DESIGN OF EXPERIMENTS FOR COMPUTER CODE CALIBRATION

Adama BARRY

Thesis supervisors : François BACHOC[1] & Clémentine PRIEUR[2]

Promoters : Sarah BOUQUET[3] & Miguel MUNOZ ZUNIGA[3]

[1]Institut de Mathématiques de Toulouse

[2]Université Grenoble Alpes & INRIA Grenoble

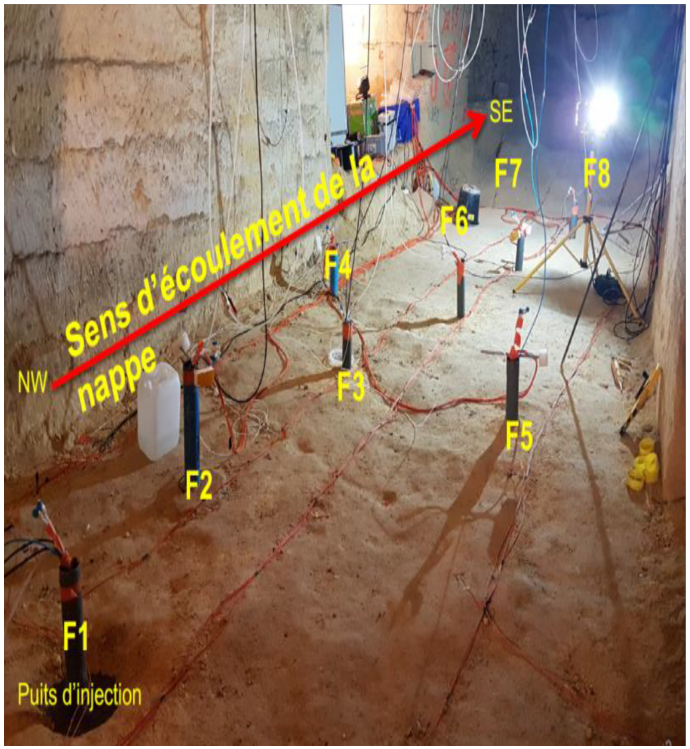
[3]IFP Énergies Nouvelles

GDR MASCOTNUM 31 Mai-2023

Plan

Plan

Application context : monitoring a CO_2 leak



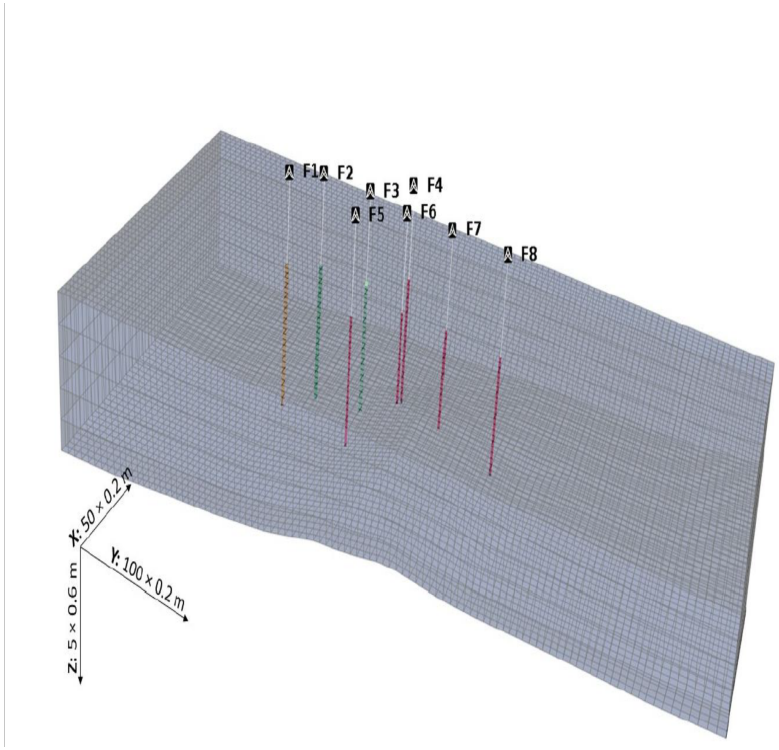
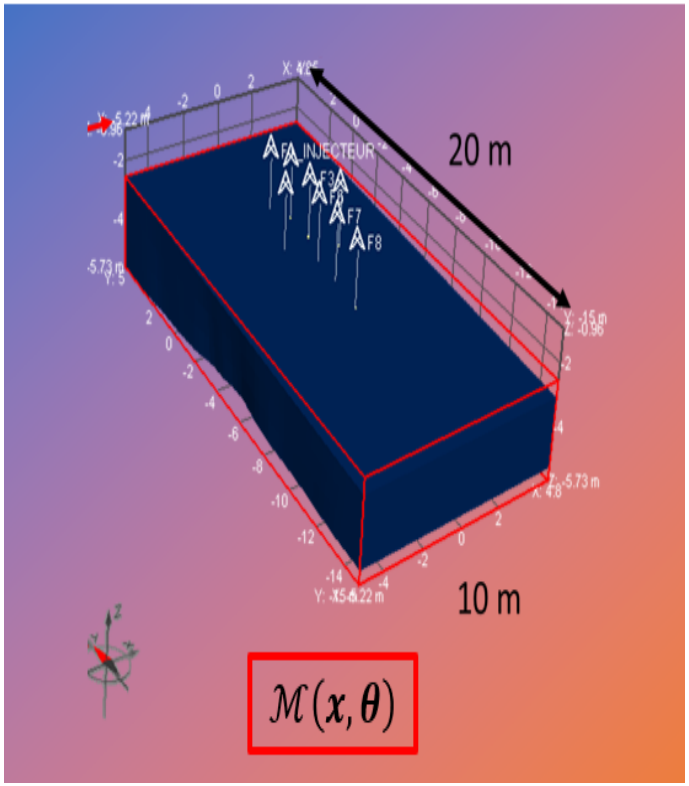


