High-Dimensional Bayesian Optimization with a Combination of Kriging Models

Tanguy APPRIOU [(1), (2)], David GAUDRIE [(1)], Didier RULLIERE [(2)]

(1) STELLANTIS

(2) École des Mines de Saint-Etienne, LIMOS

MASCOT-NUM 2024

April 3rd 2024

OUTLINE

1) **Context**
   - Design optimization
   - Kriging
   - Bayesian Optimization

2) **Issues in high-dimension**

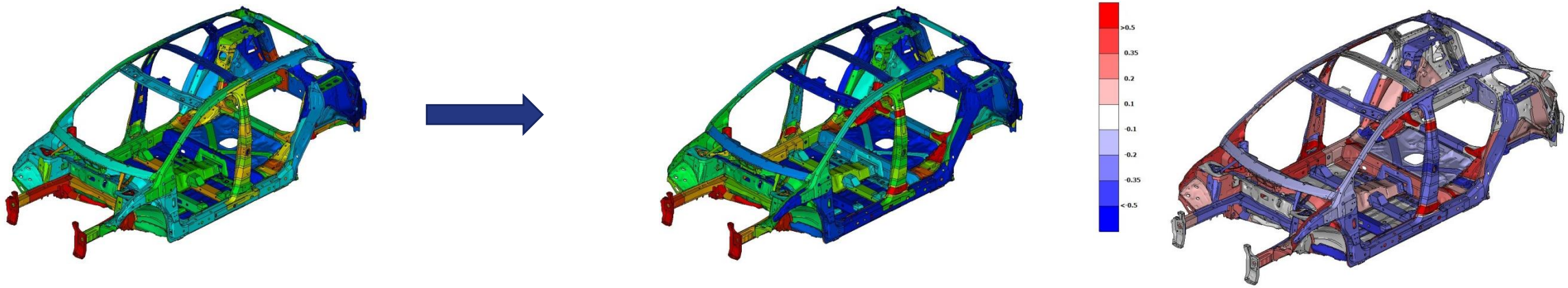
3) **Combination of Kriging models with random length-scales**

4) **Numerical results**

DESIGN OPTIMIZATION

- Design optimization is used to improve the performances of an engineering design.

Initial model **Optimized model** **Thickness difference**



Example: optimization of the Peugeot 3008 to minimize the vehicle weight while satisfying the norms for chock resistance.

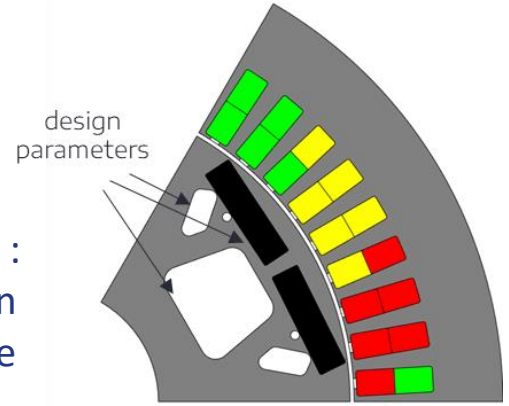
- Formally, we are interested in the optimization of a black-box function:

$$y : \boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d \rightarrow y(\boldsymbol{x}) \in \mathbb{R}.$$

→ We want to find the best design:

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x} \in \mathcal{X}} y(\boldsymbol{x}) .$$

Another example : optimization of an electrical machine



KRIGING

- We are in the context where **the black-box function y is expensive to evaluate**:

→ Evaluating the function for a single design can take hours.

↳ **We can only afford few observations**.

↳ We cannot use the usual optimization methods which require a large number of these evaluations.

Build a surrogate model

$$y(x) \approx \hat{y}(x)$$

Expensive true objective function

Cheap analytical approximation

- We dispose of n observations $\boldsymbol{Y} = \left(y(\boldsymbol{x_1}), \dots, y(\boldsymbol{x_n})\right)^T$ at the sample locations $\boldsymbol{X} = (\boldsymbol{x_1}, \dots, \boldsymbol{x_n})^T$.

→ The ordinary Kriging method approximates y as the realization of a Gaussian Process :

$$Y(.) \sim \mathcal{GP}\left(\mu, k_{\sigma,\boldsymbol{\theta}}(.,.)\right).$$

- $k_{\sigma,\boldsymbol{\theta}}(.,.)$ is the covariance function (kernel) with σ^2 the variance of the GP and $\boldsymbol{\theta} \in \mathbb{R}^d$ the covariance length-scales.

- We obtain the Kriging predictors for the mean and predictive variance by conditioning the GP Y over $\mathcal{D} = (\boldsymbol{X}, \boldsymbol{Y})$:

$$\hat{y}(\boldsymbol{x}) = \mathbf{E}(Y(\boldsymbol{x})|\mathcal{D}) = \mu + k(\boldsymbol{x}, \boldsymbol{X})\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X})^{-1}(\boldsymbol{Y} - \mathbf{1}\mu),$$

$$\hat{s}^2(\boldsymbol{x}) = \mathbf{Var}(Y(\boldsymbol{x})|\mathcal{D}) = k(\boldsymbol{x}, \boldsymbol{x}) - k(\boldsymbol{x}, \boldsymbol{X})\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X})^{-1}k(\boldsymbol{X}, \boldsymbol{x}).$$

COVARIANCE FUNCTION

The choice of the covariance function is very important to obtain a good prediction.

Popular choices of 1D stationary covariance are :

- Exponential : $k_{\sigma,\theta}(x,x') = \sigma^2 \exp\left(-\frac{|x-x'|}{\theta}\right)$,
- Gaussian : $k_{\sigma,\theta}(x,x') = \sigma^2 \exp\left(-\frac{(x-x')^2}{2\theta^2}\right)$,
- Matérn 5/2 : $k_{\sigma,\theta}(x,x') = \sigma^2 \left(1+\sqrt{5}\frac{|x-x'|}{\theta}+\frac{5(x-x')^2}{3\theta^2}\right)\exp\left(-\sqrt{5}\frac{|x-x'|}{\theta}\right)$,

Typically, the hyperparameters are optimized to maximize the log-likelihood of the model:

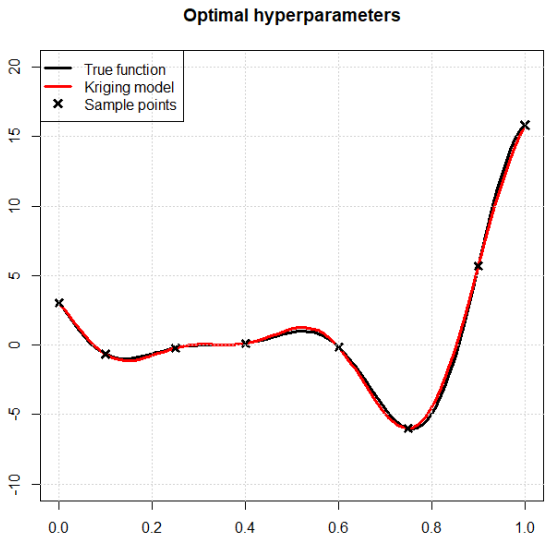
$$\mathcal{L}(\sigma,\boldsymbol{\theta}) = -\frac{1}{2}(\boldsymbol{Y}-\mu)^T \boldsymbol{K}_{\sigma,\boldsymbol{\theta}}^{-1}(\boldsymbol{Y}-\mu) - \frac{1}{2}\log\left|\boldsymbol{K}_{\sigma,\boldsymbol{\theta}}\right| - \frac{n}{2}\log(2\pi)\,.$$

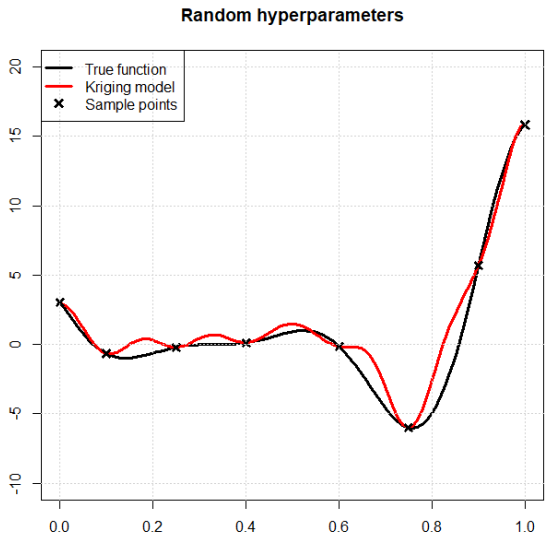
Denoting $\boldsymbol{R}$ the correlation matrix such that $\boldsymbol{K}_{\sigma,\boldsymbol{\theta}} = \sigma^2 \boldsymbol{R}_{\boldsymbol{\theta}}$, the MLE estimators for μ and σ^2 are:

$$\hat{\mu} = \frac{\mathbf{1}^T \boldsymbol{R}_{\boldsymbol{\theta}}^{-1} \boldsymbol{Y}}{\mathbf{1}^T \boldsymbol{R}_{\boldsymbol{\theta}}^{-1} \mathbf{1}}, \qquad \hat{\sigma}^2_{MLE} = \frac{1}{n}(\boldsymbol{Y}-\hat{\mu})^T \boldsymbol{R}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{Y}-\hat{\mu}).$$

And we obtain the length-scales by solving the minimization problem :

$$\hat{\theta}_{MLE} = \arg\min_{\boldsymbol{\theta}} \ \frac{n}{2}\log(\hat{\sigma}^2_{MLE}) + \frac{1}{2}\log(|\boldsymbol{R}_{\boldsymbol{\theta}}|)\,.$$





BAYESIAN OPTIMIZATION

In Bayesian optimization, we build the sampling plan sequentially by adding new training points to refine the model based on an acquisition criterion (see Jones et al., 1998).

