An overview of 5 years of research on surrogate modelling in OQUAIDO

O. Roustant[a]

Slides contributors: M. Binois, Y. Deville, N. Durrande, D. Ginsbourger, R. Le Riche, A. Lopez Lopéra, E. Padonou

[a] INSA Toulouse

—

Surrogate Days, Toulouse, 2021 July 1

Outline

Outline

Surrogate models (or metamodels) – Computer experiments

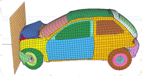
design

Surrogate models (or metamodels) – Computer experiments

design → **experiment + measurements** → **meas. output**

Surrogate models (or metamodels) – Computer experiments

design → **experiment + measurements** → **meas. output**



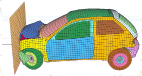
simulator inputs
$\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n$

→ **simulator**
physical model, numerical methods

→ **simulator output**
$f_{\text{sim}}(\mathbf{x}^1), f_{\text{sim}}(\mathbf{x}^2), \ldots, f_{\text{sim}}(\mathbf{x}^n)$

Surrogate models (or metamodels) – Computer experiments

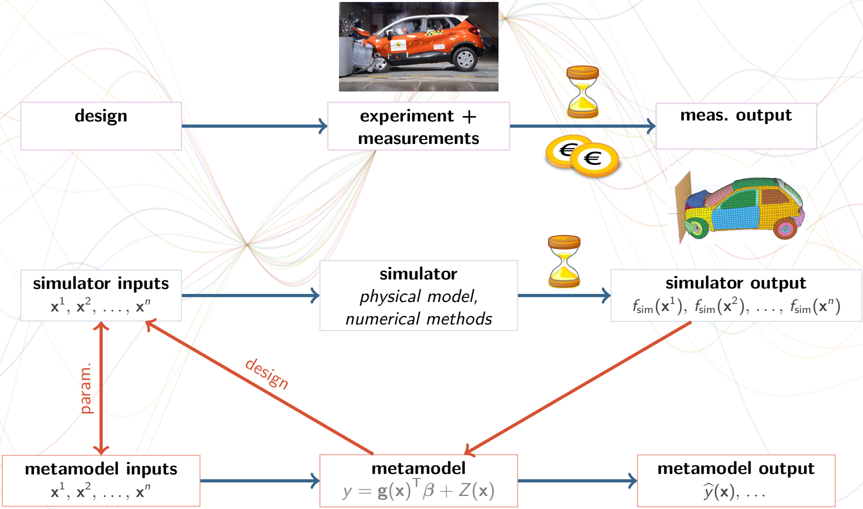
design → **experiment + measurements** → **meas. output**



simulator inputs $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n$ → **simulator** *physical model, numerical methods* → **simulator output** $f_{\text{sim}}(\mathbf{x}^1), f_{\text{sim}}(\mathbf{x}^2), \ldots, f_{\text{sim}}(\mathbf{x}^n)$

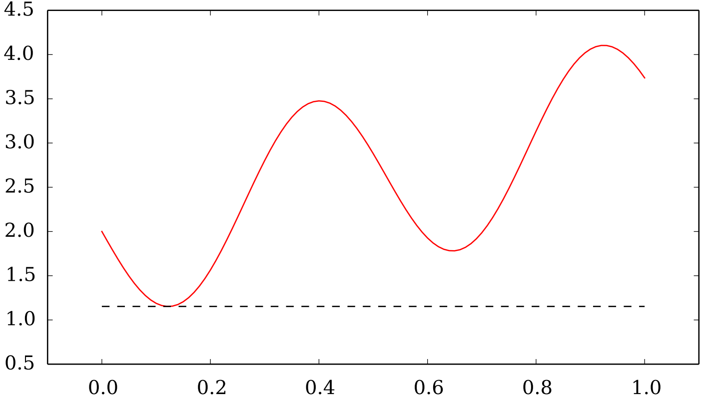
metamodel inputs $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n$ → **metamodel** $y = \mathbf{g}(\mathbf{x})^\top \beta + Z(\mathbf{x})$ → **metamodel output** $\widehat{y}(\mathbf{x}), \ldots$

Surrogate models (or metamodels) – Computer experiments



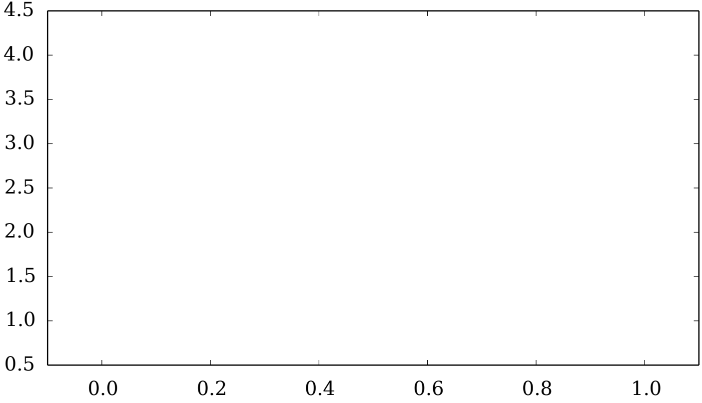
GP-based optimization

How to find the global minimum of a function... when each evaluation is costly ?



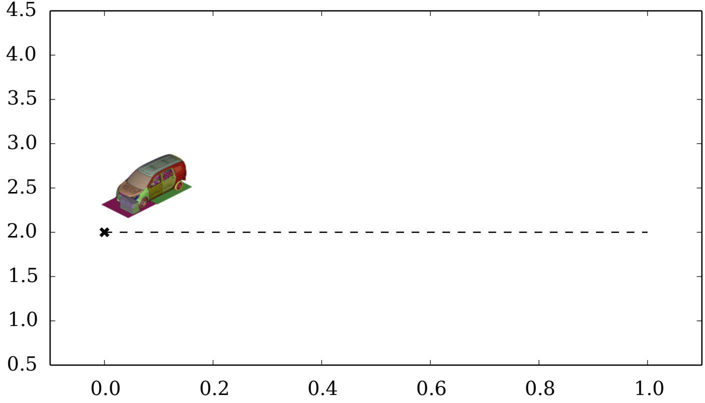
GP-based optimization

How to find the global minimum of a function... when each evaluation is costly ?