Figure: Monitoring CO_2 concentration.

- Physical phenomenon: CO_2 concentration in observation wells over time.
- Control variables : position of 6 observation wells $(F_2, F_3 \ldots, F_7)$.

Application context : computer code



- Computer code : 3D flow model.
- Input variables :
  - Control variable x : position of observation wells.
  - Calibration parameters θ :

Paramètres	**Min**	**Max**
αL /disp X (m)	0.13	4.5
αT /disp X (m)	0.04	1.5
Porosité	0.2	0.45
Perméabilité	0.1	48.24

Application context : questions and challenges

Questions

- Where to place the 6 observation wells ?
- At what times should the measuring instruments be moved to characterize the displacement of CO_2?
- What parameter values should be set for the simulator?
- Arrival time of CO_2 in well F_8? (critical case)

Challenges

- Measurement of physical phenomenon very costly.
- Costly simulation (several minutes or hours per simulation).
- Functional code output and potentially functional parameters.

Computer code as black box function

- $\mathcal{X}$: experimental domain.
- Θ : space of parameters.

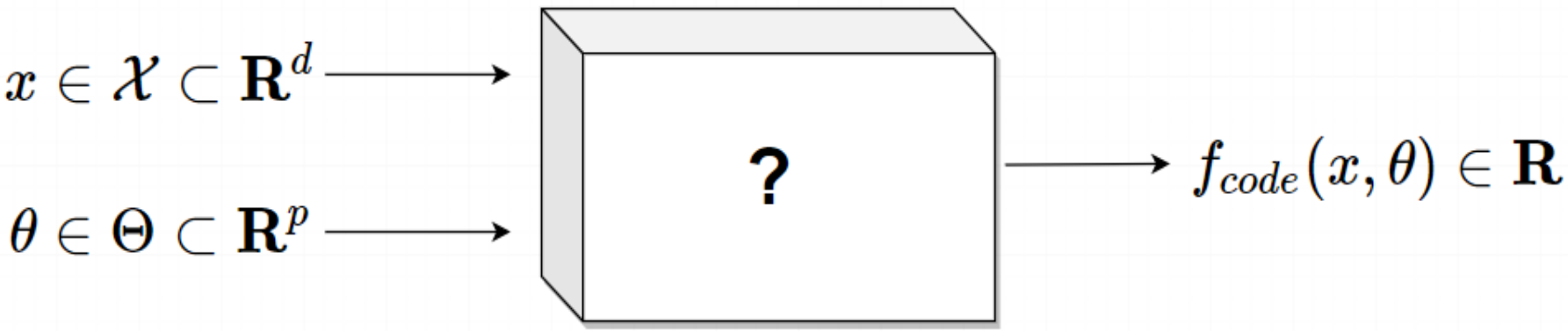


Figure: Scalar output computer code.