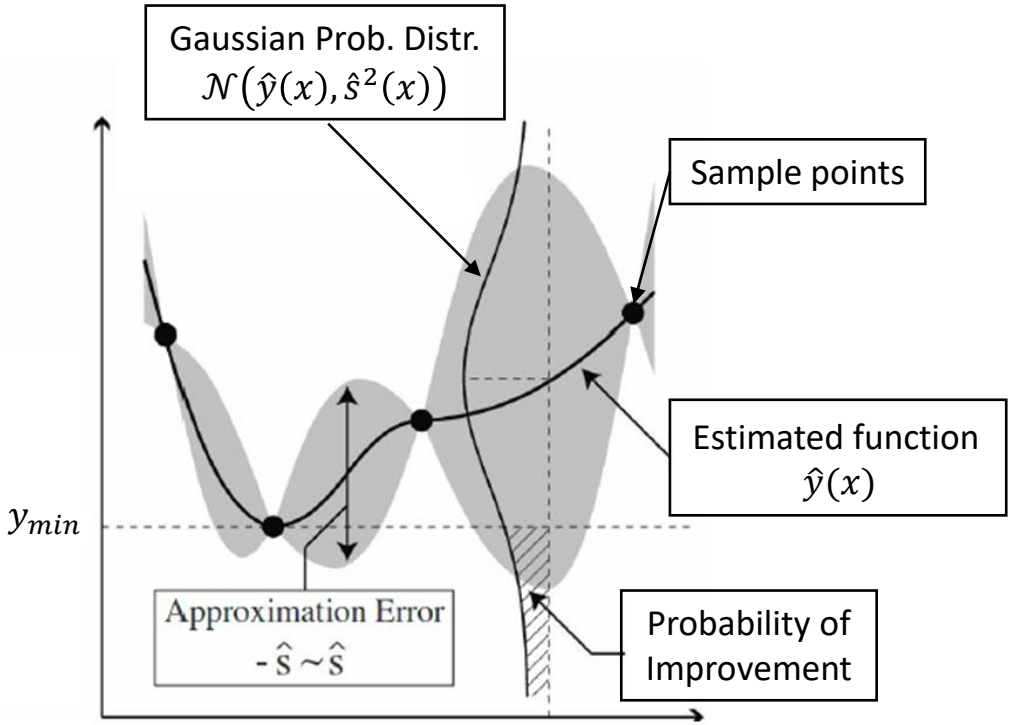
→ A popular acquisition criterion is the **Expected Improvement** (EI).

- The expected improvement is computed with both the mean estimate value and the model error estimate:

$$\mathbf{E}[I(\boldsymbol{x})] = \mathbf{E}\left(\left(y_{min} - Y(\boldsymbol{x})\right)^{+}\right)$$
$$= \left(y_{min} - \hat{y}(\boldsymbol{x})\right)\Phi\left(\frac{y_{min} - \hat{y}(\boldsymbol{x})}{\hat{s}(\boldsymbol{x})}\right) + \hat{s}(\boldsymbol{x})\phi\left(\frac{y_{min} - \hat{y}(\boldsymbol{x})}{\hat{s}(\boldsymbol{x})}\right).$$

- Φ and ϕ are respectively the cdf and the density of a standard normal distribution.

- EI balances local search around the optimum and global search where the model is not very accurate.

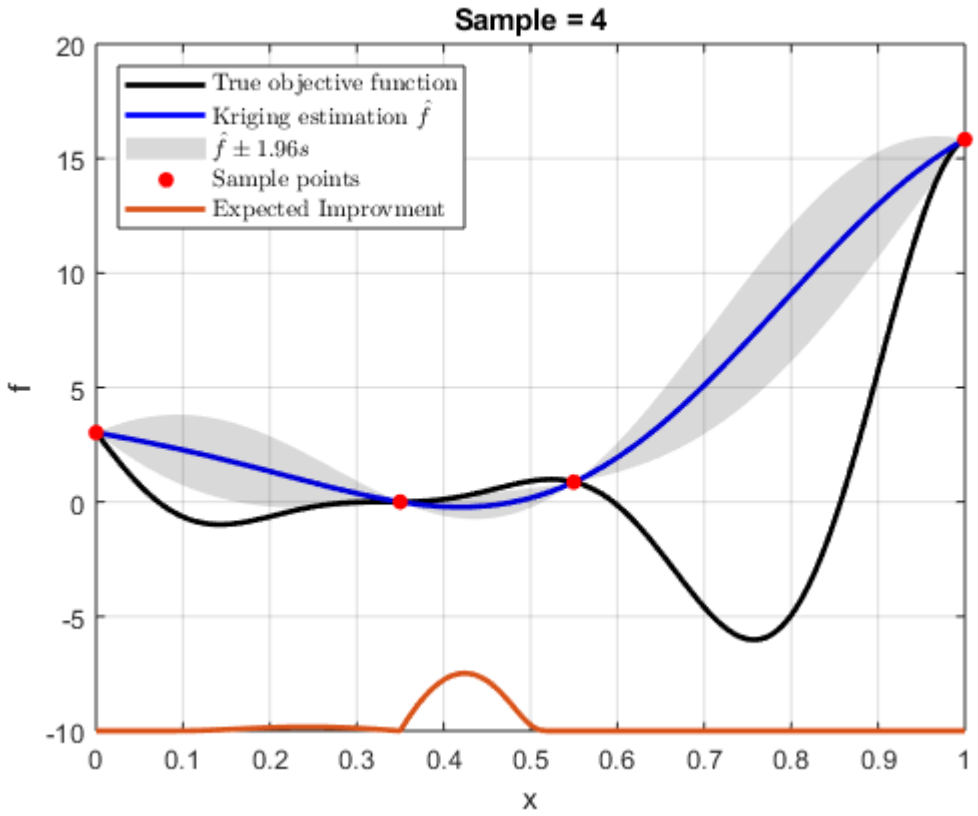


EXAMPLE

Here is an example of the optimization process for a 1D test function using the EGO algorithm.

$$f(x) = (6x - 2)^2 \sin(12x - 4)$$

We begin with 4 sample points, then we successively add 6 additional sample points.

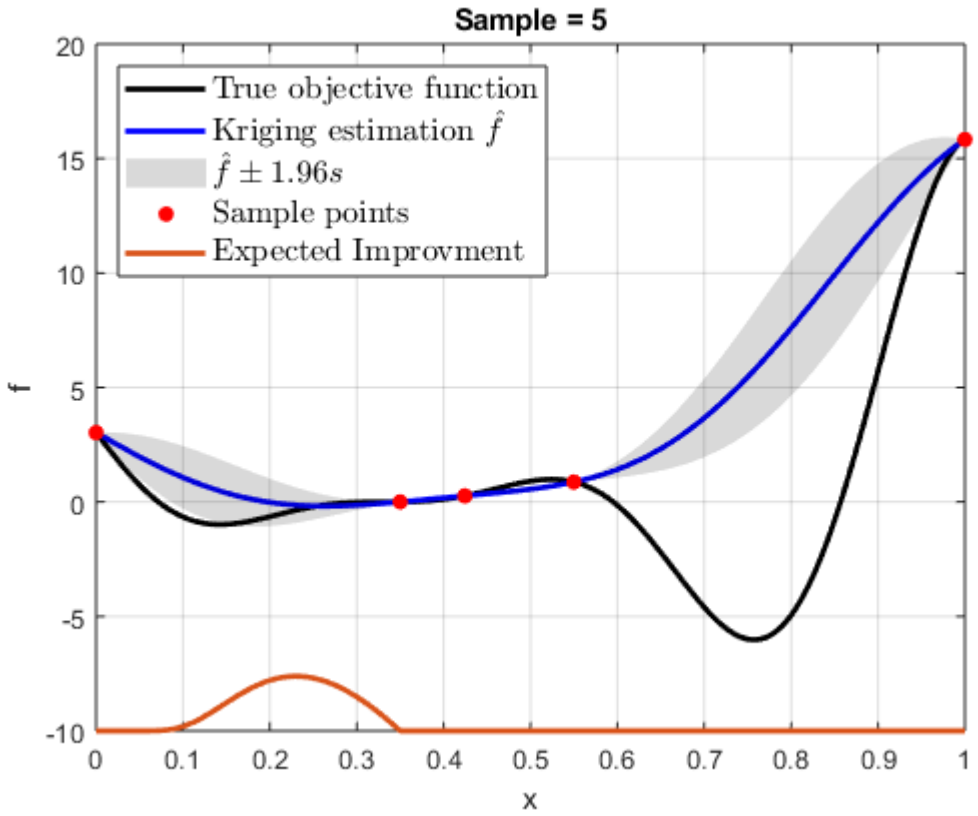


EXAMPLE

Here is an example of the optimization process for a 1D test function using the EGO algorithm.

$$f(x) = (6x - 2)^2 \sin(12x - 4)$$

We begin with 4 sample points, then we successively add 6 additional sample points.

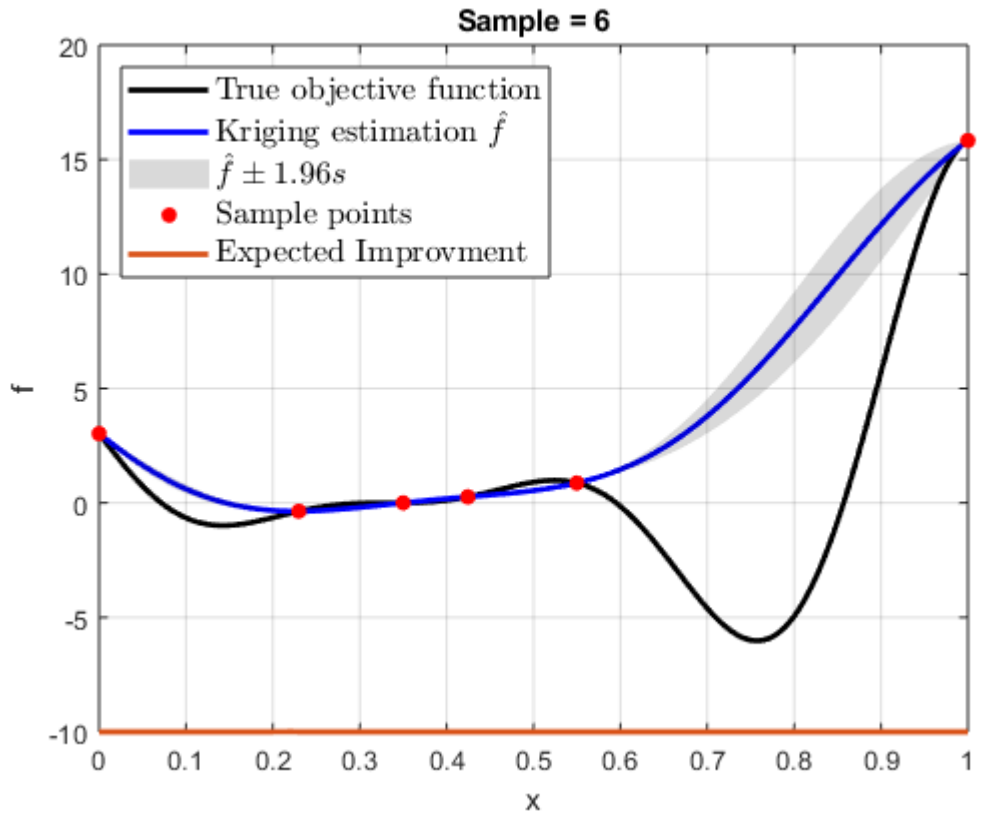


EXAMPLE

Here is an example of the optimization process for a 1D test function using the EGO algorithm.

$$f(x) = (6x - 2)^2 \sin(12x - 4)$$

We begin with 4 sample points, then we successively add 6 additional sample points.

