Multi-objective optimization using Deep Gaussian Processes : Application to Aerospace Vehicle Design

A. Hebbal [1,2], L. Brevault [1], M.Balesdent[1]
E-G. Talbi[2], N. Melab[2]

[1]ONERA - The French Aerospace Lab
[2]Université de Lille, CNRS/CRIStAL, Inria Lille

AIAA SciTech Forum 2019

Table of Contents

Context

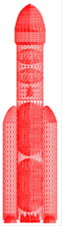
Resolution of an optimization problem characterized by :

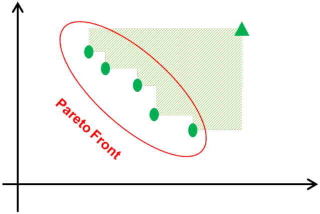
- Multiple objectives to optimize.
- Black box and computationally expensive functions.
- Non-stationary functions.

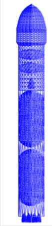
Context

Resolution of an optimization problem characterized by :

- Multiple objectives to optimize.
- Black box and computationally expensive functions.
- Non-stationary functions.



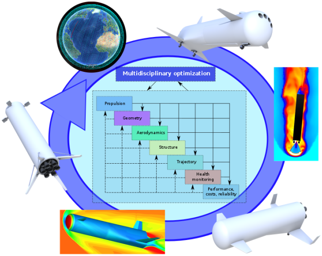




Context

Resolution of an optimization problem characterized by :

- Multiple objectives to optimize.
- Black box and computationally expensive functions.
- Non-stationary functions.



Multi-disciplinary optimization of an aerospace vehicle

Context

Resolution of an optimization problem characterized by :

- Multiple objectives to optimize.
- Black box and computationally expensive functions.
- Non-stationary functions.

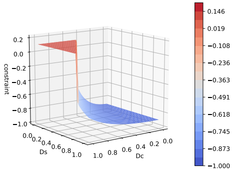
✗ Gradient based optimization approaches	✗ Classic evolutionary algorithms	✓ Bayesian Optimization

Context

Resolution of an optimization problem characterized by :

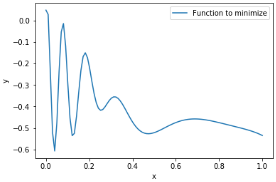
- Multiple objectives to optimize.
- Black box and computationally expensive functions.
- Non-stationary functions.

Abrupt change in the function



Non stationary modelisation of a constraint

Variation of the smoothness along the input space



Non stationary objective function

Context

Resolution of an optimization problem characterized by :

- Multiple objectives to optimize.
- Black box and computationally expensive functions.
- Non-stationary functions.

✗ Gradient based optimization approaches

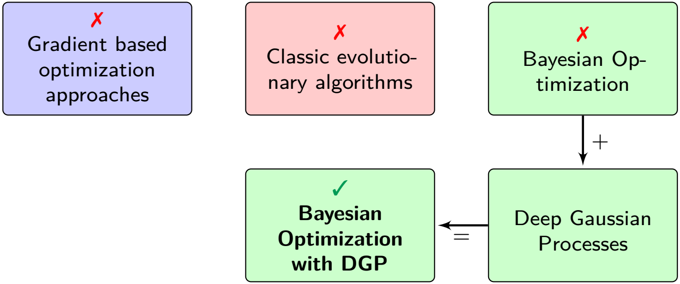
✗ Classic evolutionary algorithms

✗ Bayesian Optimization

Context

Resolution of an optimization problem characterized by :

- Multiple objectives to optimize.
- Black box and computationally expensive functions.
- Non-stationary functions.



[Damianou and Lawrence, 2013]

MO-BO framework

$$(P)\left|\begin{array}{lll} \text{Minimize}_{\mathbf{x}} & \mathbf{y} & = \mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), \ldots, f_n(\mathbf{x})] \\ \text{subject to} & g_i(\mathbf{x}) & \leq 0, \; i = 1, \ldots, n_c \end{array}\right.$$

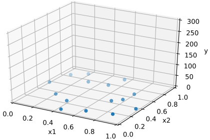
with :

$\mathbf{x} = (x_1, \ldots, x_D) \in \mathbb{X} \subseteq \mathbb{R}^D$ and $\mathbf{y} = (y_1, \ldots, y_n) \in \mathbb{Y} \subseteq \mathbb{R}^n$

MO-BO framework

Multi-Objective Bayesian Optimization (MO-BO) [Emmerich et al., 2006]

