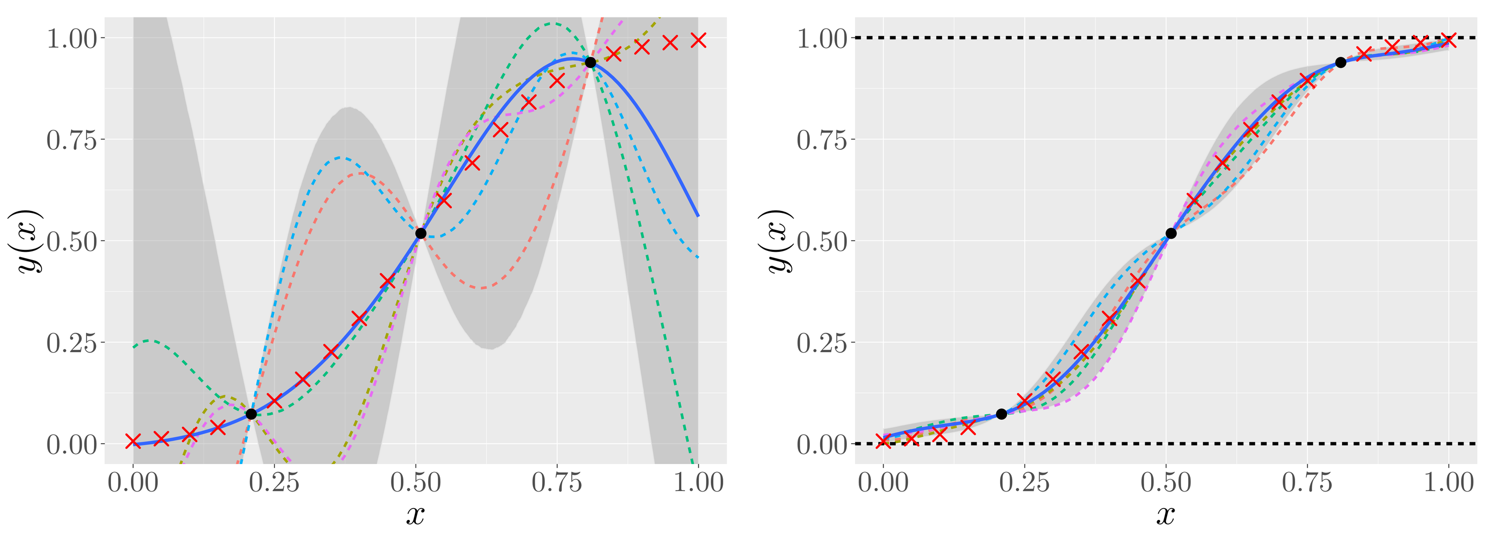
Materials and Methods

Gaussian process (GP) regression models

A GP is a collection of random variables, any finite number of which have a joint Gaussian distribution [1]. Let Y be a GP. Then, Y is completely defined by its mean function m and covariance function k

$$Y(x) \sim \mathcal{GP}(m(x), k(x, x')), \quad (1)$$

where $m(x) = \mathbb{E}\{Y(x)\}$ and $k(x,x') = \mathbb{E}\{[Y(x) - m(x)][Y(x') - m(x')]\}$.



(a) Unconstrained GP (b) Constrained GP

Figure 1: Examples GP regression models.

GP regression models under linear inequality conditions [3]

1) Define the finite-dimensional GP Y_m as the piecewise linear interpolation of Y at knots $t_1, \cdots, t_m$ (equally-spaced)

$$Y_m(x) = \sum_{j=1}^{m} Y(t_j)\phi_j(x), \ \text{s.t.} \begin{cases} Y_m(x_i) = y_i & \text{(interpolation conditions)}, \\ Y_m \in \mathcal{E} & \text{(inequality conditions)}, \end{cases} \quad (2)$$

where $x_i \in [0,1]$, $y_i \in \mathbb{R}$ for $i = 1, \cdots, n$, $\boldsymbol{\xi} = [Y(t_1), \cdots, Y(t_m)] \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma})$ with covariance matrix $\boldsymbol{\Gamma}$, and $\phi_1 \cdots, \phi_m$ are hat basis functions.