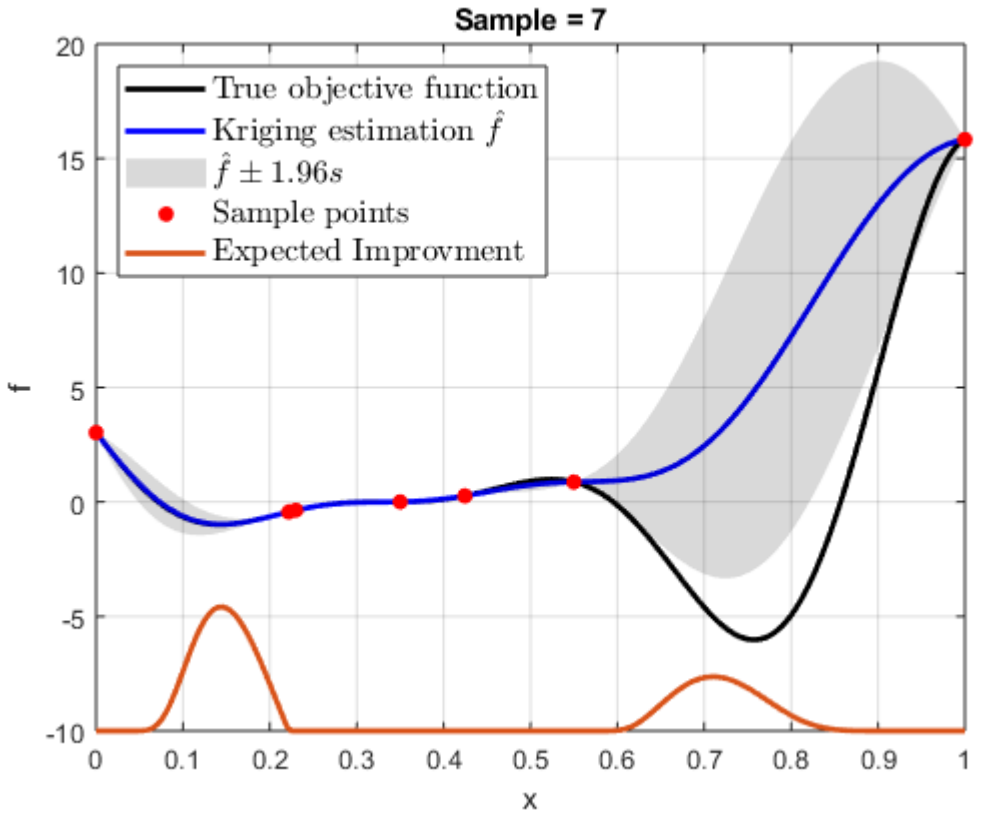


EXAMPLE

Here is an example of the optimization process for a 1D test function using the EGO algorithm.

$$f(x) = (6x - 2)^2 \sin(12x - 4)$$

We begin with 4 sample points, then we successively add 6 additional sample points.

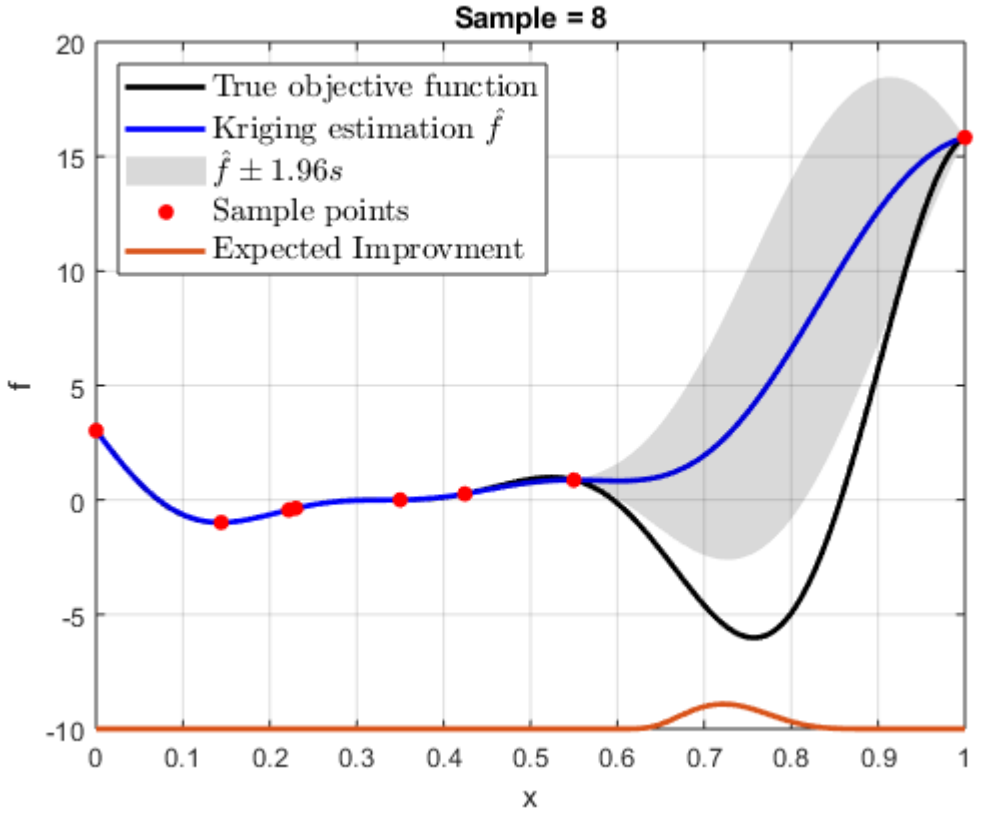


EXAMPLE

Here is an example of the optimization process for a 1D test function using the EGO algorithm.

$$f(x) = (6x - 2)^2 \sin(12x - 4)$$

We begin with 4 sample points, then we successively add 6 additional sample points.

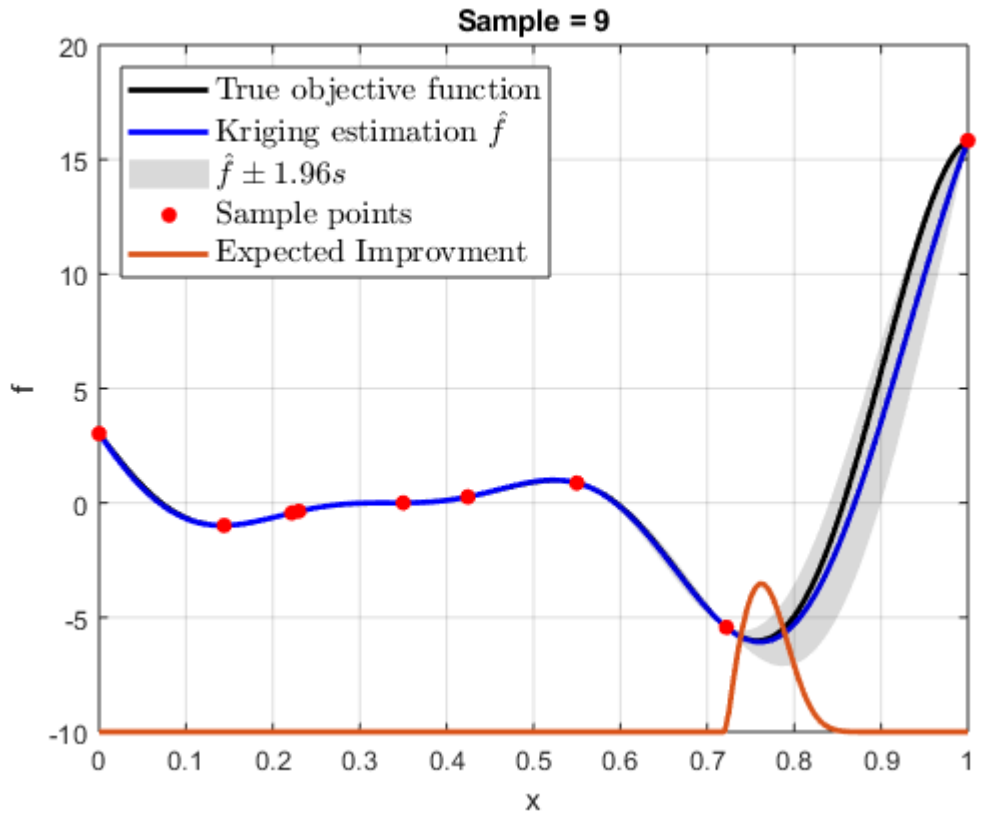


EXAMPLE

Here is an example of the optimization process for a 1D test function using the EGO algorithm.

$$f(x) = (6x - 2)^2 \sin(12x - 4)$$

We begin with 4 sample points, then we successively add 6 additional sample points.

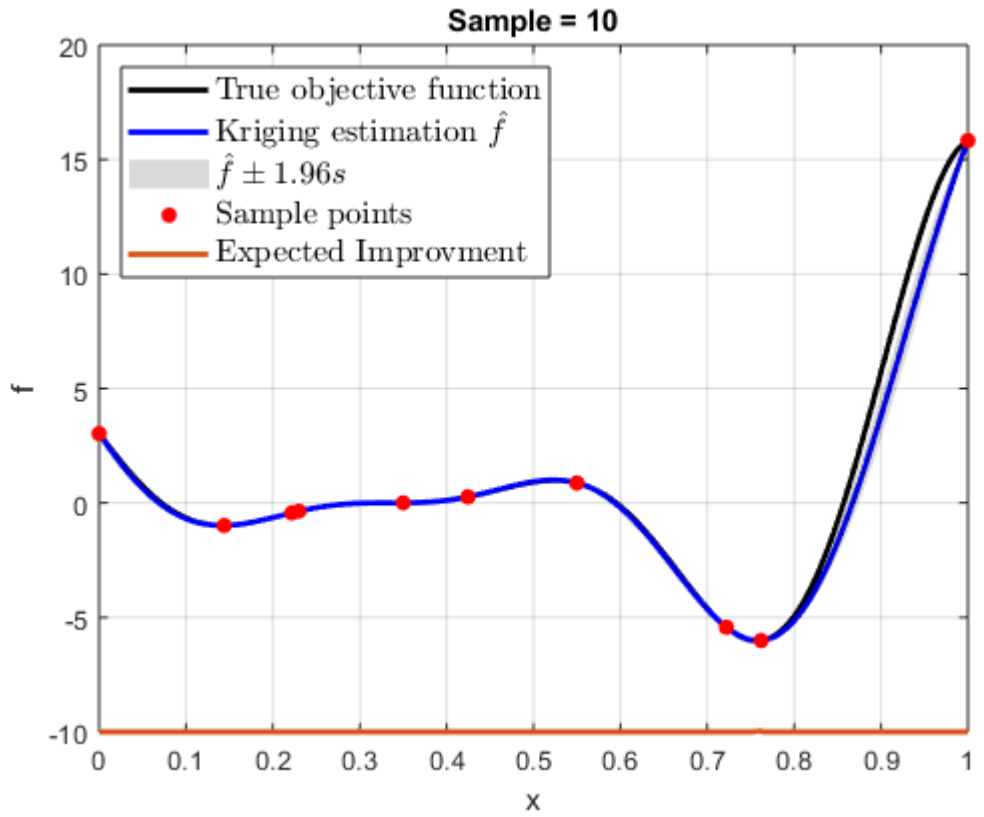


EXAMPLE

Here is an example of the optimization process for a 1D test function using the EGO algorithm.

$$f(x) = (6x - 2)^2 \sin(12x - 4)$$

We begin with 4 sample points, then we successively add 6 additional sample points.



