LOD 2021

Numerical issues in maximum likelihood parameter estimation for Gaussian process interpolation

Subhasish Basak[1,2], Sébastien Petit[1,3], Julien Bect[1] & Emmanuel Vazquez[1]

October 7, 2021

1. Laboratoire des Signaux et Systèmes, CNRS, CentraleSupeléc, Univ. Paris-Saclay

2. ANSES, France

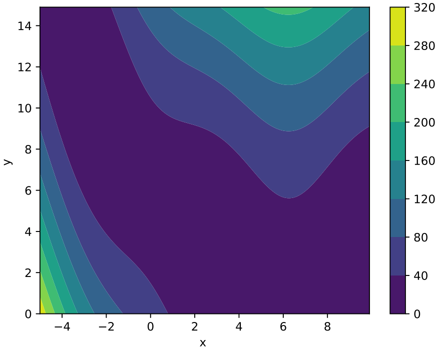
3. Safran Aircraft Engines, France

Motivation & Scope

- Gaussian processes (GP): popular tool for interpolation and regression in statistics and ML
  - **Geostatistics** (Stein, 1999)
  - **Design & analysis of computer experiments** (Santner et al., 2003)
  - **Machine Learning** (Rasmussen & Williams, 2006)
  - **Bayesian optimization** (Mockus, 1975; Jones, 1998; Emmerich et al., 2006; ...)
- Users rely on off-the-shelf GP implementations
- Problem: lack of consistency and robustness (see Erickson et al., 2018) among available software packages (Python, R, Matlab ...)

Example : the Branin function

$$f : \begin{cases} [-5,10] \times [0,15] & \rightarrow \quad \mathbb{R} \\ (x_1, x_2) & \mapsto \quad (x_2 - \frac{5.1}{4\pi^2} x_1^2 + \frac{5}{\pi} x_1 - 6)^2 + 10(1 - \frac{1}{8\pi}) \cos(x_1) + 10 \end{cases}$$



- Gaussian process model $\xi \sim \mathrm{GP}(0,\, k_\theta)$,
- with $k_\theta(x,y) = \sigma^2 \mathcal{M}_{5/2}\left(\sqrt{\left(\frac{x_1-y_1}{\rho_1}\right)^2 + \left(\frac{x_2-y_2}{\rho_2}\right)^2}\right)$, $\theta = (\sigma^2, \rho_1,\, \rho_2)$,
- σ^2 the process variance, ρ_1 and ρ_2 the lengthscales,
- and $\mathcal{M}$ the Matérn correlation function (with $\nu = 5/2$)
- Noisy observations with fixed noise σ_ϵ^2
- Estimate θ by optimizing the NLL

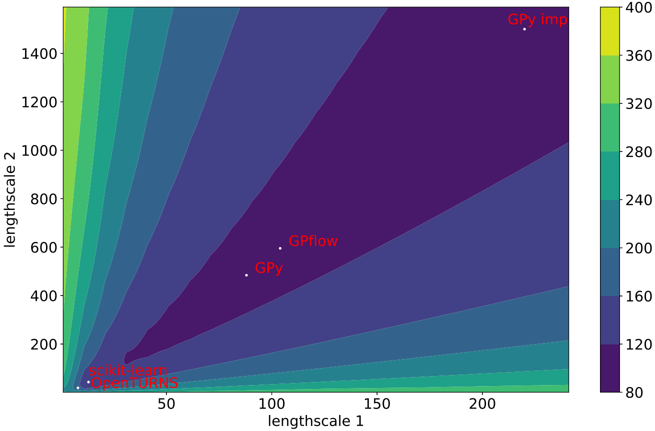
$$\begin{aligned} \mathcal{L}(\underline{Z}_n|\ \sigma^2, \rho_1, \rho_2) &= -2\log(p(\underline{Z}_n|\ \sigma^2, \rho_1, \rho_2)) && (1) \\ &= \log(\det(K_\theta)) + \underline{Z}_n^T K_\theta^{-1} \underline{Z}_n + C && (2) \end{aligned}$$

- K_θ is the covariance matrix associated to the design

Results

- size of training set $= 50$, size of testing set $= 500$

	scikit-learn	GPy	GPflow	GPyTorch	OpenTURNS	GPy "improved"
NLL	132.421	113.707	113.223	$2 \cdot 10^5$	163.125	**112.050**
RMSE	1.482	0.259	0.236	12.87	3.301	**0.175**

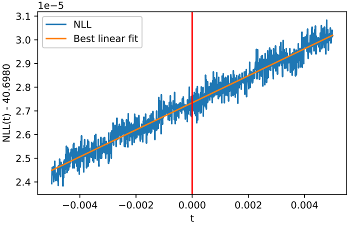


Plot of the NLL

- Efficient optimization of the NLL is critical for obtaining good GP interpolation.
- The objective of our article is two-fold:
  - investigate the origins of these inconsistencies
  - propose effective strategies for improvement