Statistical framework

- Relationship between physical observation and simulation :

$$f_{obs}(x) = f_{code}(x, \theta_0) + \varepsilon_x, \tag{1}$$

  where $\theta_0 \in \Theta$ the true value, $\varepsilon_x \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ et σ_ε^2 is assumed known.
- Design of physical experiments $X_{obs} = \left[x^{(1)}, \ldots, x^{(n)}\right]^T$.
- Physical observations $Y_{obs} = \left[y_{obs}^{(1)} := f_{obs}(x^{(1)}), \ldots, y_{obs}^{(n)} := f_{obs}(x^{(n)})\right]^T$.
- Prior distribution : $\pi(\theta)$.
- Exact posterior distribution :

$$\pi(\theta \mid Y_{obs}) \propto \frac{1}{\sigma_\varepsilon^n} \exp\Big(- \frac{1}{2\sigma_\varepsilon^2} SS(\theta)\Big) \pi(\theta), \tag{2}$$

  where $SS(\theta) = \sum_{i=1}^{n} \left(y_{obs}^{(i)} - f_{code}(x^{(i)}, \theta)\right)^2$.

Statistical framework

Goals

- The choice of an design of physical experiments $X_{obs} = \left[x^{(1)}, \ldots, x^{(n)}\right]^T$ to minimize uncertainty about θ_0.
- Estimate the vector of parameters θ_0.

Approaches :

- Quality criteria for physical experiments.
- Criterion for design of numerical experiments for building the metamodel.

Plan

Gaussian process metamodel

Goal

Build an surrogate model for the computer code.

A priori : f_{code} is the realization of a Gaussian process :

$$\mathbf{Y}_{code} \sim \mathbf{GP}(m_\beta, k),$$

where

- $m_\beta : (x,\theta) \longmapsto m_\beta(x,\theta) = h(x,\theta)^T\beta$: mean function,
- $k : \big((x,\theta),(x',\theta')\big) \longmapsto k\big((x,\theta),(x',\theta')\big)$: covariance function.

Consider the following :

- $D_M = \big[(x_1,\theta_1),\ldots,(x_M,\theta_M)\big]^T$ the design of numerical experiment,
- $f_{code}(D_M) = \big[f_{code}(x_1,\theta_1),\ldots,f_{code}(x_M,\theta_M)\big]^T$ the simulations.

A posteriori : $Y_M := \mathbf{Y}_{code} \mid f_{code}(D_M) \sim \mathbf{GP}\Big(\mu^M, k^M\Big)$,

where μ^M et k^M are respectively the a posteriori mean function and the a posteriori covariance function.

Approximation of the exact a posteriori distribution

Preliminary :

- The metamodel parameters are estimated using the modularization technique (Liu, Bayarri and Berger 2009).
- Assumption: A priori on θ and f_{code} independent.

Approximation of the exact posterior distribution

$$\pi(\theta \mid Y_{obs}, f_{code}(D_M)) \propto \mathcal{L}^c(Y_{obs} \mid f_{code}(D_M), \theta)\pi(\theta). \tag{3}$$

with

$$\begin{aligned}\mathcal{L}^c\left(Y_{obs} \mid \theta, f_{code}(D_M)\right) = & \frac{1}{(2\pi)^{n/2}\det\left(k^M(D_\theta) + \sigma_\varepsilon^2 I_n\right)^{1/2}} \\ & \times \exp\left[-\frac{1}{2} \mid\mid Y_{obs} - \mu^M(D_\theta) \mid\mid_{k^M(D_\theta)+\sigma_\varepsilon^2 I_n}\right],\end{aligned} \tag{4}$$

where $D_\theta = \{(x^{(1)}, \theta), \ldots, (x^{(n)}, \theta)\}$.