OUTLINE

1) Context
   - Design optimization
   - Gaussian Process regression
   - Bayesian Optimization

2) **Issues in high-dimension**

3) Combination of Kriging models with random length-scales

4) Numerical results

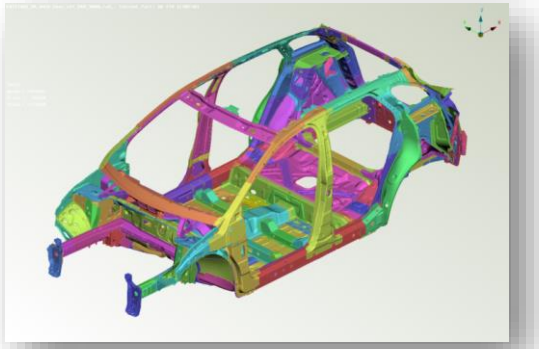
ISSUES IN HIGH DIMENSION

- The main issue is the **optimization of the hyperparameters**.

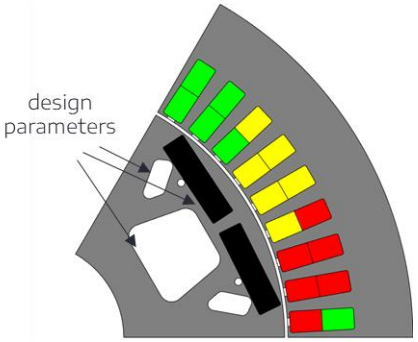
There is one length-scale hyperparameter per dimension, and all these hyperparameters need to be optimized.
→ The optimization of the hyperparameters is difficult :

- d-dimensional problem (with $d > 20$ up to $\approx 100 - 150$).

$d = 130$



$d = 37$



- **The optimization can be costly** due to the cost of the cost for the evaluation of the objective (log-likelihood) and its gradient is in $O(n^3)$.

- When the training data is sparse, **the likelihood criterion can lead to a bad estimation of the hyperparameters.**

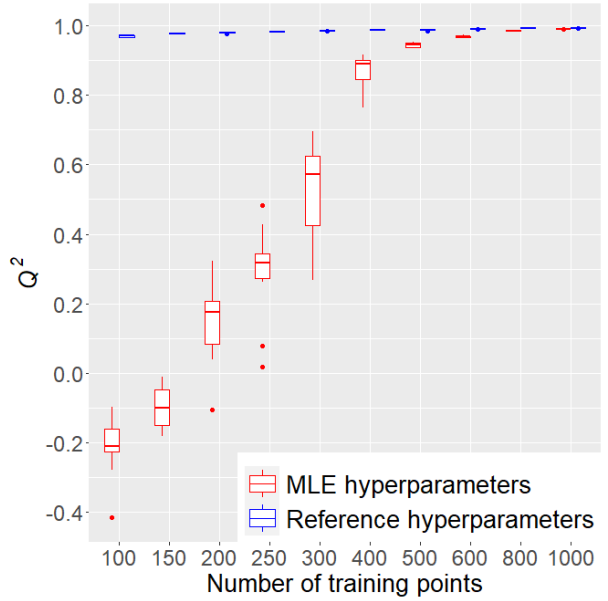
ILLUSTRATION ON AN EXAMPLE

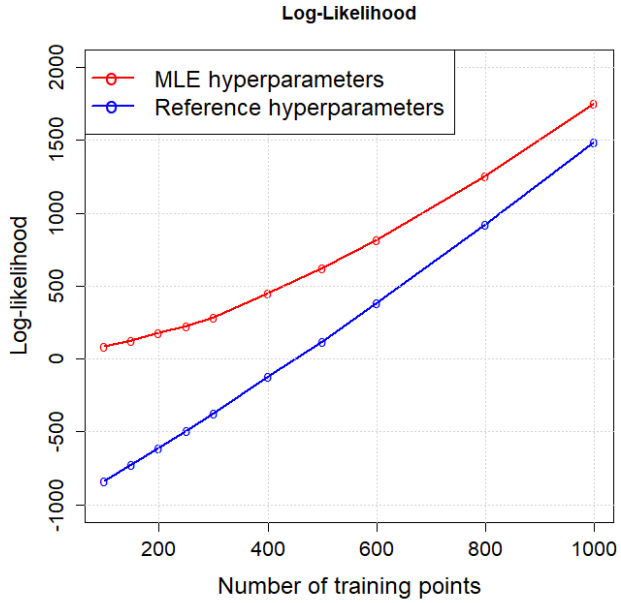
- An illustration of this difficulty: approximating the 50D sphere function:

$$f_{sphère}(x_1, \dots, x_d) = \sqrt{\sum_{i=1}^{d} (x_i - 0.5)^2}, \qquad 0 \leq x_i \leq 1.$$

We fit a Kriging model with MLE hyperparameters using a varying number of training points and compare to a Kriging model with reference hyperparameters :

- 500 iterations for the hyperparameter optimization using the DiceKriging package in R.
- The reference hyperparameters are obtained by doing the optimization with 5000 points.
- The boxplots give the results for 10 different runs.





ISSUES IN HIGH DIMENSION

- Several methods have been proposed to solve this issue:

- Reduction of the problem's dimension by embedding the design space into a lower-dimension space (see for example Constantine et al., 2015, Bouhlel et al., 2016).
- Additive Kriging where the function is assumed to be a sum of one-dimensional components (see for example Durrande et al., 2012).
- Penalized version of the likelihood to improve the robustness of the hyperparameter optimization (see for example RobustGaSP in Gu et al., 2018).
- …

→ We proposed a method to **bypass the hyperparameter optimization** by combining Kriging sub-models with fixed length-scales.

This method is both:
- **Fast** since it avoids the expensive hyperparameter optimization,
- **Easily generalizable** since it does not assume a particular form of the underlying function.

OUTLINE

1) Context
   - Design optimization
   - Gaussian Process regression
   - Bayesian Optimization

2) Issues in high-dimension

3) **Combination of Kriging models with random length-scales**
   - Sampling the random length-scales
   - Weights of the combination
   - Variance of the combination

4) Numerical results

COMBINATION OF KRIGING SUB-MODELS

→ We propose a model which is a combination of Kriging models with random length-scales

$$M_{tot}(\boldsymbol{x}) = \sum_{i=1}^{p} w_i(\boldsymbol{x}) M_i(\boldsymbol{x}) \,,$$

with $M_i(x) = \mu_i + k_{\boldsymbol{\theta}_i}(\boldsymbol{x}, \boldsymbol{X}) \boldsymbol{K}_{\boldsymbol{\theta}_i}^{-1}(\boldsymbol{Y} - \mu_i)$ Kriging model with fixed length−scale vector $\boldsymbol{\theta}_i$.

COMBINATION OF KRIGING SUB-MODELS

→ We propose a model which is a combination of Kriging models with random length-scales

Choice of the sub-models

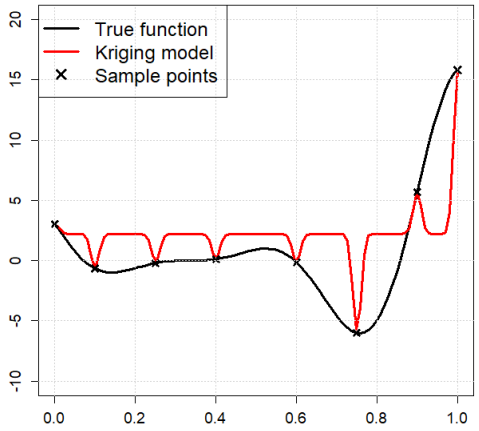
$$M_{tot}(\boldsymbol{x}) = \sum_{i=1}^{p} w_i(\boldsymbol{x}) M_i(\boldsymbol{x}),$$

with $M_i(x) = \mu_i + k_{\boldsymbol{\theta}_i}(\boldsymbol{x}, \boldsymbol{X}) K_{\boldsymbol{\theta}_i}^{-1}(\boldsymbol{Y} - \mu_i)$ Kriging model with fixed length−scale vector $\boldsymbol{\theta}_i$.

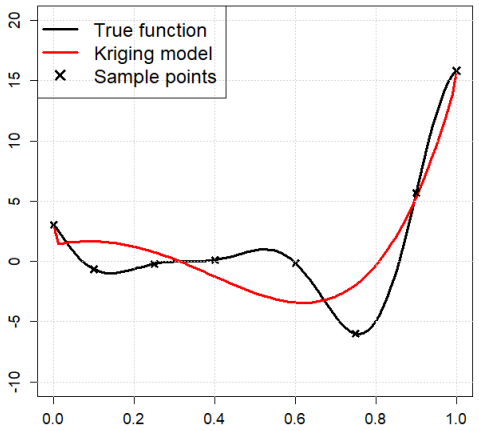
- We want to sample the length-scales in **a range of appropriate values** to avoid degenerate cases.

- For too small values: $k_\theta(x_i, x_j) \longrightarrow 0$ for all $i \neq j$, and $\boldsymbol{K}_\theta \longrightarrow \sigma^2 \boldsymbol{I}_n$.
- For too large values: $k_\theta(x_i, x_j) \longrightarrow 1$, and $\boldsymbol{K}_\theta \longrightarrow \boldsymbol{1}_{n \times n}$.

Small θ



Large θ (+nugget)



We sample the length-scales using an entropy-based criterion.

First, we study the case for a Gaussian correlation where analytical expressions can be obtained.

- Assume that design points are distributed as a random vector $\boldsymbol{X} = \left(X^{(1)}, \dots, X^{(d)}\right)$ with i.i.d components with common variance σ_X^2 and kurtosis κ_X.