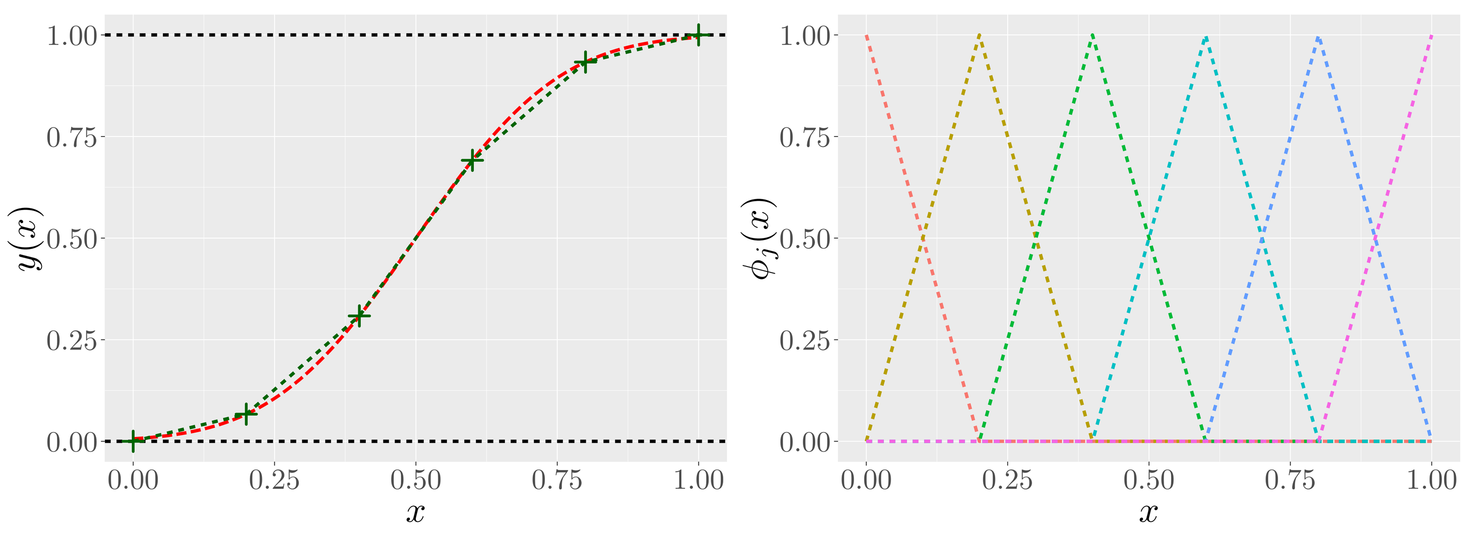


Figure 2: Finite-dimensional approximation of GP regression models.

Property: the function $Y_m(x) \in \mathcal{E} \leftrightarrow$ the vector $\boldsymbol{\xi} \in \mathcal{C}$ [2].

2) Since linearity preserves Gaussian distributions, quantifying uncertainty on Y_m relies on simulating a truncated Gaussian vector $\boldsymbol{\xi} \in \mathcal{C}$ (e.g. MC, MCMC).

Result 1. Performance of MC and MCMC samplers

Table 1: Samplers: Rejection Sampling from the Mode (RSM) [2], Exponential Tilting (ET), Gibbs Sampling (Gibbs), Metropolis-Hasting (MH), Hamiltonian Monte Carlo (HMC) [4]. **Indicators:** effective sample size: ESS $= n/(1 + 2\sum_{\forall k} \widehat{\rho}_k)$, time normalised (TN)-ESS.

Toy Example	Method	CPU Time [s]	ESS [$\times 10^4$] ($q_{10\%}, q_{50\%}, q_{90\%}$)	TN-ESS [$\times 10^4 s^{-1}$]	Hyperparameter
Figure 1(b)	RSM	-	-	-	-
	ET	41.16	$(0.99, 1.00, 1.00)$	0.02	-
	Gibbs	40.28	$(0.37, 0.6, 0.91)$	0.01	thinning = 1000
	MH	-	-	-	-
	HMC	**12.92**	$(0.85, 0.93, 1.00)$	**0.07**	-

Result 2. Constrained Maximum Likelihood (CML)

The conditional log-likelihood is written

$$\mathcal{L}_{\mathcal{C},m}(\boldsymbol{\theta}) = \log p_{\boldsymbol{\theta}}(\mathbf{Y}_m) + \log P_{\boldsymbol{\theta}}(\boldsymbol{\xi} \in \mathcal{C} | \boldsymbol{\Phi}\boldsymbol{\xi} = \mathbf{Y}_m) - \log P_{\boldsymbol{\theta}}(\boldsymbol{\xi} \in \mathcal{C}), \quad (3)$$

where the first term is the unconstrained log-likelihood.