The final objective ?

Hybrid algorithm for computer code calibration

Step 1 : Build an initial metamodel for f_{code} with simulations on LHS-maxin design on $\mathcal{X} \times \Theta$.

Step 2 : Use the metamodel to select the design of physical experiments using quality criteria.

Step 3 : Select the design of numerical experiments to improve the metamodel.

Step 4 : Approximation of the exact a posteriori distribution and estimation of the CIs of the calibration parameters.

Plan

Criterion of Design Of Physical Experiments

- $X = \{x^{(1)}, \ldots, x^{(n)}\} \in \mathcal{X}^n$ is a design of physical experiments.

Goal

Choose the optimal design of physical experiments X_{obs} such that :

$$X_{obs} = \arg\min_{X \in \mathcal{X}^n} \mathbf{C}(X) \text{ or } \arg\max_{X \in \mathcal{X}^n} \mathbf{C}(X),$$

where $\mathbf{C}(X)$ is the quality criterion to be defined.

Two types of criterion in litterature :

1. Criterion based on Fisher's information matrix.
2. Criterion based on posterior distribution (DAP).

Fisher information matrix's criteria for DOPE

- Computer code approximation : $\hat{f}_{code,D}$ built using $(D, f_{code}(D))$, where D is LHS-maximin design on $\mathcal{X} \times \Theta$,
- The information matrix in θ_0 is given by :

$$M(X, \theta_0) = J(X, \theta_0)^T C_\varepsilon^{-1} J(X, \theta_0). \tag{5}$$

where $J(X, \theta_0) = \left(\frac{\partial \hat{f}_{code,D}(x^{(i)}, \theta_0)}{\partial \theta_j} \right)_{1 \leq i \leq n, 1 \leq j \leq p}$ is the Jacobian.

- The Bayesian information matrix in θ_0 :

$$M_b(X, \theta_0) = M(X, \theta_0) + M_0(\theta_0), \tag{6}$$

with $M_0(\theta_0) = \frac{\partial^2}{\partial \theta^2} \log \pi(\theta_0)$ the precision matrix.

- Optimality function :

$$\psi : M \in \mathcal{S}_p^+(\mathbf{R}) \longmapsto \psi(M) \in \mathbf{R}$$

Examples : determinant, trace, maximum eigenvalue, minimum eigenvalue, etc.

Fisher information matrix's criteria for DOPE

1. Eψ-optimality criterion :

$$\mathbf{C}(X) = \mathbf{E}_{\theta_0}\left[\psi(M_b(X, \theta_0))\right]. \tag{7}$$

2. Robuste min-max-optimality criterion :

$$\mathbf{C}(X) = \min_{\theta_0 \in \Theta} \psi(M_b(X, \theta_0)) \text{ ou } \max_{\theta_0 \in \Theta} \psi(M_b(X, \theta_0)). \tag{8}$$

Examples :

- $\psi(M) = det(M)$, ED-optimality, to be maximized (Fedorov 1980; Pronzato and Walter 1985),
- $\psi(M) = Tr(M^{-1})$ MMIT-optimality, to be maximized (Pronzato and Walter 1985).

Inconveniences

- Local criteria : we use linear approximation of the code in the neighbourhood of θ_0.
- Uncertainty of physical observations not taken into account.

Kullback Leibler criterion for DOPE

Definition (Divergence de Kullback-Leibler)

Let p and q be two probability distributions on the same Ω space. The Kullback-Leibler divergence is defined as :

$$\mathbf{KL}(p \mid\mid q)) = \int_{\Omega} \log \frac{p(x)}{q(x)} p(dx).$$

Information measurement of design X (Abellan and Noetinger 2010) :

$$\mathbf{KL}\left[\pi(\theta \mid Y(X,\theta_0), f_{code}(D)) \mid\mid \pi(\theta)\right] \tag{9}$$

where

- $Y(X,\theta_0) = \hat{f}_{code,D}(X,\theta_0) + \varepsilon_X$, où $\varepsilon_X \sim \mathcal{N}(0, C_\varepsilon)$ et $\theta_0 \sim \pi(\theta)$.