- We note D^2 the random square distance between two independent points $\boldsymbol{X}$ and $\boldsymbol{X}'$ of the design. For a large enough dimension:

$$D^2 = \sum_{k=1}^{d} (X_k - X_k')^2 \sim \mathcal{N}\left(2d\sigma_X^2, 2d\sigma_X^4(\kappa_X + 1)\right).$$

- For a Gaussian correlation:

$$R_\theta = e^{-\frac{1D^2}{2\theta^2}} \sim \log\mathcal{N}\left(\frac{-\sigma_X^2}{\theta^2}d, \frac{\sigma_X^4}{2\theta^4}(\kappa_X + 1)d\right).$$

- We can finally obtain the entropy of the correlation:

$$H(R_\theta) = \mathbf{E}\left(-\log f_{R_\theta}(R_\theta)\right) = -\frac{\sigma_X^2}{\theta^2}d + \frac{1}{2}\ln\left(\frac{\sigma_X^4}{2\theta^4}d(\kappa_X + 1)2\pi\right) + \frac{1}{2}.$$

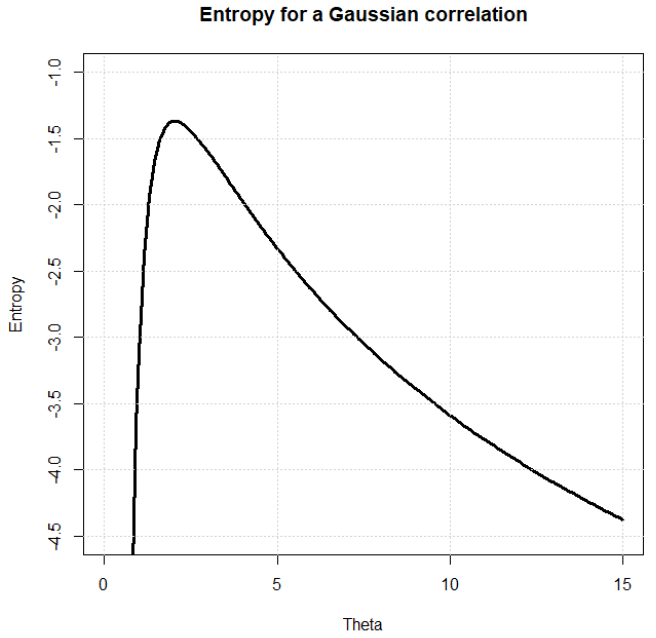
$$H(R_\theta) = \mathbf{E}\big(-\log f_{R_\theta}(R_\theta)\big) = -\frac{\sigma_X^2}{\theta^2}d + \frac{1}{2}\ln\left(\frac{\sigma_X^4}{2\theta^4}d(\kappa_X+1)2\pi\right) + \frac{1}{2}.$$

How to use the knowledge about this entropy ?

- When sampling the length-scales, we want to favor θ corresponding to high entropy values, which result in a high variability in the correlation.

- Finally, we will sample the length-scales using a positive transformation of the entropy:

$$f(\theta) \propto \exp\big(H(R_\theta)\big).$$



Entropy of a Gaussian correlation in 50D for a uniform design ($\sigma_X^2 = 1/12$ and $\kappa_X = 9/5$).

COMBINATION OF KRIGING SUB-MODELS

→ We propose a model which is a combination of Kriging models with random length-scales

Choice of the sub-models

$$M_{tot}(\boldsymbol{x}) = \sum_{i=1}^{p} w_i(\boldsymbol{x}) M_i(\boldsymbol{x}),$$

with $M_i(x) = \mu_i + k_{\boldsymbol{\theta}_i}(\boldsymbol{x}, \boldsymbol{X}) \boldsymbol{K}_{\boldsymbol{\theta}_i}^{-1}(\boldsymbol{Y} - \mu_i)$ Kriging model with fixed length−scale vector $\boldsymbol{\theta}_i$.

COMBINATION OF KRIGING SUB-MODELS

→ We propose a model which is a combination of Kriging models with random length-scales

Choice of the sub-models

Choice of the weights

$$M_{tot}(\boldsymbol{x}) = \sum_{i=1}^{p} w_i(\boldsymbol{x}) M_i(\boldsymbol{x}),$$

with $M_i(x) = \mu_i + k_{\boldsymbol{\theta}_i}(\boldsymbol{x}, \boldsymbol{X}) \boldsymbol{K}_{\boldsymbol{\theta}_i}^{-1} (\boldsymbol{Y} - \mu_i)$ Kriging model with fixed length−scale vector $\boldsymbol{\theta}_i$.

- One method uses constant weights obtained by minimizing the LOOCV error of the combination (see Viana et al., 2009) :

$$e_{LOOCV}(M_{tot}) = \frac{1}{n} \sum_{k=1}^{n} \left(\sum_{i=1}^{p} w_i M_{i_{-k}}(\boldsymbol{x}_k) - y(\boldsymbol{x}_k) \right)^2 = \boldsymbol{w}^T \boldsymbol{C} \boldsymbol{w}.$$

→ The components of the matrix $\boldsymbol{C}$ are : $c_{ij} = \frac{1}{N} e_i^T e_j$, with $e_i^{(k)} = \left[\boldsymbol{K}_{\boldsymbol{\theta}_i}^{-1} \boldsymbol{Y} \right]_k / \left[\boldsymbol{K}_{\boldsymbol{\theta}_i}^{-1} \right]_{k,k}$, $k = 1, \dots, n$.

The weights are then obtained by :

$$\boldsymbol{w}_{LOOCV} = \arg\min_{w} \boldsymbol{w}^T \boldsymbol{C} \boldsymbol{w}, \qquad \text{subject to } \boldsymbol{1}^T \boldsymbol{w} = 1 \qquad \Longrightarrow \boldsymbol{w}_{LOOCV} = \frac{\boldsymbol{1}^T \boldsymbol{C}^{-1}}{\boldsymbol{1}^T \boldsymbol{C}^{-1} \boldsymbol{1}}.$$

(See Appriou et al., 2022 for more details and comparison with other methods).

COMBINATION OF KRIGING SUB-MODELS

→ We propose a model which is a combination of Kriging models with random length-scales

Choice of the sub-models

Choice of the weights

$$M_{tot}(\boldsymbol{x}) = \sum_{i=1}^{p} w_i(\boldsymbol{x}) M_i(\boldsymbol{x}),$$

with $M_i(x) = \mu_i + k_{\boldsymbol{\theta}_i}(\boldsymbol{x}, \boldsymbol{X}) \boldsymbol{K}_{\boldsymbol{\theta}_i}^{-1}(\boldsymbol{Y} - \mu_i)$ Kriging model with fixed length−scale vector $\boldsymbol{\theta}_i$.

Variance of the combination

- Kriging models naturally provides a measure of the model error. For a Kriging model with $Y(.) \sim \mathcal{GP}\left(\mu, k_{\sigma,\boldsymbol{\theta}}(.,.)\right)$:

$$\mathbf{E}\left(\left(M(x) - Y(x)\right)^2\right) = \mathbf{Var}(Y(\boldsymbol{x})|Y(X)) = k(\boldsymbol{x}, \boldsymbol{x}) - k(\boldsymbol{x}, \boldsymbol{X}) \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X})^{-1} k(\boldsymbol{X}, \boldsymbol{x})$$

→ **This prediction error is essential** when performing Bayesian optimization.

We can obtain the prediction error for every individual sub-model, but **the covariance structure between the sub-models is unknown**.

→ We cannot directly access the prediction error of the combination.

VARIANCE OF THE COMBINATION

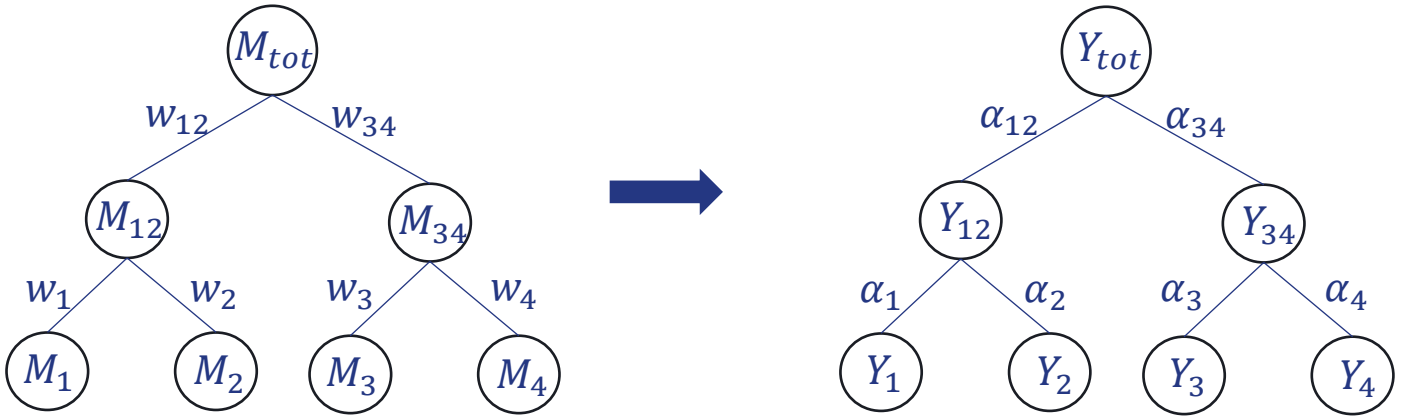
- To obtain the variance of the combination, we add the hypothesis that **the underlying Gaussian Process Y_{tot} is a combination (with different weights) of independent Gaussian Processes**:

$$Y_{tot} = \sigma_{tot}^2 \sum_{i=1}^{p} \alpha_i Y_i \,, \qquad \text{with } Y_i \sim \mathcal{GP}\left(\mu_i, r_{\boldsymbol{\theta}_i}(.,.)\right), \qquad \sum_{i=1}^{p} \alpha_i = 1, \qquad \text{and } \sigma_{tot}^2 \text{ the variance of the GP.}$$

Thus, the covariance of this GP is:

$$k_{tot}(.,.) = \sigma_{tot}^2 \sum_{i=1}^{p} \alpha_i^2 r_{\boldsymbol{\theta}_i}(.,.)\,.$$

- To simplify the upcoming expressions, we will also assume that the sub-models (and the associated GPs) are combined following a binary tree structure:



VARIANCE OF THE COMBINATION

- The weights α in the combination of GPs are chosen to **minimize the expected mean-square error of the combined model** with respect to $Y_{tot} = \alpha Y_1 + (1-\alpha)Y_2$:

$$\alpha^* = \arg\min_{\alpha} \mathbf{E}\left[\mathbf{E}\left[\left(wM_1(\boldsymbol{x}) + (1-w)M_2(\boldsymbol{x}) \; - \alpha Y_1(\boldsymbol{x}) + (1-\alpha)Y_2(\boldsymbol{x})\right)^2 \mid Y_1, Y_2\right]\right].$$