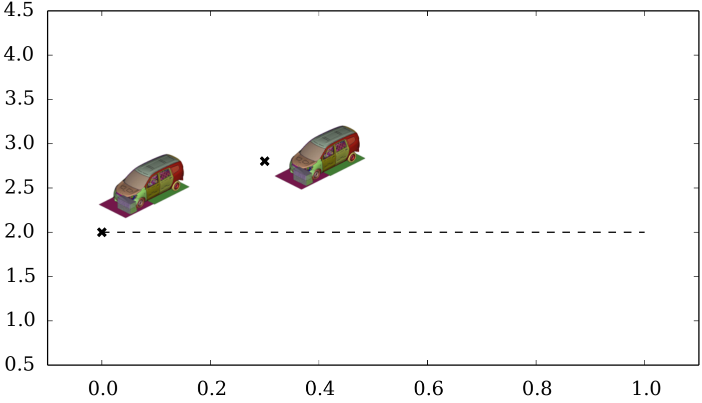
GP-based optimization

How to find the global minimum of a function... when each evaluation is costly ?



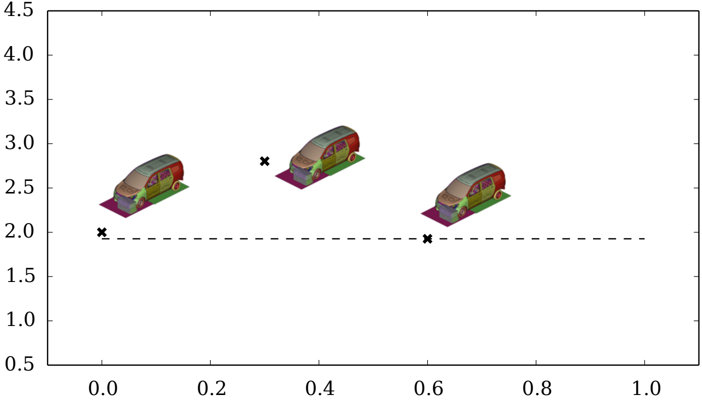
GP-based optimization

How to find the global minimum of a function... when each evaluation is costly ?



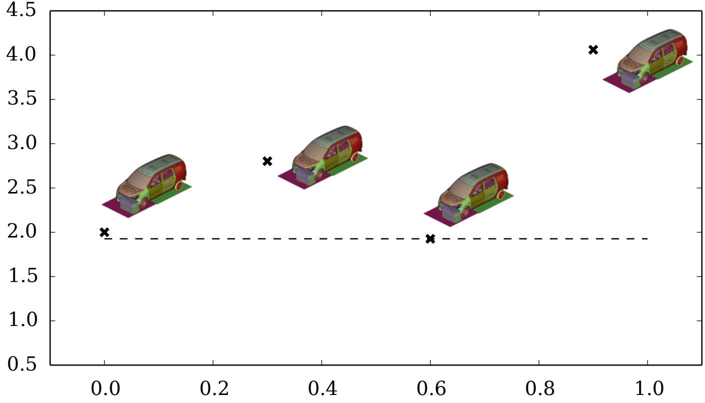
GP-based optimization

How to find the global minimum of a function... when each evaluation is costly ?



GP-based optimization

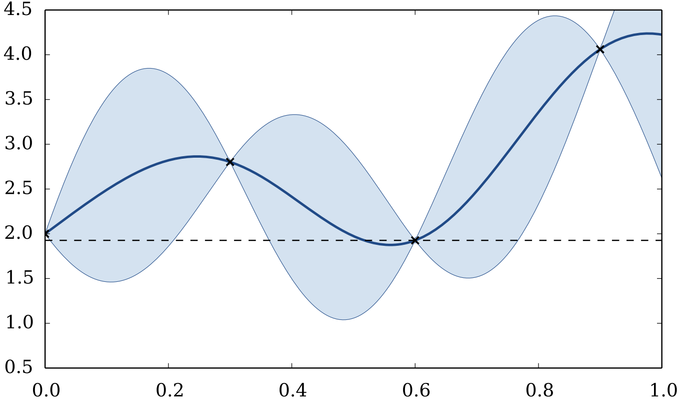
How to find the global minimum of a function... when each evaluation is costly ?



GP-based optimization

A solution : **GP-based (or "Bayesian") optimization** [Močkus, 1975, Jones et al., 1998]

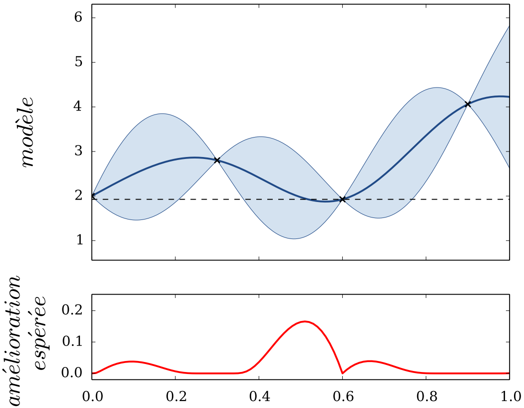
First ingredient : a GP model Y



GP-based optimization

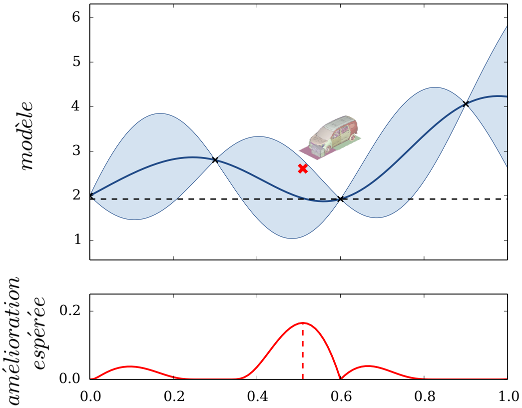
Second ingredient : an **easy-to-compute** criterion **accounting for uncertainty at unknown regions**, e.g. here "expected improvement"

$$EI(x) = \mathrm{E}([f_0 - Y(x)]^+ | Y(x_1), \ldots, Y(x_n)) \qquad f_0 : \textit{ current minimum}$$



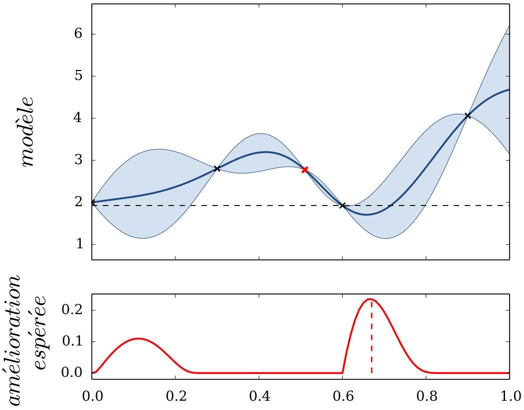
GP-based optimization

The algorithm (here "EGO") : (1) Find the next point by maximizing the criterion $\rightarrow$ (2) Evaluate the function $\rightarrow$ (3) Update the GP model $\uparrow$