Contents

1 Numerical noise



- Our paper shows that numerical noise is linked to the condition number κ of the covariance matrix, here $\kappa = 10^{11}$ (double precision)
- Numerical experiments support the conclusion that jitter is not a satisfactory solution to tackle numerical noise

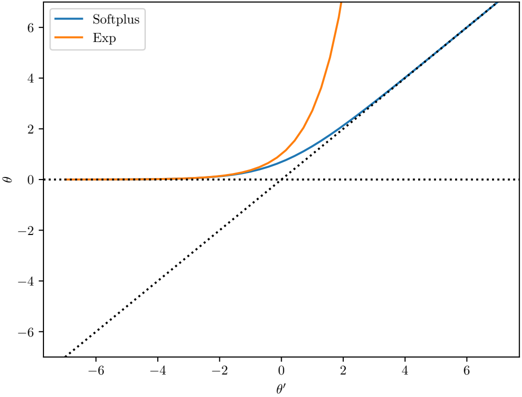
2 Example of lever for better NLL optimization: the parameterization

- Stationary covariance function k_θ, $\theta = (\sigma^2, \rho_1, \cdots, \rho_d) \in \mathbb{R}_+^{d+1}$
- Constant mean function $m(\cdot) = c \in \mathbb{R}$
- No numerical noise: $\sigma_\epsilon^2 = 0$
- In implementations, a monotonic one-to-one mapping $\Delta : \Theta' \to \Theta$ is used to optimize the NLL:

$$\theta'_{\text{opt}} = \underset{\theta' \in \Theta'}{\arg\min} - \log(\mathcal{L}(\underline{Z}_n | \Delta(\theta'), c)) \tag{3}$$

- Advantages:
  - Optimization on $\mathbb{R}^{d+1}$ instead of $\mathbb{R}_+^{d+1}$
  - Facilitates convergence

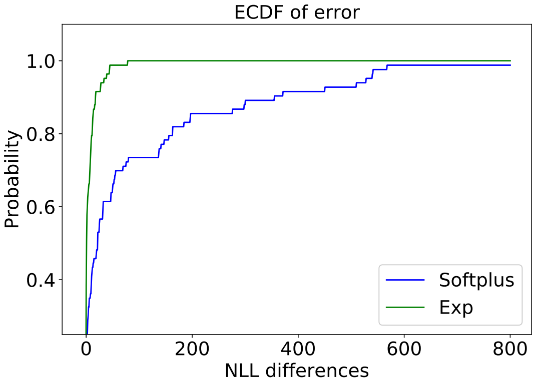
Two usual transformations



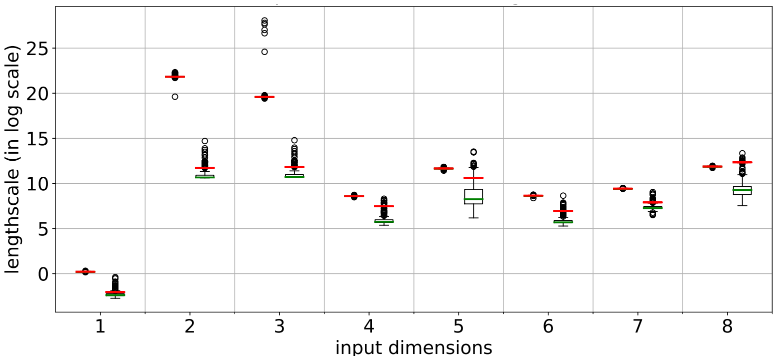
Softplus: $\theta' \mapsto \log(1 + \exp(\theta'))$ and Exp: $\theta' \mapsto \exp(\theta')$

3 Numerical study

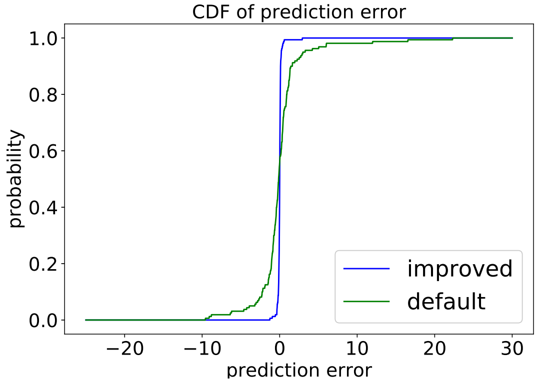
- Levers studied
  - Effect of parameterization
  - Effect of the initialization
  - Restarts/multi-starts
- Benchmark of 21 functions from 6 optimization problems
  - Data size $n \in \{3d, 5d, 10d, 20d\}$



Impact of the parameterization on GPy



LOO estimated lengthscales on the Borehole function $d = 8$, $n = 160$



ECDFs of LOO prediction errors

Default GPy implementation vs “improved” (Exp parameterization . . .)

4 Concluding remarks

- In GP interpolation, parameter estimation is difficult because of numerical noise
- Adaptive jitter cannot be considered as a do-it-all solution
- The ML estimation can be significantly improved using some simple and effective strategies
- This study intends to encourage practitioners to develop more robust GP implementations