By approximation the global MSE using the LOOCV error, we obtain:

$$\alpha^* = \frac{a_1(w)}{a_1(w) + a_2(w)}, \qquad \text{with: } \begin{cases} a_1(w) = w^2\mathbf{E}(e_{LOOCV}(M_1)|Y_2) + (1-w^2)\mathbf{E}(e_{LOOCV}(M_2)|Y_2), \\ a_2(w) = (1-w)^2\mathbf{E}(e_{LOOCV}(M_2)|Y_1) + (1-(1-w)^2)\mathbf{E}(e_{LOOCV}(M_1)|Y_1). \end{cases}$$

Finally, the variance of the combination is obtained as:

$$\hat{s}^2(\boldsymbol{x}) = \mathbf{Var}(Y_{tot}(\boldsymbol{x})|\mathcal{D}) = k_{tot}(\boldsymbol{x},\boldsymbol{x}) - k_{tot}(\boldsymbol{x},\boldsymbol{X})\boldsymbol{K}_{tot}(\boldsymbol{X},\boldsymbol{X})^{-1}k_{tot}(\boldsymbol{X},\boldsymbol{x}).$$

Relation to other methods:

- Relation to additive models: as we use an additive structure for obtaining the variance, why not use it for the prediction as well ?

$$\widetilde{M}(x) = \mathbf{E}(Y_{tot}(\boldsymbol{x})|\mathcal{D}) = \mu_{tot} + k_{tot}(\boldsymbol{x}, \boldsymbol{X})\boldsymbol{K}_{tot}(\boldsymbol{X}, \boldsymbol{X})^{-1}(\boldsymbol{Y} - \mu_{tot}).$$

→ $\boldsymbol{K}_{tot}(\boldsymbol{X}, \boldsymbol{X})^{-1}$ is the inverse of a sum of matrices and **there is no direct formula for the inverse of a sum of matrices**.

↳ Estimating the weights will **involve a large number of matrix inversions and an inner optimization** which we aim to avoid with our method.

- Relation to mixture models: a mixture of GPs will give the same mean prediction as the linear combination, and we can directly obtain the variance of the mixture.

→ There is a relation between the MSE of both models:

$$\mathbf{E}\left[\left(M_{mix}(x) - Y_{mix}(x)\right)^2\right] = \mathbf{E}\left[\left(M_{tot}(x) - Y_{tot}(x)\right)^2\right], \qquad \text{when } k_{tot}(.,.) = \sum_{i=1}^{p} w_i k_{\theta_i}(.,.)\,.$$

→ By tuning the weights $\boldsymbol{\alpha}$, we achieve **better calibrated confidence intervals** than a mixture model.

OUTLINE

1) **Context**
   - Design optimization
   - Gaussian Process regression
   - Bayesian Optimization

2) **Issues in high-dimension**

3) **Combination of Kriging models with random length-scales**
   - Sampling the random length-scales
   - Weights of the combination
   - Variance of the combination

4) **Numerical results**

NUMERICAL RESULTS

We test the combination for high-dimensional Bayesian optimization and compare it to ordinary Kriging models using MLE:

- Number of initial samples : $2 \times d$,
- Number of iterations : $10 \times d$,
- MLE optimization of the hyperparameters at each iteration:
  - R package DiceKriging (L-BFGS-B, 300 max iterations),
- Number of sub-models : $p = 16$,
- Optimization of the EI with the package DiceOptim along with TREGO trust regions (see Diouane et al., 2023),
- 10 optimization runs with different initializations.

NUMERICAL RESULTS – TEST FUNCTIONS

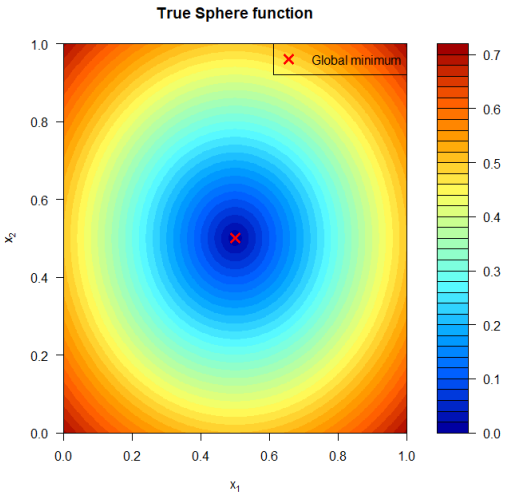
In this section, we consider two test functions (with varying dimensions) for the optimization:

- The sphere function:

$$f_{sphère}(x_1, \dots, x_d) = \sqrt{\sum_{i=1}^{d} (x_i - 0{,}5)^2}\,, \qquad 0 \leq x_i \leq 1.$$

→ Deceptively difficult to model with Gaussian Processes with few observations.

→ Easy to optimize (convex function).



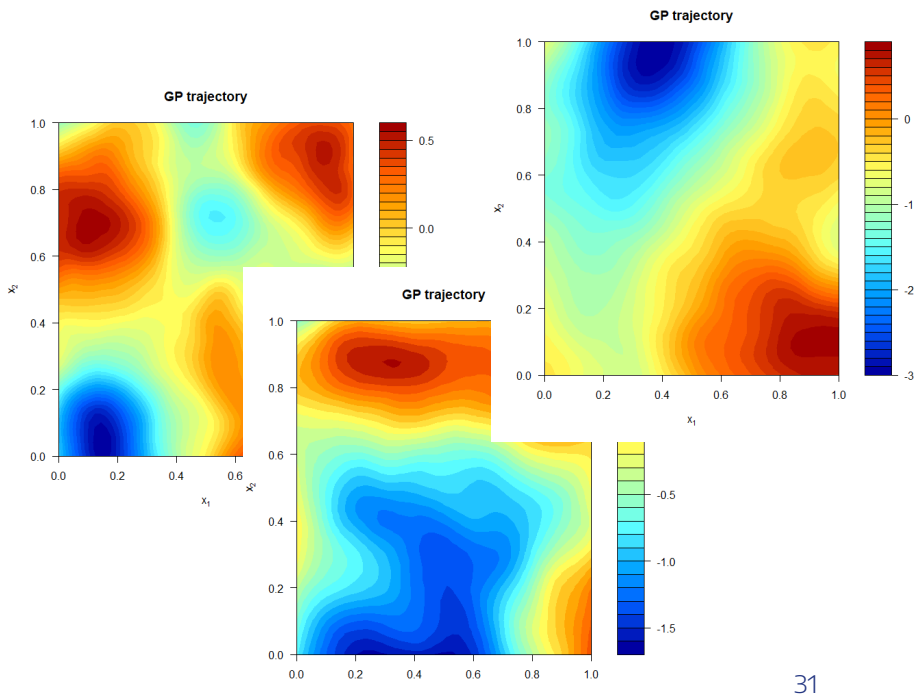
- GP trajectory:

$$f_{GP}(.) \sim GP\big(\mathbf{0}, k_\theta(.,.)\big),$$

Where k_θ is an isotropic Matérn 5/2 correlation with length-scale $\theta = \sqrt{\frac{d}{12}}$.

→ Harder to optimize (multimodal) and more representative of true functions.

→ Case where the Kriging hypothesis is verified.

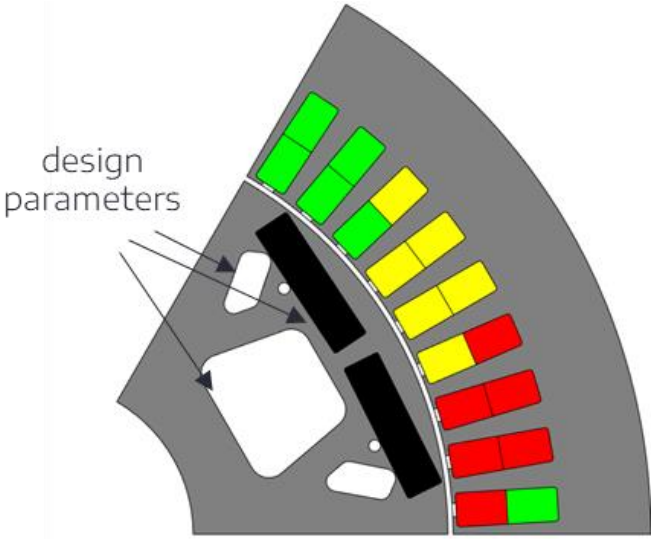


NUMERICAL RESULTS – TEST FUNCTIONS

- Design of an electrical machine:

- 37 design variables:

Position and size of air holes and magnets, radius of the machine.



Full problem:

- 2 objectives:

Consumption and cost of the machine.

- 10 constraints:

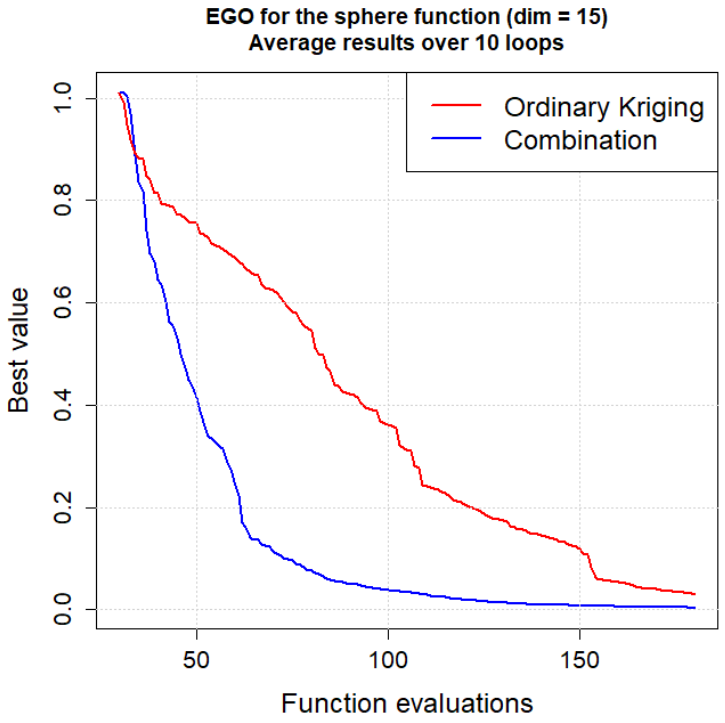
Related to the dynamics of the vehicle and to the dynamics of the machine.