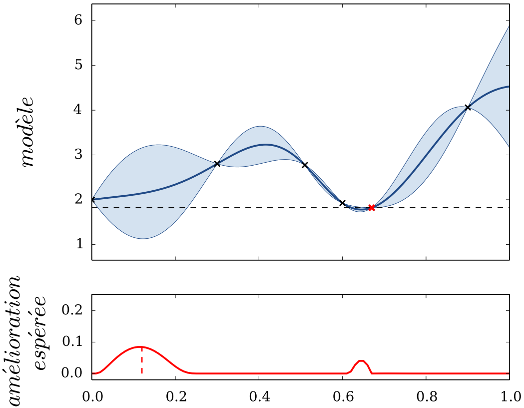
GP-based optimization

Iteration 2 :



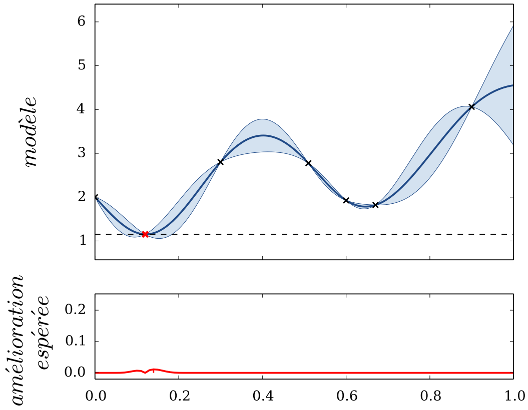
GP-based optimization

Iteration 3 :



GP-based optimization

Theory shows that **EGO algorithm provides a dense sequence of points**, up to a slight condition on the kernel used for GPs [Vazquez and Bect, 2010] .

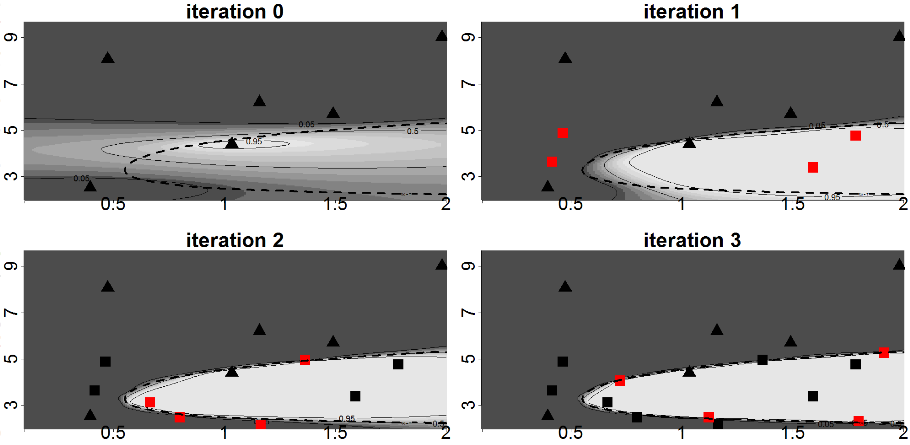


GP-based inversion

Same receipt for estimating a probability of failure ("SUR" strategy).

See [Chevalier et al., 2014] for details and [Bect et al., 2019] for a convergence analysis with supermartingales.

Illustration : Estimation of the nuclear criticity region $k_{\text{eff}} > 0.95$



Gaussian processes

Gaussian processes are stochastic processes (or random fields) s.t. every finite dimensional distribution is Gaussian. → Parameterized by two functions

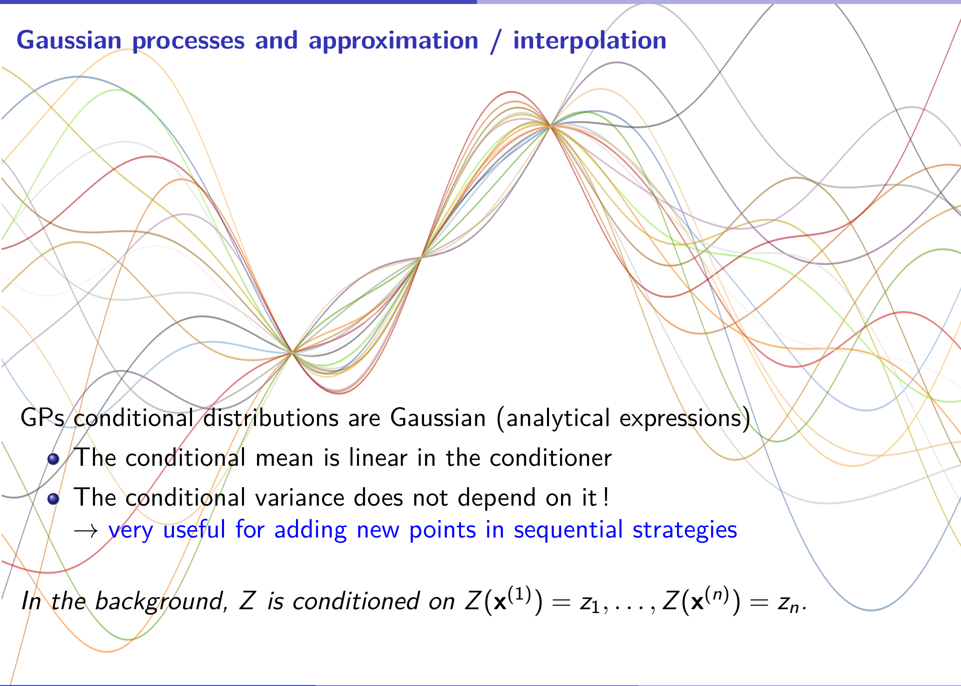
$$Z_{\mathbf{x}} \sim GP(\underbrace{m(\mathbf{x})}_{trend}, \underbrace{k(\mathbf{x}, \mathbf{x}')}_{kernel})$$

- The trend can be any function.
- The kernel is positive semidefinite :

$$\forall n, \alpha_1, \dots, \alpha_n, \mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}, \qquad \sum_{i=1}^{n} \alpha_i \alpha_j k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) \geq 0.$$

It contains the spatial dependence.

Gaussian processes and approximation / interpolation



GPs conditional distributions are Gaussian (analytical expressions)

- The conditional mean is linear in the conditioner
- The conditional variance does not depend on it !
  → very useful for adding new points in sequential strategies

In the background, Z is conditioned on $Z(\mathbf{x}^{(1)}) = z_1, \dots, Z(\mathbf{x}^{(n)}) = z_n$.