We can then define the EKL-optimality criteria :

$$\mathbf{C}_{KL}(X) = \mathbf{E}_{\theta_0}\left[\mathbf{E}_{\varepsilon}\left(\mathbf{KL}\left[\pi(\theta \mid Y(X,\theta_0), f_{code}(D)) \mid\mid \pi(\theta)\right]\right)\right].$$

Posterior distribution's criteria for DOPE

- Simulation of physical observations :

$$\hat{f}_{code,D}(X, \theta_0) + \varepsilon_X,$$

where $\varepsilon_X \sim \mathcal{N}(0, C_\varepsilon)$ and $\theta_0 \sim \pi(\theta)$.

- Assessment of design of physical experiments quality :

$$\psi\Big(Cov(\theta \mid \hat{f}_{code,D}(X, \theta_0) + \varepsilon_X, f_{code}(D))\Big),$$

$$\mathbf{E}_\theta\Big[l(\theta, \theta_0) \mid \hat{f}_{code,D}(X, \theta_0) + \varepsilon_X, f_{code}(D)\Big],$$

where ψ is the optimality function and l is a loss function.

Examples :

- $\psi(M) = Tr(M)$: sum of a posteriori variances,
- $l(\theta, \theta_0) = \frac{1}{n} \mid\mid \theta - \theta_0 \mid\mid_1$: Mean Absolute Error a posteriori,
- $l(\theta, \theta_0) = \frac{1}{n} \mid\mid \theta - \theta_0 \mid\mid_2^2$: Mean Square Error a posteriori.

Posterior distribution's criteria for DOPE

- Considering uncertainties :

$$\mathbf{C_{cov}}(X) = \mathbf{E}_{\theta_0}\left\{\mathbf{E}_{\varepsilon_X}\left[\psi\left(Cov(\theta \mid \hat{f}_{code,D}(X,\theta_0) + \varepsilon_X, f_{code}(D))\right)\right]\right\}, \quad (10)$$

$$\mathbf{C_{loss}}(X) = \mathbf{E}_{\theta_0}\left\{\mathbf{E}_{\varepsilon_X}\left[\mathbf{E}_{\theta}\left(l(\theta,\theta_0) \mid \hat{f}_{code,D}(X,\theta_0) + \varepsilon_X, f_{code}(D)\right)\right]\right\}. \quad (11)$$

- Monte Carlo calculation costs : $o(N_{\theta_0} N_{\varepsilon_x} N_{mcmc})$.
- Fast calculation of criterion : $o(N_{\theta_0})$
  - Joint random sampling of $\theta_0 \sim \pi(\theta)$ and $\varepsilon_X \sim \mathcal{N}(0, \sigma_\varepsilon I_n)$ (we assume that $\theta \perp \varepsilon_x$),
  - Uniform sampling of θ and weighting by the approximation of the posterior distribution $\pi(\theta \mid \hat{f}_{code,D}(X,\theta_0) + \varepsilon_X, f_{code}(D))$.
- Sequential versions of these criteria are possible.

Optimization algorithm for optimal DOPE

- **Optimization problem :**

$$X^* = \underset{X \in \mathcal{X}^n}{\arg\max}\, \mathbf{C}(X)$$

  We recall $\mathcal{X} \subset \mathbf{R}^d$.

- **Some constraints :**
  - The Design $X \in \mathcal{X}^n \subset \mathbf{R}^{d \times n}$ is a matrix :

$$X = \begin{pmatrix} x_1^{(1)} & \dots & x_d^{(1)} \\ \vdots & \ddots & \vdots \\ x_1^{(n)} & \dots & x_d^{(n)} \end{pmatrix}, \tag{12}$$

  - An evaluation of the criteria $C(X)$ takes time,
  - No analytical expression for the gradient or the hessian of $C(X)$.

$\Longrightarrow$ A variant of the simulated annealing optimization algorithm.

Simulated Annealing for optimal DOPE

- Key idea : random perturbation per row.