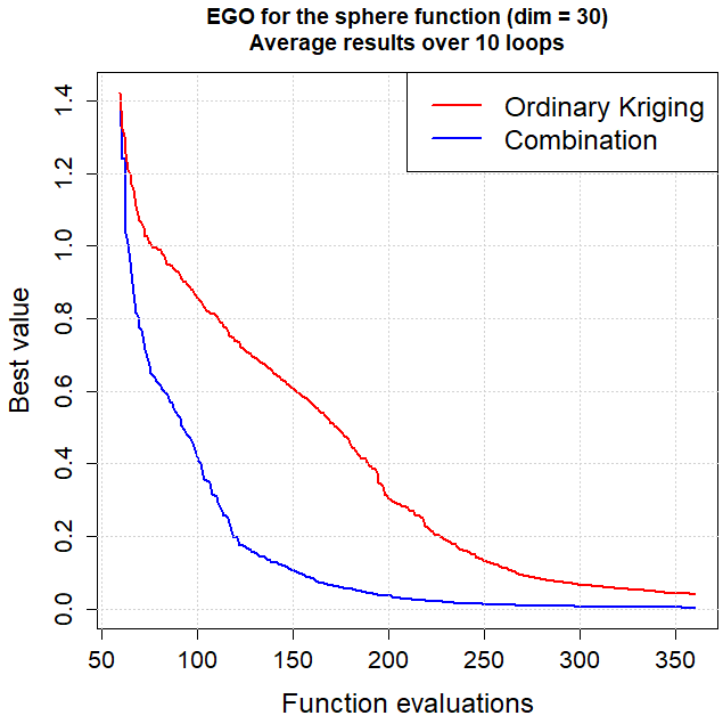
→ Here we test the method only for single objective optimization.

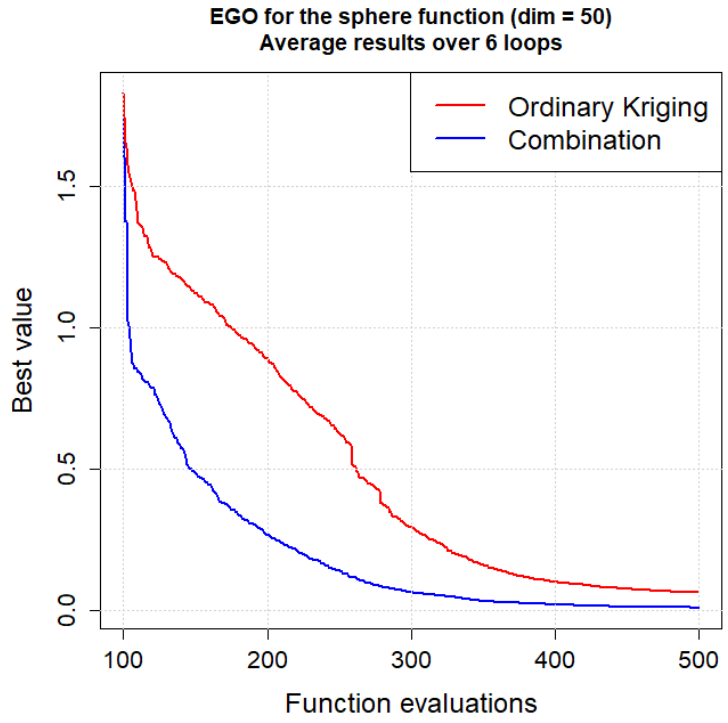
→ We only optimize the first constraint (maximum speed of the car).

NUMERICAL RESULTS – SPHERE FUNCTION

- Sphere function



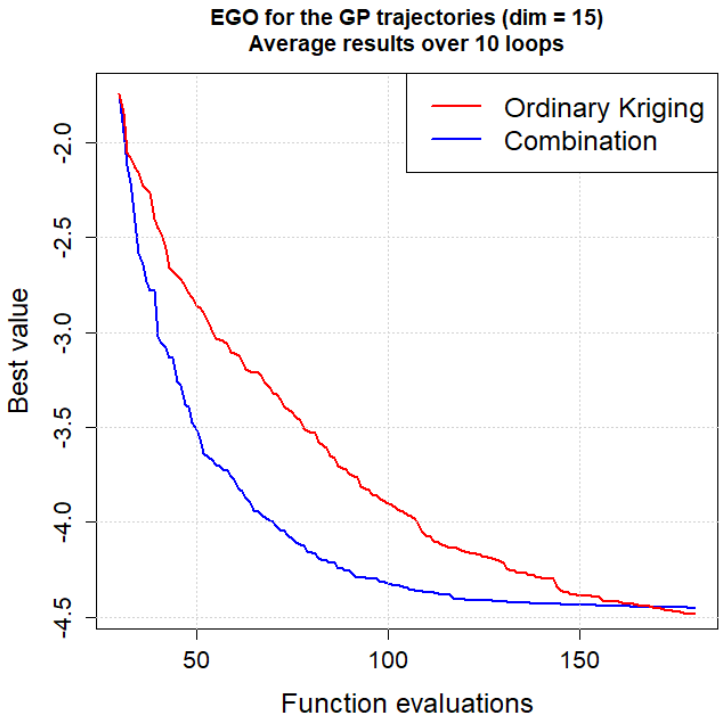


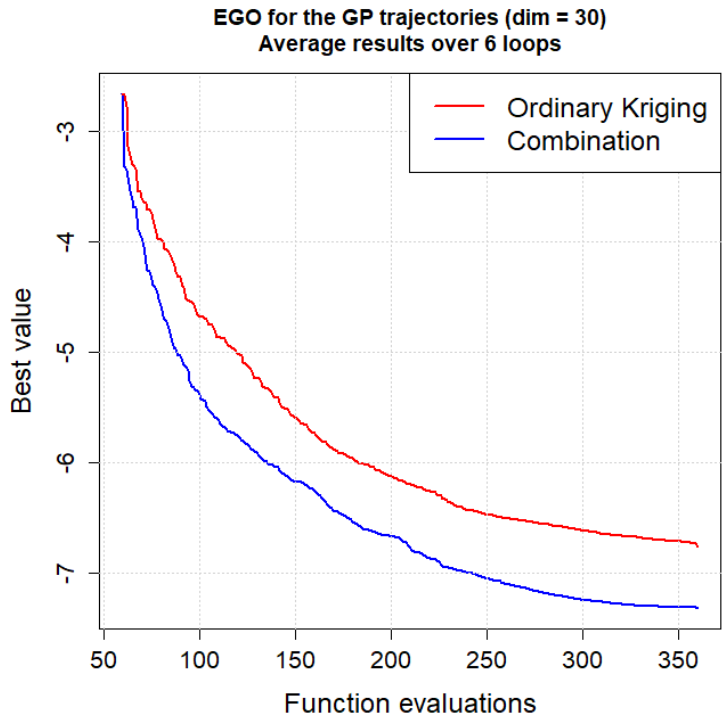


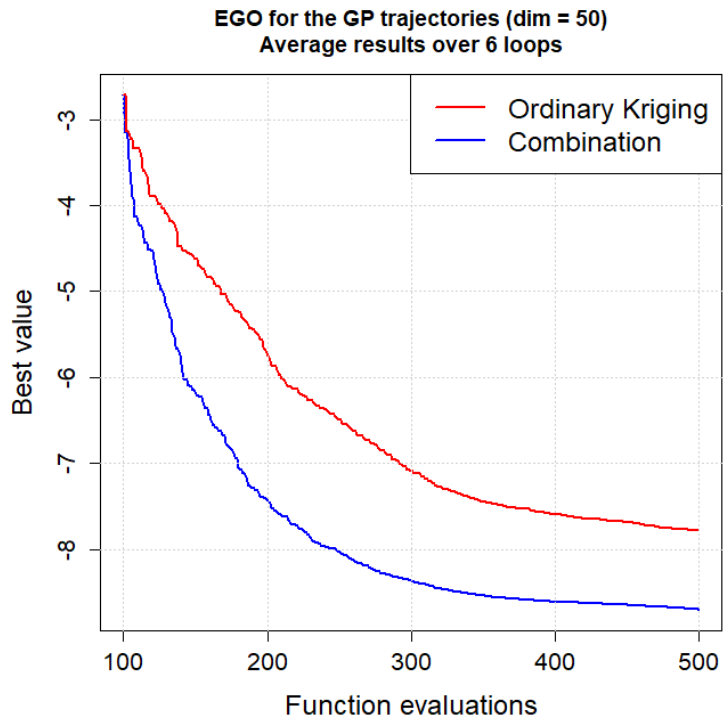
→ The combination converges faster at the beginning of the optimization (few points) because the models are more accurate.

NUMERICAL RESULTS – GP TRAJECTORY

- GP trajectories





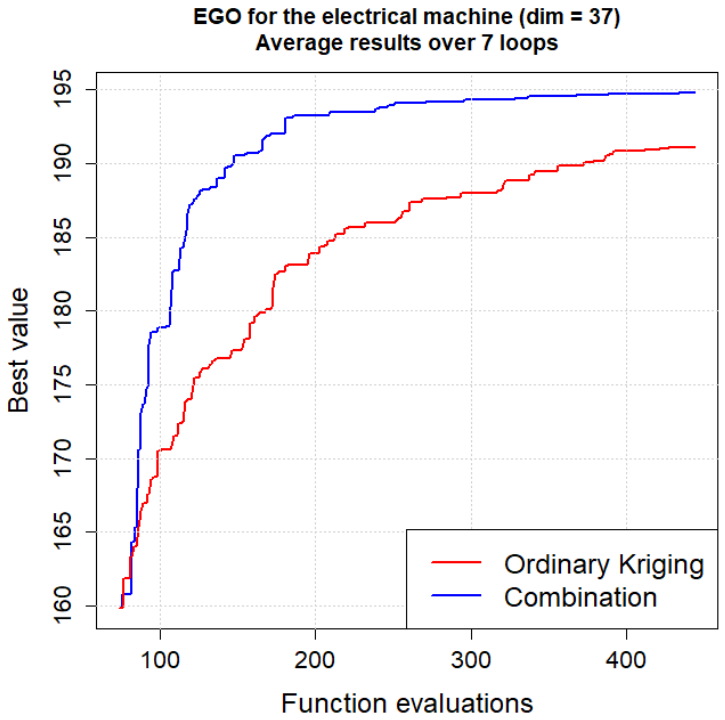


→ The combination still converges faster.

→ For the multi-modal GP trajectories, the combination converges to a better optimum.

NUMERICAL RESULTS – ELECTRICAL MACHINE

- Electrical machine



→ Similar results to the optimization of GP trajectories.

PERSPECTIVES

- The combination of Kriging models shows promising results on the test functions and **outperforms the ordinary Kriging especially at the start of the optimization**.

- **A benchmark against other high-dimensional optimization methods** such that additive models or dimension reduction techniques still need to be conducted.

- The method was tested on an **industrial test cases** only for single objective optimization. Tests on **multi-objective problems with constraints** can be conducted by adapting the acquisition strategy (for example EHVI).

For more details, see our preprint: https://hal.science/hal-04477236

Thank you for your attention !