Gaussian processes, splines and RKHS

The 3 faces of a kernel

$GP(0, k(\mathbf{x}, \mathbf{x}')) \quad \Leftrightarrow \quad$ p.s.d. functions $k \quad \Leftrightarrow \quad$ RKHS : $\mathcal{H} = \overline{\mathrm{span}\{k(., \mathbf{x}), \mathbf{x} \in D\}}$

where $\mathcal{H}$ is a "Reproducing Kernel" Hilbert Space with dot product :

$$\langle k(\mathbf{x}, .), k(\mathbf{x}', .)\rangle = k(\mathbf{x}, \mathbf{x}') \qquad (*)$$

RKHS can be also defined as Hilbert spaces of functions such that evaluations $f \to f(\mathbf{x})$ are continuous : By Riesz theorem, there exists a unique $k(., \mathbf{x})$ s.t.

$$f(\mathbf{x}) = \langle f, k(., \mathbf{x})\rangle$$

Choosing $f = k(., \mathbf{x}')$ gives the reproducing identity (*).

Ref : [Aronszajn, 1950], [Berlinet and Thomas-Agnan, 2011].

Gaussian processes, splines and RKHS

Correspondence between interpolation spline and GP conditional mean

[Kimeldorf and Wahba, 1971]

The interpolation spline is defined by the functional problem

$$(*) \qquad \min_{h \in \mathcal{H}} \|h\| \qquad s.t. \quad h(\mathbf{x}^{(i)}) = z_i, \; i = 1, \dots, n$$

If $\mathcal{H}$ is the RKHS of kernel k, and if $K = \left(k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})\right)_{1 \leq i,j \leq n}$ is invertible, then (*) has a unique solution in the finite dimensional space spanned by the $k(., \mathbf{x}^{(i)})$:

$$h_{\mathrm{opt}}(\mathbf{x}) = \mathrm{E}\left[Z_{\mathbf{x}} \middle| Z(\mathbf{x}^{(i)}) = z_i, \; i = 1, \dots, n \right]$$

$\rightarrow$ In this sense, GPs are generalizing interpolation splines.

The first part (reduction to finite dimension) is know as Representer theorem.

Playing with kernels

A lot of flexibility can be obtained with kernels !

Building a kernel from other ones (basic examples)

Sum, tensor sum	$k_1 + k_2, \ k_1 \oplus k_2$
Product, tensor product	$k_1 \times k_2, \ k_1 \otimes k_2$
ANOVA	$(1 + k_1) \otimes (1 + k_2)$
Warping	$k(\mathbf{x}, \mathbf{x}') = k_1(f(\mathbf{x}), f(\mathbf{x}'))$
...	...

See examples in [Rasmussen and Williams, 2006]... and in this talk !

Why Gaussian processes are popular in (geo)statistics / machine learning ?

GP strengths

- Probabilistic models that **traduce spatial dependence**
  → provide realistic uncertainty in unvisited area
- **Conditional distributions are analytical**, and the cond. var. is constant
  → Useful for prediction and sequential strategies
- **Parameterized by functions** : mean and covariance **(kernel)**
  → Flexibility
- At the crossing between **rich mathematical theories**
  → Stochastic proc., Reproducing Kernel Hilbert Spaces, Positive Definite Functions

Outline

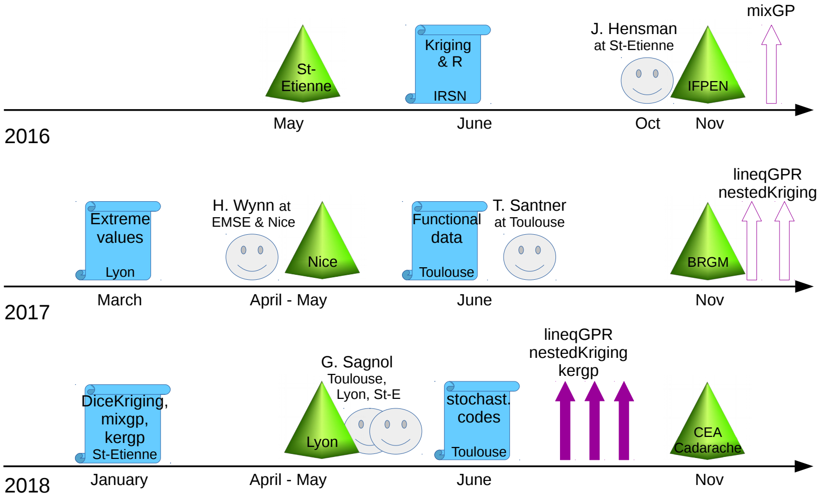
The OQUAIDO Chair

The research Chair OQUAIDO – for "Optimisation et QUAntification d'Incertitudes pour les Données Onéreuses" in French – has gathered academic and technological research partners to work on surrogate modeling and application to optimization and uncertainty quantification.

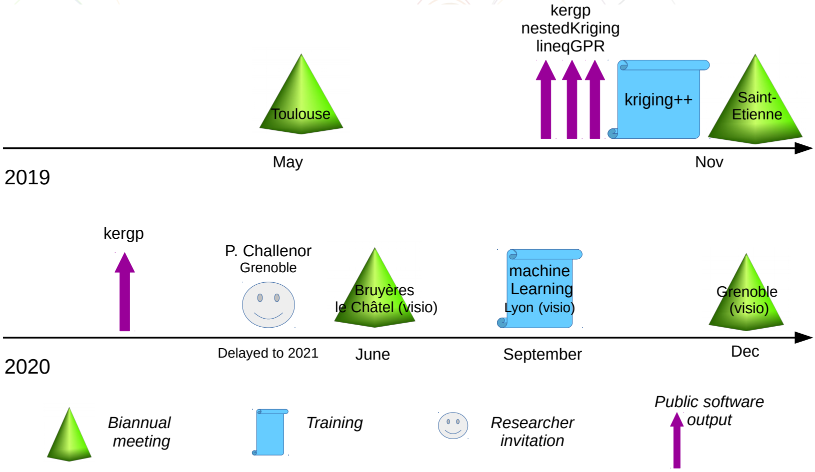
- Mines Saint-Étienne (that hosts the Chair), École Centrale de Lyon, CNRS, Univ. Grenoble Alpes, Univ. de Nice, Univ. de Toulouse III
- BRGM, CEA, IFPEN, IRSN, Safran, Storengy
- Y. Deville (AlpeStat), J. Garnier (École Polytechnique), D. Ginsbourger (Idiap), L. Polès (XtraFormation), as experts, and L. Carraro, as advisor.

Website : oquaido.emse.fr

OQUAIDO Chair Life



OQUAIDO Chair Life



Scientific program and production

\ **Application** **Framework** \	Optimization	Inversion / Calibration	Uncertainty Quantification	Modeling and other
Categorical inputs	PhD 4	Case 2 - Case 2' PhD 4		Post-doc 1
Stochastic codes	Case 4 - Post-doc 2		Case 4 - Post-doc 2	
Functional inputs/outputs High nb of inputs	Case 3 - PhD 1	Case 1 - Case 3 PhD 1, MSc	PhD 1	
Specific constraints			PhD 2	PhD 2
High nb of data				
Other topics				PhD 3

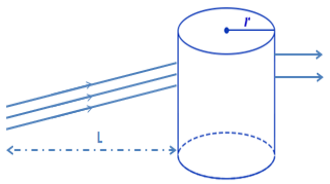
Table – Scientific program, case studies (Case) and extra manpower : post-doc, PhD thesis (PhD), master thesis (MSc).