1. **Initialization** :
   - Choose k_{max} the maximum number of iterations, T_0 the initial temperature, σ^2 the variance parameter and X_0 the initial design of physical experiment of size n by LHS-maximin.
2. **While** $k \leq k_{max}$ **do** :
   - Random perturbation :

$$X^* = X_k + \begin{pmatrix} 0 & \dots & 0 \\ \vdots & \ddots & \vdots \\ \varepsilon_1 & \dots & \varepsilon_d \\ \vdots & \ddots & \vdots \\ 0 & \dots & 0 \end{pmatrix},$$

où $\varepsilon_i \sim \mathcal{N}(0, \sigma^2), i = 1, \dots, d.$

Simulated Annealing for optimal DOPE

2. Continuation of step 2 :
   - Evaluating deterioration $\Delta_k = C(X_k) - C(X^*)$
   - Calculate $p = \max(e^{-\Delta_k / T_k}, 1)$
   - Draw $u \sim \mathcal{U}_{[0,1]}$
   - Accept-reject step :

$$X_{k+1} \longleftarrow \begin{cases} X^* & \text{if } p \geq u \\ X_k & \text{else.} \end{cases}$$

   - Update the temperature $T_{k+1} = cT_k$, with $0 \leq c \leq 1$
   - Update $k \longleftarrow k+1$

   End While.

3. Return $X_{k_{max}}$.

Plan

Design Of Numerical Experiments

- At this stage, physical data (X_{obs}, Y_{obs}) are available.

Goal

Choose the design of numerical experiments $D_M = \left[(x_1, \theta_1), \ldots, (x_M, \theta_M)\right]^T$.

Two approaches :

- Space filling design such as LHS-maximin (Pronzato and Müller 2012),
- Sequential design based on minimization of some criterion.

Damblin et al. 2018 criterion for DONE

Damblin et al. 2018 criterion :

$$D_M = \underset{D\in(\mathcal{X}\times\Theta)^M}{\arg\min}\ \mathbf{KL}(\pi(\theta \mid Y_{obs}) \mid\mid \pi(\theta \mid Y_{obs}, f_{code}(D))). \tag{13}$$

Heuristic : KL minimization (Damblin et al. 2018)

- Choose

  $D_M = D_0 \cup D_{M-M_0}$ where $D_0 \in (\mathcal{X}\times\Theta)^{M_0}$ and $D_{M-M_0} \in (X_{obs}\times\Theta)^{M-M_0}$,

  M_0 : the size of the initial design for building an initial metamodel.
- Equivalence between (13) and the following optimization problem :

$$\min_{\theta\in\Theta}\quad SS(\theta) = \sum_{i=1}^{n}\left(y_{obs}^{(i)} - f_{code}(x^{(i)},\theta)\right)^2 = \mid\mid Y_{obs} - f_{code}(D_\theta) \mid\mid_2^2 . \tag{14}$$

$\Longrightarrow$ **Resolution of** (14) **by the EGO algorithm based on the EI criterion.**

Plan

Analytical test case

Test function :

$$f_{code} : [0,1] \times [0,10] \times [5,10] \to \mathbf{R}$$
$$(x, \theta_1, \theta_2) \longmapsto \left(0.1(\theta_1 l_1(x) + l_2(\theta_2) l_1(x)^2)\right)^3,$$

où :

$$l_1 : x \in [0,1] \longmapsto l_1(x) = \frac{e^{10x-5}}{1+e^{10x-5}} \in [0,1],$$
$$l_2 : t \in [5,10] \longmapsto l_2(t) = \frac{e^{2t-15}}{1+e^{2t-15}} \in [0,1].$$