Asymptotic property [3]: Let

$$\mathcal{L}_{\mathcal{C},n}(\boldsymbol{\theta}) = \mathcal{L}_n(\boldsymbol{\theta}) + \log P_{\boldsymbol{\theta}}(Y \in \mathcal{E}_\kappa | \mathbf{Y}_n) - \log P_{\boldsymbol{\theta}}(Y \in \mathcal{E}_\kappa),$$

where $\mathcal{E}_\kappa$ is the set of boundedness, monotonicity, and convexity constraints for $\kappa = 1, 2, 3$ (resp.). Assume that $\forall \varepsilon > 0$ and $\forall M < \infty$ (Consistency of the ML),

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

Then, (Consistency of the conditional CML)

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

Consequently, $\text{argmax}_{\boldsymbol{\theta} \in \Theta} \ \mathcal{L}_n(\boldsymbol{\theta}) \xrightarrow[n \to \infty]{P} \boldsymbol{\theta}^*$ and $\text{argmax}_{\boldsymbol{\theta} \in \Theta} \ \mathcal{L}_{\mathcal{C},n}(\boldsymbol{\theta}) \xrightarrow[n \to \infty]{P|Y \in \mathcal{E}_\kappa} \boldsymbol{\theta}^*$.

Result 3. 2D Nuclear Criticality Example

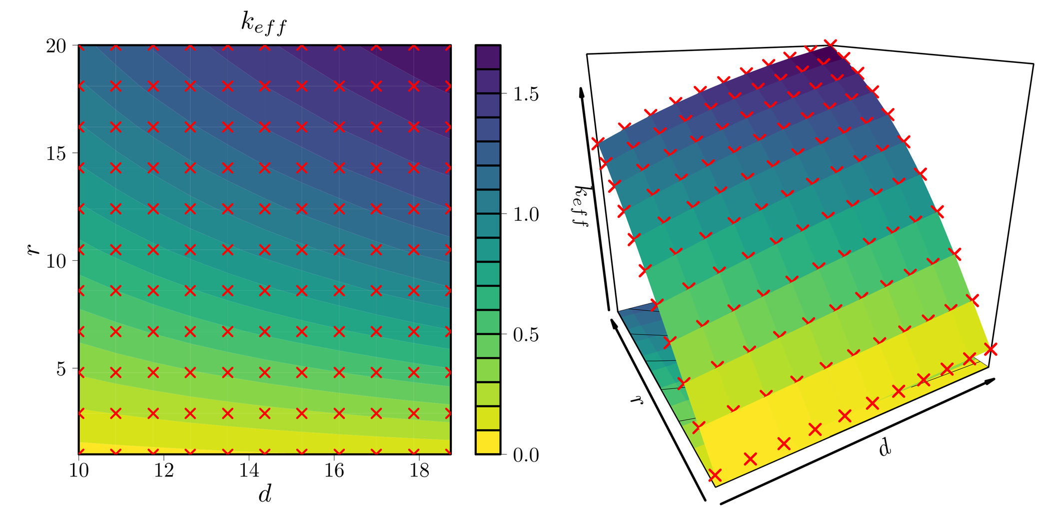
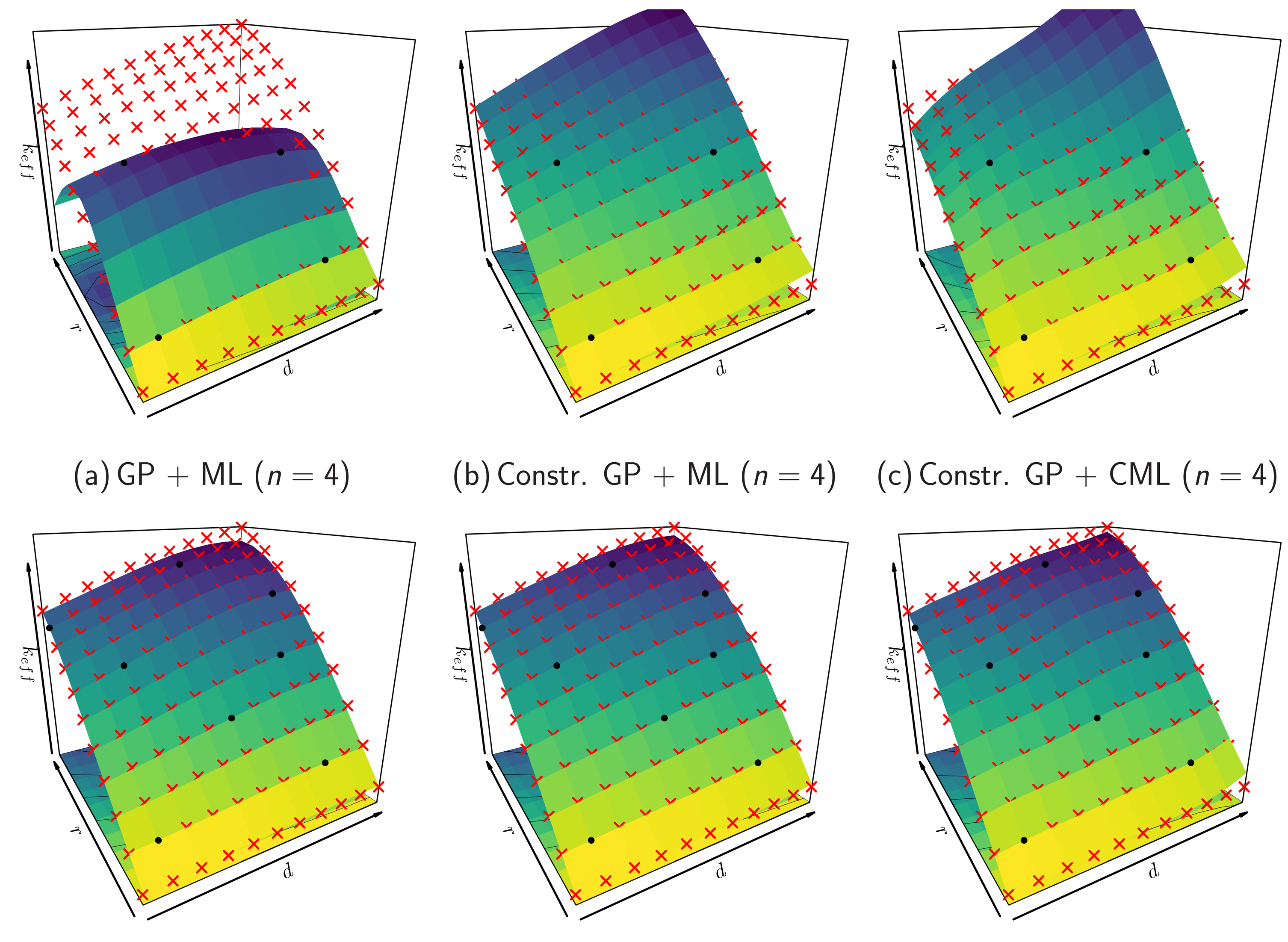