Outline

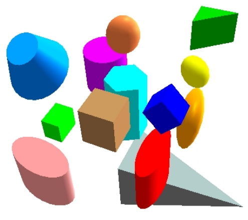
GP for categorical inputs : groups of levels & trees

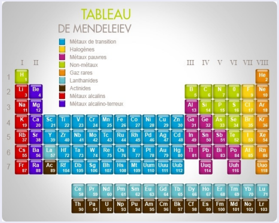
A guiding case study : Nuclear activity quantification

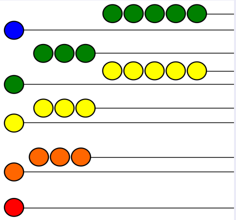
1. Computation using Monte Carlo
2. 4 continuous inputs : L, density, mean width, lateral surface
3. 3 categorical inputs : energy, form, chemical element.



Specific problem : large number of levels





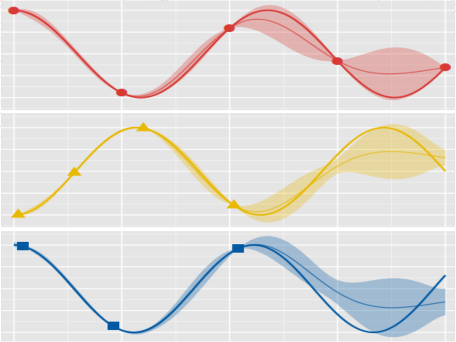


Form (s, c, p) $Z \in \{1, \ldots, 94\}$ $E \in \{E_1, E_2, E_3, E_4, E_5, E_6\}$

GP interpretation when no distance is available

A GP for $(x, u) \in [0, 1] \times \{"red", "yellow", "blue"\}$ can be defined with :

- a kernel on $[0, 1]$, i.e. a **covariance function**
- a kernel on $\{"red", "yellow", "blue"\}$, i.e. a **covariance matrix**
- a valid operation between them, such as *, $+$, ...



Example : $\mathrm{Cov}(Y(x, "blue"), Y(x', "red")) = k(x, x') \times 0.8$

GP for categorical inputs : group kernels for partitioned levels

A group kernel is a block covariance matrix

Block covariance matrix

$$\mathbf{T} = \begin{pmatrix} \mathbf{W}_1 & \mathbf{B}_{1,2} & \cdots & \mathbf{B}_{1,G} \\ \mathbf{B}_{2,1} & \mathbf{W}_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbf{B}_{G-1,G} \\ \mathbf{B}_{G,1} & \cdots & \mathbf{B}_{G,G-1} & \mathbf{W}_G \end{pmatrix}$$

with constant between-group blocks

Hierarchical (group/level) process

$$\eta_{g/\ell} = \mu_g + \lambda_{g/\ell}$$

with :

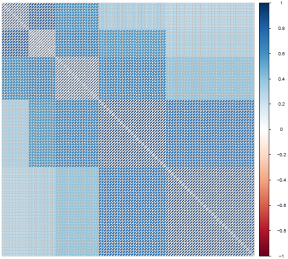
- Gaussian ind. priors for $\boldsymbol{\mu}, \boldsymbol{\lambda}_{g/.}$
- Centering cond. : $\sum_\ell \lambda_{g/\ell} = 0$

Main results [Roustant et al., 2020]

- Connection with hierarchical GPs : $\mathbf{T} = \mathrm{cov}(\boldsymbol{\eta})$.
- Characterization & parameterization of valid group kernels

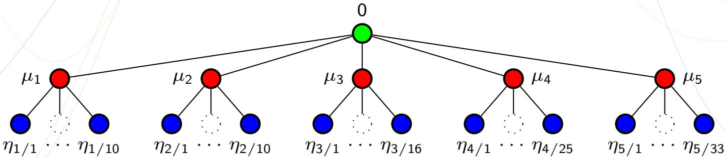
Result on the case study

For the categorical input 'chemical element', 5 groups are identified by experts
$\rightarrow$ Parsimonious parametrization with only 20 parameters (instead of $94 \times 95/2$)

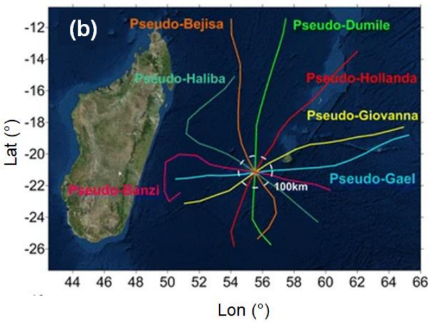


- 10 between-group covariances
- 5 within-group covariances
- 5 within-group variances

With a stratified LHS design of size 3×94, the Q^2 of the whole model is > 0.95

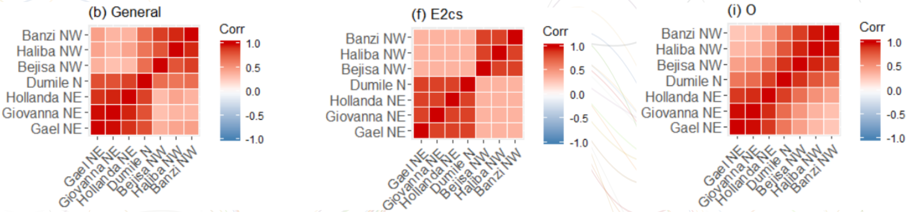


Application to cyclones [Rohmer et al., 2020]



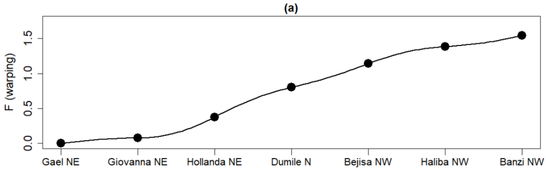
Output : significant wave height (maximum value over time)
Inputs : 5 continuous inputs + 1 categorical (cyclone profile)

Application to cyclones



Here two kernels for the categorical input give a 'good' GP model : a group kernel, or an ordinal kernel. Global kernel used for the second case :

$$k((x,u),(x',u')) = k_{\text{cont}}(x,x')k_{\text{cont2}}(F(u),F(u'))$$



Outline

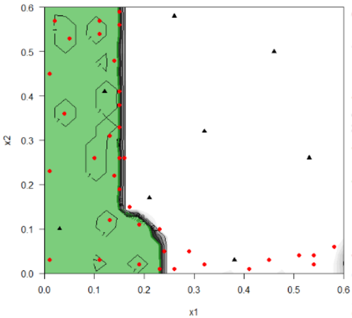
Inversion under uncertainty [El Amri et al., 2020]

Automotive test-case : Estimate

$$\Gamma_{f,V} = \{x, \mathbb{E}_V(f(x, V)) \leq c\},$$

where f is the pollutant concentration, and V represents uncontrolled driving cycles.