Contact :

Tanguy APPRIOU
tanguy.appriou@stellantis.com

REFERENCES

- Appriou, T., Rullière, D. and Gaudrie, D., 2022. Combination of High-Dimensional Kriging Sub-models.
- Appriou, T., Rullière, D. and Gaudrie, D., 2024. High-Dimensional Bayesian Optimization with a Combination of Kriging models.
- Bouhlel, M.A., Bartoli, N., Otsmane, A. and Morlier, J., 2016. Improving kriging surrogates of high-dimensional design models by Partial Least Squares dimension reduction. *Structural and Multidisciplinary Optimization*, *53*(5), pp.935-952.
- Constantine, P.G., Dow, E. and Wang, Q., 2014. Active subspace methods in theory and practice: applications to kriging surfaces. *SIAM Journal on Scientific Computing*, *36*(4), pp.A1500-A1524.
- Diouane, Y., Picheny, V., Riche, R.L. and Perrotolo, A.S.D., 2023. TREGO: a trust-region framework for efficient global optimization. *Journal of Global Optimization*, *86*(1), pp.1-23.
- Durrande, N., Ginsbourger, D. and Roustant, O., 2012. Additive covariance kernels for high-dimensional Gaussian process modeling. In *Annales de la Faculté des sciences de Toulouse: Mathématiques* (Vol. 21, No. 3, pp. 481-499).
- Gu, M., Palomo, J. and Berger, J.O., 2018. RobustGaSP: Robust Gaussian stochastic process emulation in R. *arXiv preprint arXiv:1801.01874*.
- Gu, M., Wang, X. and Berger, J.O., 2018. Robust Gaussian stochastic process emulation. *The Annals of Statistics*, *46*(6A), pp.3038-3066.

REFERENCES

- Jones, D.R., Schonlau, M. and Welch, W.J., 1998. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, *13*, pp.455-492.

- Roustant, O., Ginsbourger, D. and Deville, Y., 2012. DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization. *Journal of statistical software*, *51*, pp.1-55.

- Rasmussen, C.E. and Williams, C.K., 2006. *Gaussian processes for machine learning*. Cambridge, MA: MIT press.

- Viana, F.A., Haftka, R.T. and Steffen, V., 2009. Multiple surrogates: how cross-validation errors can help us to obtain the best predictor. *Structural and Multidisciplinary Optimization*, *39*(4), pp.439-457.

In practice, for any correlation function R_θ and any design plan $\boldsymbol{X}$.

1. For a given length-scale θ. We sample N values of the correlation for the design plan $\boldsymbol{X}$: $r_\theta^{(1)}, \dots r_\theta^{(N)}$.

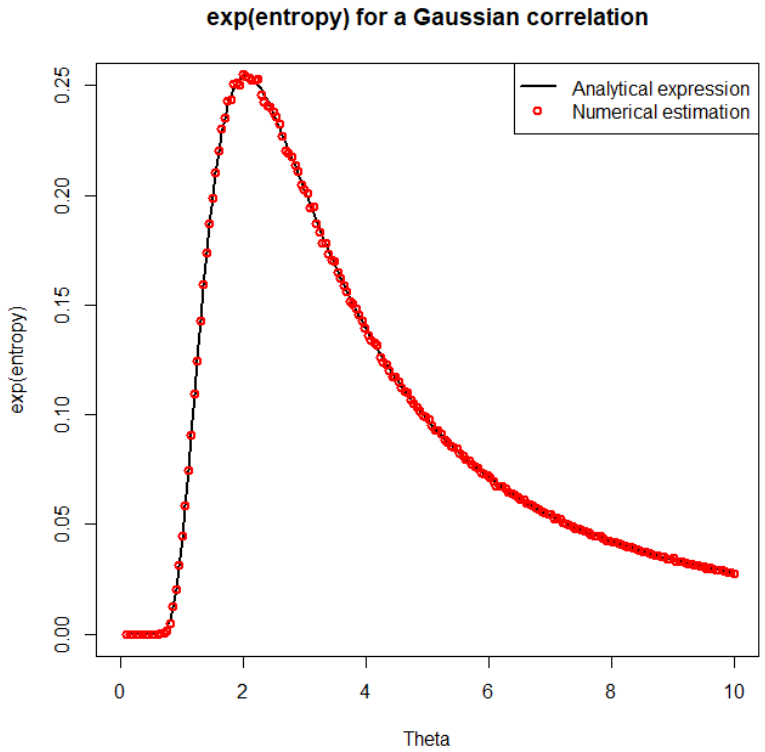
2. We make a kernel estimation $\hat{f}_{R_\theta}$ of the density of R_θ based on these samples.

3. We compute the empirical entropy:

$$\widehat{H}(R_\theta) = -\frac{1}{n}\sum_{i=1}^{n} \ln \hat{f}_{R_\theta}\left(r_\theta^{(i)}\right).$$

4. We define a grid of possible values for the length-scales $\theta_{grid}^{(\ell)},\ \ell = 1, \dots, q$, and we sample with probability:

$$P\left(\theta_{grid}^{(\ell)}\right) \propto \exp\left(H(R_\theta)\right).$$



→ We sample d length-scale values (one for each dimension) for each of the sub-models.

- Finally, the last step is to **calibrate the amplitude of the variance** using the amplitude hyperparameter σ_{tot}^2.

For LOO strategies, typically this is done by observing that **the normalized LOO errors should be normally distributed if the model is well-specified**:

$$\frac{e_{LOO}}{\sqrt{\sigma_{tot}^2 Var_{LOO}}} \sim \mathcal{N}(0,1).$$

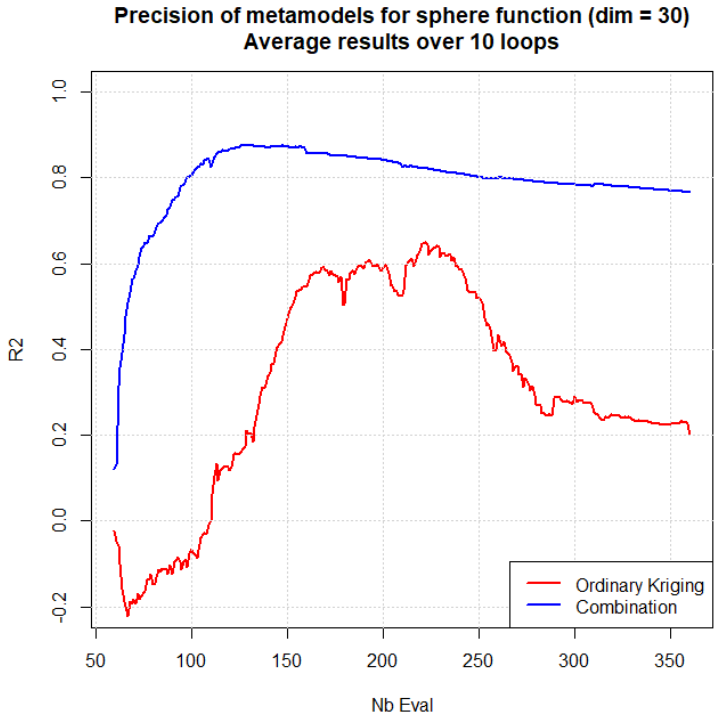
→ Thus, by setting the empirical variance of the normalized residuals to 1: $\sigma_{tot}^2 = \frac{1}{n}\sum_{i=1}^{n} \frac{{e_{LOO}}_i^2}{{Var_{LOO}}_i}$.

However, this definition tends to give too large amplitudes due to the presence of many outliers in the LOO error.

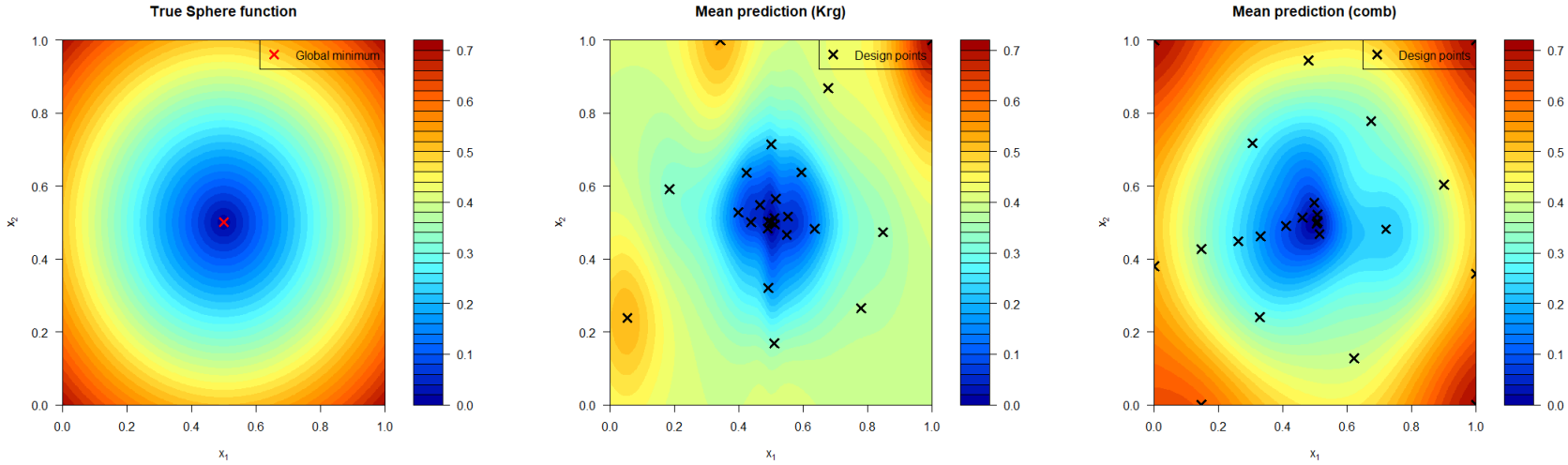
To have an **expression for the amplitude more robust to outliers** and which overall give prediction interval that are better calibrated, we fit **the empirical inter-quartile distance** of the LOO error to that of a Gaussian distribution:

$$IQ\left(\frac{e_{LOO}}{\sigma_{tot}\sqrt{Var_{LOO}}}\right) = IQ_{norm} \iff \sigma_{tot} = \frac{IQ\left(\frac{e_{LOO}}{\sqrt{Var_{LOO}}}\right)}{IQ_{norm}} = \frac{q_{0,75}\left(\frac{e_{LOO}}{\sqrt{Var_{LOO}}}\right) - q_{0,25}\left(\frac{e_{LOO}}{\sqrt{Var_{LOO}}}\right)}{IQ_{norm}}.$$

APPENDICE – RESULTS FOR SPHERE FUNCTION



2D illustration of an EGO optimization for Ordinary Kriging and the Combination (4 initial points and 20 iterations):



→ The Ordinary Kriging needs many iterations to achieve a reasonable precision.

→ The 2D example shows how ordinary Kriging still finds the global minimum with good precision despite a poor global accuracy of the surrogate model.