Test function

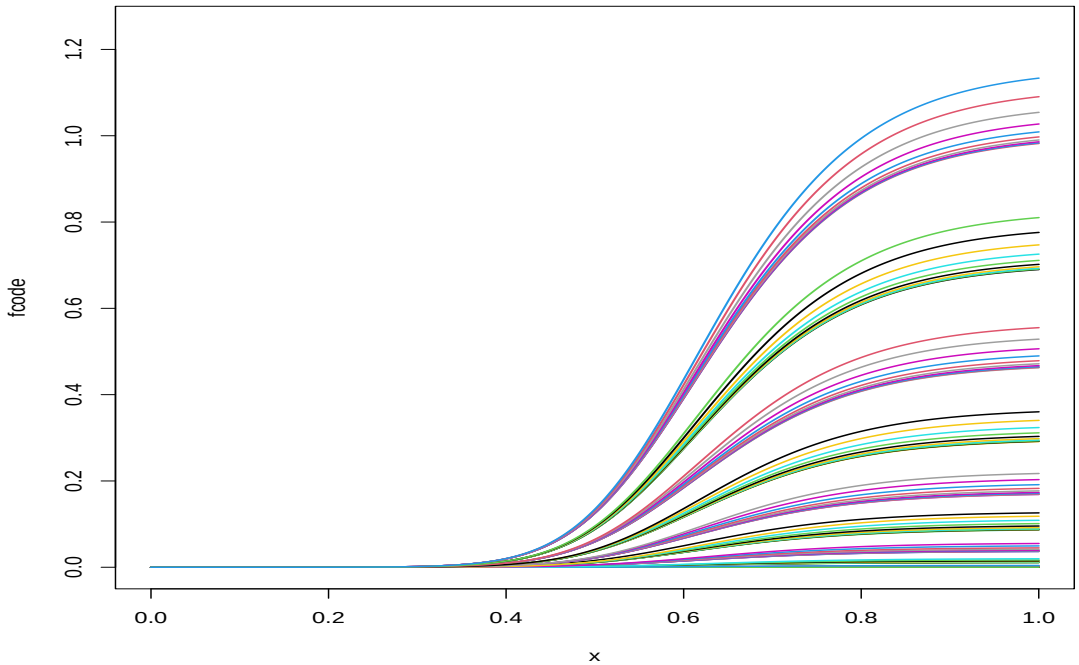


Figure: Test function for different values of $\theta_0 \in [0, 10] \times [5, 10]$.

Comparison of criteria for DOPE

We choose :

- The true value of $\theta_0 = (5, 7)$ and the noise variance $\sigma_\varepsilon^2 = 0.05^2$,
- The prior distribution $\pi(\theta) = \mathcal{U}_{[0,10]} \times \mathcal{U}_{[5,10]}$,
- The size of design of physical experiments $n = 10$,
- The size of design of numerical experiments $M = 80$,
- The Monte-Carlo size $L = 10^5$ for sampling the postrior.

To select the design of physical experiment :

- We build the initial metamodel with simulations on LHS-maximin design on $[0, 1] \times [0, 10] \times [5, 10]$ size $M_0 = 40$,
- We use the initial metamodel to choose the design of physical experiments using the following criteria :
  - $\mathbf{C}_{Det}$: criterion based on the determinant of information matrix,
  - $\mathbf{C}_{Trace}$: criterion based on the trace of information matrix,
  - $\mathbf{C}_{Var}$: criterion based on the sum of variance a posteriori,
  - $\mathbf{C}_{MSE}$: criterion based on the mean quadratic error a posteriori,
  - $\mathbf{C}_{KL}$: criterion based on the the Kullback-Leibler divergence.

Comparison of criteria for DOPE

For each design selected, we repeat $N = 50$ times the following procedure :

1. Simulate the physical observations :

$$Y = f_{code}(X, \theta_0) + \varepsilon_X, \text{ where } \varepsilon_X \sim \mathcal{N}(0, \sigma_\varepsilon^2 I_{10}).$$

2. Apply the calibration procedure and estimate credibility intervals.

We use two comparison metrics :

- Coverage rate **CR** : $\%\theta_0 \in IC_{95\%}$.
- Average lengths of CIs **ALCI**.

Criterion	**CR(%)**	**ALCI(θ_1)**	**ALCI(θ_2)**
Uniform	92	2.93	6.93
$\mathbf{C}_{Det}$	96	1.32	4.49
$\mathbf{C}_{Trace}$	92	1.85	4.74
$\mathbf{C}_{SVar}$	96	1.47	4.27
$\mathbf{C}_{MSE}$	98	1.18	4.26
$\mathbf{C}_{KL}$	98	2.12	4.16

Table: Coverage rates and Average lengths of CIs.

$\Longrightarrow$ In this example, it makes sense to select the physical experiments.