Mathematical tools : GP + random sets.



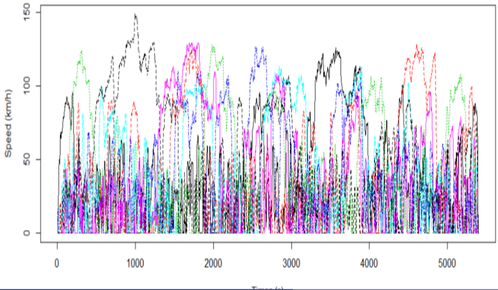
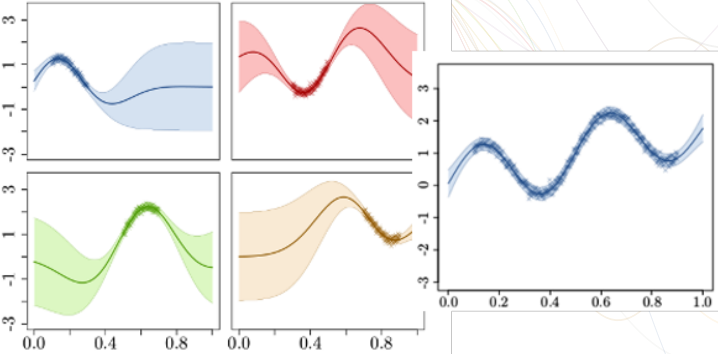


Figure – Left : examples of driving cycles. Top : estimation of $\Gamma_{f,V}$ (green area), after 37 iterations. Black triangles : initial data ; Red points : added data.

Outline

1. Context and motivation
2. The OQUAIDO Chair
3. Categorical inputs
4. Functional inputs/outputs
5. **High nb of data**
6. Specific constraints
7. Other topics and software
8. The CIROQUO consortium

Nested Kriging : Aggregating submodels



Building a GP model on a large dataset is done by linearly aggregating submodels.
Description and proof of consistency can be found in [Rullière et al., 2018].
Tested with 400 000 points in dimension 34 on a Storengy test case.

Outline

GP under linear inequalities : Impact on uncertainty quantification

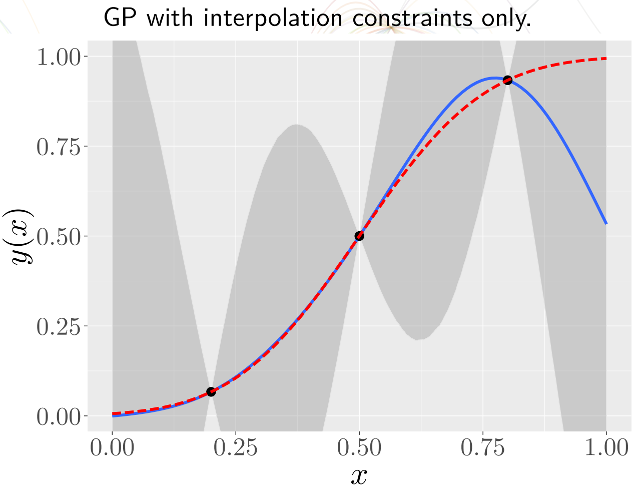


Illustration on a toy example (cdf of a Normal distribution)

GP under linear inequalities : Impact on uncertainty quantification

GP with boundedness + monotonicity additional constraints.

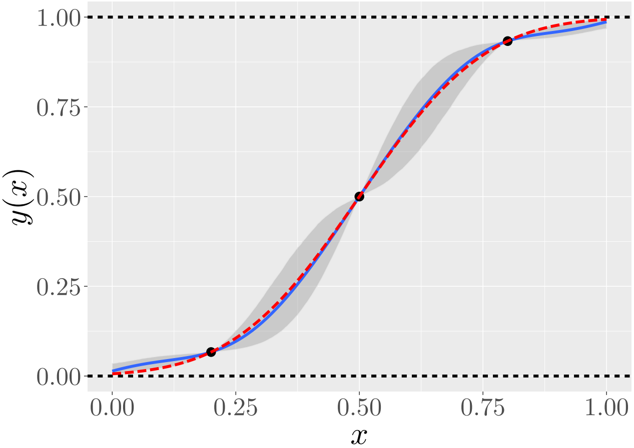


Illustration on a toy example (cdf of a Normal distribution)

GP and linear inequalities : Some theory

A finite elements (P_1) model for 1D GPs

[Maatouk and Bay, 2017, López-Lopera et al., 2018]

Each sample path of a GP Y is approximated by a **piecewise affine function**

$$Y_m(x) = \sum_{j=1}^{m} \xi_j \phi_j(x)$$

where ϕ_j are "hat" functions and $\boldsymbol{\xi}$ is a Gaussian vector extracted from Y

- Key point : Boundedness, monotonicity (and others) for a piecewise affine function can be checked only at knots → finite number of conditions only

GP and linear inequalities : Some theory

A finite elements (P_1) model for 1D GPs

[Maatouk and Bay, 2017, López-Lopera et al., 2018]

Each sample path of a GP Y is approximated by a **piecewise affine function**

$$Y_m(x) = \sum_{j=1}^{m} \xi_j \phi_j(x)$$

where ϕ_j are "hat" functions and $\boldsymbol{\xi}$ is a Gaussian vector extracted from Y

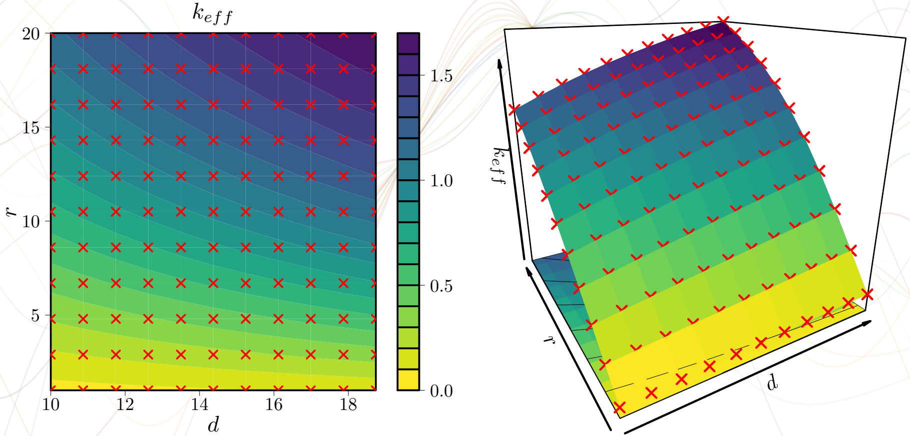
- Key point : Boundedness, monotonicity (and others) for a piecewise affine function can be checked only at knots → finite number of conditions only

Key feature

All paths / predictions **fullfill inequality constraints everywhere in the space**.