Figure 3: Nuclear criticality safety dataset. k_{eff} is positive and non-decreasing.



(a) GP + ML ($n = 4$) (b) Constr. GP + ML ($n = 4$) (c) Constr. GP + CML ($n = 4$)

(d) GP + ML ($n = 8$) (e) Constr. GP + ML ($n = 8$) (f) Constr. GP + CML ($n = 8$)

Figure 4: 2D GP regression models using different number of training points n. ML: Maximum Likelihood. CML: Constrained Maximum Likelihood.

Table 2: Performance of GPs for different n and using 20 random Latin hypercube designs. The accuracy is evaluated using the mean μ and the standard deviation σ of the Q^2 results.

n	GP + MLE $\mu \pm \sigma$	Constr. GP + MLE $\mu \pm \sigma$	Constr. GP + CMLE $\mu \pm \sigma$
2	-0.128 ± 1.004	$\mathbf{0.967 \pm 0.026}$	0.952 ± 0.043
4	0.558 ± 0.260	0.981 ± 0.014	$\mathbf{0.996 \pm 0.006}$
6	0.858 ± 0.139	0.940 ± 0.059	$\mathbf{0.995 \pm 0.004}$
8	0.962 ± 0.035	$\mathbf{0.995 \pm 0.003}$	0.981 ± 0.011

Conclusions

- We extended the framework proposed in [2] to deal with any set of linear inequality constraints.
- We suggested an efficient MCMC sampler based on HMC [4] to approximate the truncated posterior distribution.
- We further investigated theoretical/numerical properties of a constrained likelihood. The asymptotic properties are detailed in [3].

Future works

- To scale the proposed framework for higher dimensions and for a high number of observations.
- To study more theoretical properties of the constrained likelihood.

Acknowledgment

- This work was funded by the chair of applied mathematics OQUAIDO. We thank Yann Richet (IRSN) for providing the nuclear criticality safety data.

References

[1] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005.

[2] H. Maatouk and X. Bay, "Gaussian process emulators for computer experiments with inequality constraints," *Mathematical Geosciences*, vol. 49, no. 5, pp. 557–582, 2017.

[3] A. F. López-Lopera, F. Bachoc, N. Durrande, and O. Roustant, "Finite-dimensional Gaussian approximation with linear inequality constraints," *to appear in SIAM/ASA Journal on Uncertainty Quantification*, 2018.

[4] A. Pakman and L. Paninski, "Exact Hamiltonian Monte Carlo for truncated multivariate Gaussians," *Journal of Computational and Graphical Statistics*, vol. 23, no. 2, pp. 518–542, 2014.

[5] F. Bachoc, A. Lagnoux, and A. F. López-Lopera, "Maximum likelihood estimation for Gaussian processes under inequality constraints," *ArXiv e-prints*, Apr. 2018.

andres-felipe.lopez@emse.fr