Plan

Conclusion

Synthesis :

- Gaussian process metamodel and parameter estimation technique,
- Quality criteria for design of physical experiments,
- Hybrid algorithms for designing experiments to calibrate computer codes,
- Fast optimization algorithms for physical design of experiments,
- There is a potential benefit in choosing the design of the physical experiments before the calibration procedure.

Thanks for your attention !!!

Références I

[1] A. Abellan and Benoit Noetinger. 'Optimizing Subsurface Field Data Acquisition Using Information Theory'. In: *Mathematical geosciences* 42 (Aug. 2010), pp. 603–630. DOI: 10.1007/s11004-010-9285-6.

[2] Mathieu Carmassi et al. 'Bayesian calibration of a numerical code for prediction'. working paper or preprint. Jan. 2018. URL: https://hal.archives-ouvertes.fr/hal-01677167.

[3] Xiaowu Dai and Peter Chien. *Another Look at Statistical Calibration: A Non-Asymptotic Theory and Prediction-Oriented Optimality*. 2018. DOI: 10.48550/ARXIV.1802.00021. URL: https://arxiv.org/abs/1802.00021.

[4] Guillaume Damblin et al. 'Adaptive Numerical Designs for the Calibration of Computer Codes'. In: *SIAM/ASA Journal on Uncertainty Quantification* 6.1 (2018), 151–179. ISSN: 2166-2525. DOI: 10.1137/15m1033162. URL: http://dx.doi.org/10.1137/15M1033162.

[5] Valery V. Fedorov. 'Convex design theory 1'. In: *Statistics* 11 (1980), pp. 21–43.

Références II

[6] Marc Kennedy and Anthony O'Hagan. 'Bayesian Calibration of Computer Models'. In: *Journal of the Royal Statistical Society Series B* 63 (Feb. 2001), pp. 425–464. DOI: 10.1111/1467-9868.00294.

[7] J. Kiefer. 'General Equivalence Theory for Optimum Designs (Approximate Theory)'. In: *The Annals of Statistics* 2.5 (1974), pp. 849–879. DOI: 10.1214/aos/1176342810. URL: https://doi.org/10.1214/aos/1176342810.

[8] J. Kiefer. 'Optimum Experimental Designs'. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 21.2 (1959), pp. 272–304. DOI: https://doi.org/10.1111/j.2517-6161.1959.tb00338.x. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1959.tb00338.x. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1959.tb00338.x.

[9] F. Liu, M. Bayarri and J. Berger. 'Modularization in Bayesian analysis, with emphasis on analysis of computer models'. In: *Bayesian Analysis* 4 (Mar. 2009). DOI: 10.1214/09-BA404.

Références III

[10] Luc Pronzato. 'Synthèse d'expériences robustes pour modèles à paramètres incertains'. Thèse de doctorat dirigée par Walter, Éric Informatique Paris 11 1986. PhD thesis. 1986, 1 vol. (222 p.) URL: http://www.theses.fr/1986PA112260.

[11] Luc Pronzato and Werner G. Müller. 'Design of computer experiments: space filling and beyond'. In: *Statistics and Computing* 22 (2012), pp. 681–701.

[12] Luc Pronzato and Eric Walter. 'Robust experiment design via stochastic approximation'. In: *Mathematical Biosciences* 75.1 (1985), pp. 103–120. ISSN: 0025-5564. DOI: https://doi.org/10.1016/0025-5564(85)90068-9. URL: https://www.sciencedirect.com/science/article/pii/0025564859006 89.

[13] Olivier Roustant, David Ginsbourger and Yves Deville. 'DiceKriging, DiceOptim: Two R Packages for the Analysis of Computer Experiments by Kriging-Based Metamodeling and Optimization'. In: *Journal of Statistical Software* 51.1 (2012), pp. 1–55. URL: https://www.jstatsoft.org/v51/i01/.

Références IV

[14] Miguel Munoz Zuniga & Delphine Sinoquet. *Optimization and bayesian approaches for model calibration. Application to oil and gas fiel management*. Présenté à l'Ateliers de Modélisation de l'Atmosphère (AMA). 2019.

[15] Valery V.Fedorov. *Theory of Optimal Experiments Designs*. Jan. 1972.