Remark : Immediate extension in 2D (and higher) by using tensors $\phi_{j_1}(x_1)\phi_{j_2}(x_2)$.

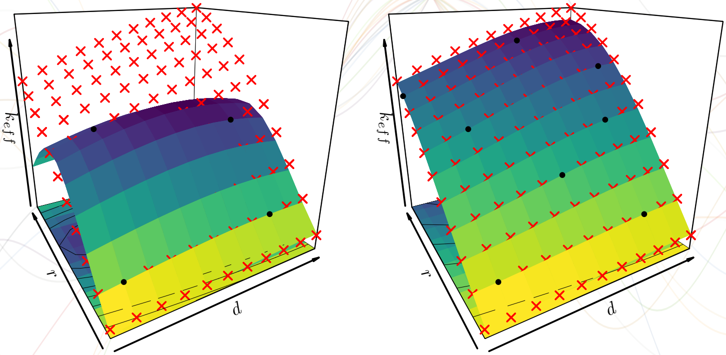
Example of application in 2D



Nuclear criticality safety assessments : IRSN's dataset.

Extra information : k_{eff} is positive and non-decreasing.

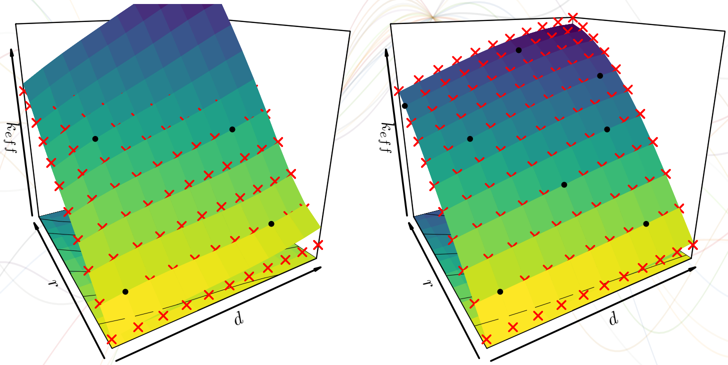
Example of application in 2D



Unconstrained model + MLE.

Monotonicity constraints are nearly learnt with 8 points, but not with 4 points.

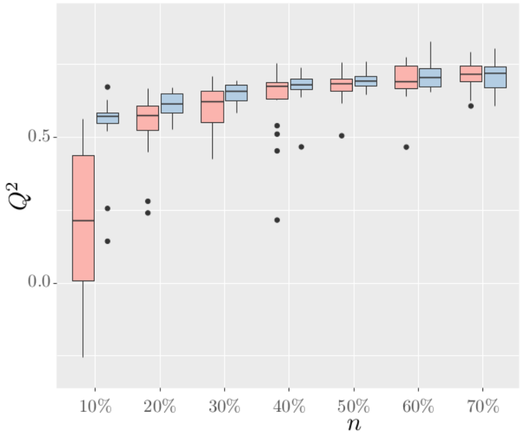
Example of application in 2D



Constrained model + constrained MLE.

Monotonicity constraints are fulfilled everywhere in the space, whatever the size of the training set.

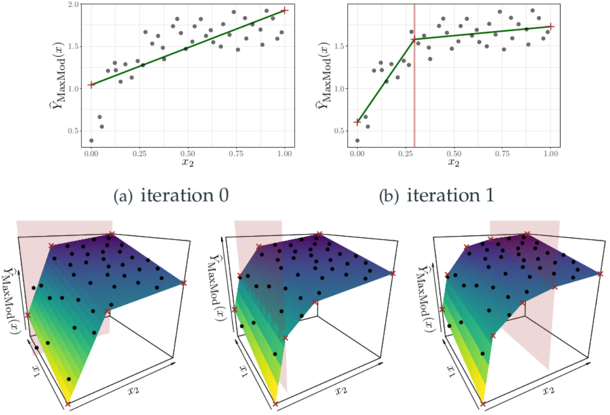
A 5D application (BRGM coastal flooding case study [López-Lopera et al., 2020])



With a fraction (e.g. 10%) of the total budget ($n = 200$), the constrained model (blue boxplots) outperforms the unconstrained one $\Rightarrow$ save budget !

A sequential algorithm to go to higher dimensions [Bachoc et al., 2020]

The MaxMod algorithm adds a knot / variable s.t. the L^2 variation of the mode a posteriori is maximum. Works in dimension 20 when 5 variables are really active.



(a) iteration 0 (b) iteration 1

(c) iteration 2 (d) iteration 3 (e) iteration 4

• training points + knots ■ MAP estimate

Theoretical results

Correspondence with spline under inequality [Bay et al., 2016]

Let $\widehat{Y}_m$ be the mode a posteriori (MAP) of $Y_m(x) = \sum_{j=1}^{m} \xi_j \phi_j(x)$, defined by replacing ξ by the mode of the distribution of ξ conditional on the constraints. Then, when the number of knots m tends to infty,

$$\hat{Y}_m \underset{\text{unif.}}{\longrightarrow} \underset{f \in \mathcal{H} \cap \mathcal{C} \cap \mathcal{I}}{\operatorname{argmin}} \|f\|$$

where $\mathcal{C}$ is a convex set of inequality constraints, $\mathcal{I}$ the set of interpolation constraints $f(x_i) = y_i$ $(i = 1, \dots, n)$, and $\mathcal{H}$ is the RKHS associated to Y.

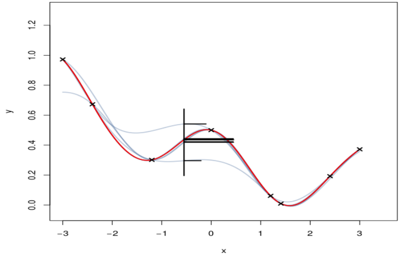
Convergence of the MaxMod algorithm [Bachoc et al., 2020]

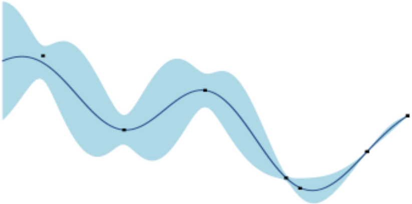
Let $\widehat{Y}_{\text{MaxMod},m}$ be the MAP at iteration m of the MaxMod algorithm. Then,

$$\hat{Y}_{\text{MaxMod},m} \underset{\text{unif.}}{\longrightarrow} \underset{f \in \mathcal{H} \cap \mathcal{C} \cap \mathcal{I}}{\operatorname{argmin}} \|f\|$$

Outline

{General Surrogate}-based optimization [Ben Salem et al., 2017]

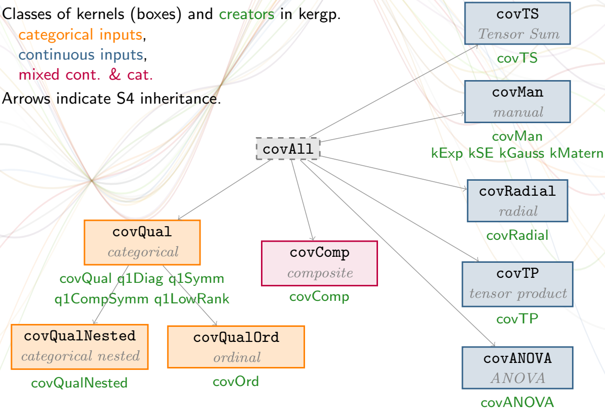




Bayesian optimization can be used with any surrogate model (e.g. SVM) via the Universal Prediction distribution, built by cross-validation.

Open source R packages

- 4 R packages released : *lineqGPR*, *nestedKriging*, *specgp* and *kergp*



Outline

The future of OQUAIDO : CIROQUO

Further information on OQUAIDO (2016 - 2020)

This talk has presented a selection of works done in OQUAIDO.
Read more in our activity report, now on HAL [Roustant et al., 2021]

Next iteration : CIROQUO (2021 - 2025)

- Academics : Centrale Lyon, INRIA, Mines St-Etienne, Univ. Nice & Toulouse
- Technos : IFPEN, BRGM, CEA, IRSN, Storengy
- Experts : M. Mougeot (ENSIIE), P. Havé
- Research lines
  - Calibration, validation and transposition of computer codes
  - Metamodeling for complex environments
  - Optimization and inversion under uncertainty
- New directions
  - more interaction with machine learning
  - wide language software *libKriging*
- Website : ciroquo.ec-lyon.fr

Références I

Aronszajn, N. (1950).
Theory of reproducing kernels.
Transactions of the American mathematical society, 68(3) :337–404.

Bachoc, F., Lopera, A. F. L., and Roustant, O. (2020).
Sequential construction and dimension reduction of gaussian processes under inequality constraints.
arXiv preprint arXiv :2009.04188.

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

Bect, J., Bachoc, F., and Ginsbourger, D. (2019).
A supermartingale approach to Gaussian process based sequential design of experiments.
Bernoulli, 25(4A) :2883 – 2919.

Ben Salem, M., Roustant, O., Gamboa, F., and Tomaso, L. (2017).
Universal prediction distribution for surrogate models.
SIAM/ASA Journal on Uncertainty Quantification, 5(1) :1086–1109.

Références II

Berlinet, A. and Thomas-Agnan, C. (2011).
Reproducing kernel Hilbert spaces in probability and statistics.
Springer Science & Business Media.

Chevalier, C., Bect, J., Ginsbourger, D., Vazquez, E., Picheny, V., and Richet, Y. (2014).
Fast parallel kriging-based stepwise uncertainty reduction with application to the identification of an excursion set.
Technometrics, 56(4) :455–465.

El Amri, M. R., Helbert, C., Lepreux, O., Zuniga, M. M., Prieur, C., and Sinoquet, D. (2020).
Data-driven Stochastic Inversion via Functional Quantization.
Statistics and Computing, 30(3) :525–541.

Jones, D. R., Schonlau, M., and Welch, W. J. (1998).
Efficient global optimization of expensive black-box functions.
Journal of Global Optimization, 13(4) :455–492.

Kimeldorf, G. and Wahba, G. (1971).
Some results on Tchebycheffian spline functions.
Journal of mathematical analysis and applications, 33(1) :82–95.

Références III

López-Lopera, A. F., Bachoc, F., Durrande, N., Rohmer, J., Idier, D., and Roustant, O. (2020).
Approximating gaussian process emulators with linear inequality constraints and noisy observations via mc and mcmc.
In Tuffin, B. and L'Ecuyer, P., editors, *Monte Carlo and Quasi-Monte Carlo Methods*, pages 363–381, Cham. Springer International Publishing.

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

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

Močkus, J. (1975).
On Bayesian methods for seeking the extremum.
In Marchuk, G. I., editor, *Optimization Techniques IFIP Technical Conference Novosibirsk, July 1–7, 1974*, pages 400–404, Berlin, Heidelberg. Springer Berlin Heidelberg.

Références IV

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

Rohmer, J., Roustant, O., Lecacheux, S., and Manceau, J.-C. (2020).
Revealing the dependence structure of scenario-like inputs in numerical environmental simulations using Gaussian Process regression.
Document de travail.

Roustant, O., Le Riche, R., Garnier, J., Ginsbourger, D., Deville, Y., Helbert, C., Pronzato, L., Prieur, C., Gamboa, F., Bachoc, F., Rohmer, J., Perrin, G., Marrel, A., Damblin, G., Gliere, A., Sinoquet, D., Richet, Y., Da Veiga, S., and Huguet, F. (2021).
Chair in applied mathematics OQUAIDO Activity report.
Research report, Mines Saint-Etienne ; Ecole Centrale Lyon ; BRGM (Bureau de recherches géologiques et minières) ; CEA ; IFP Energies Nouvelles ; Institut de Radioprotection et de Sûreté Nucléaire ; Safran Tech ; Storengy ; CNRS ; Université Grenoble - Alpes ; Université Nice - Sophia Antipolis ; Université Toulouse 3 (Paul Sabatier).

Roustant, O., Padonou, E., Deville, Y., Clément, A., Perrin, G., Giorla, J., and Wynn, H. (2020).
Group kernels for Gaussian process metamodels with categorical inputs.
SIAM/ASA Journal on Uncertainty Quantification, 8(2) :775–806.

Références V

Rullière, D., Durrande, N., Bachoc, F., and Chevalier, C. (2018).
Nested Kriging predictions for datasets with a large number of observations.
Statistics and Computing, 28(4) :849–867.

Vazquez, E. and Bect, J. (2010).
Convergence properties of the expected improvement algorithm with fixed mean and covariance functions.
Journal of Statistical Planning and Inference, 140(11) :3088 – 3095.