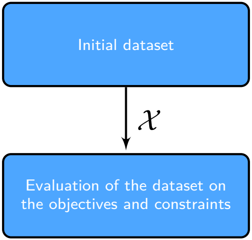
Initial dataset

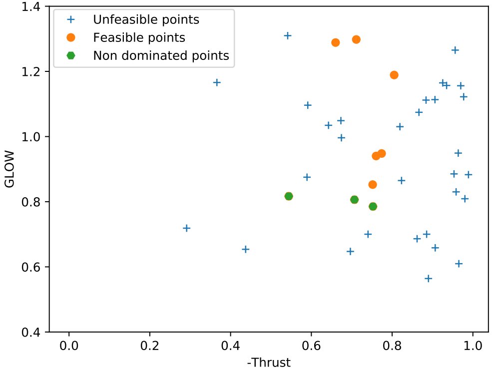


Design of Experiment depending on the dimension and the nature of the problem

MO-BO framework

Multi-Objective Bayesian Optimization (MO-BO) [Emmerich et al., 2006]





Calls the **expensive** black-box functions

MO-BO framework

Multi-Objective Bayesian Optimization (MO-BO) [Emmerich et al., 2006]

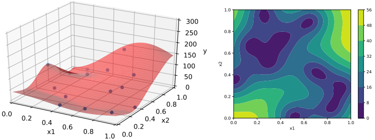
Initial dataset

$\mathcal{X}$

Evaluation of the dataset on the objectives and constraints

$\mathcal{Y}, \mathcal{G}$

Construction of the GP regression models for the objectives and constraints



Learning the hyperparameters of each surrogate model

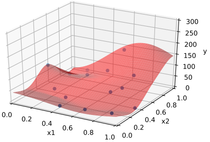
MO-BO framework

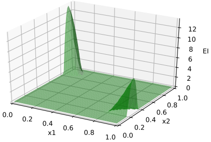
Multi-Objective Bayesian Optimization (MO-BO) [Emmerich et al., 2006]

Initial dataset

$\mathcal{X}$

Evaluation of the dataset on the objectives and constraints





Optimization of a chosen **infill sampling criterion**

$\mathcal{Y}, \mathcal{G}$

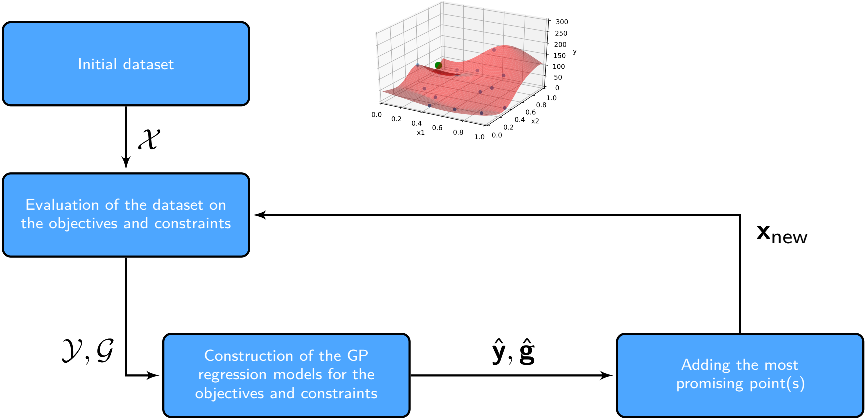
Construction of the GP regression models for the objectives and constraints

$\hat{\mathbf{y}}, \hat{\mathbf{g}}$

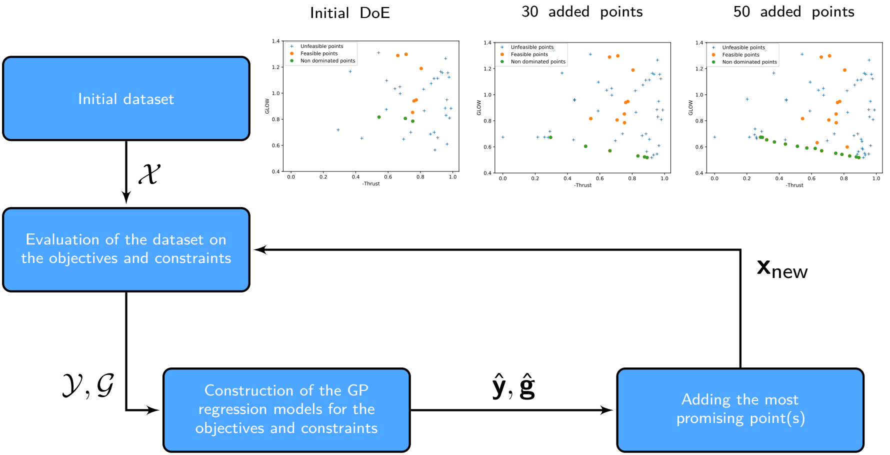
Adding the most promising point(s)

MO-BO framework

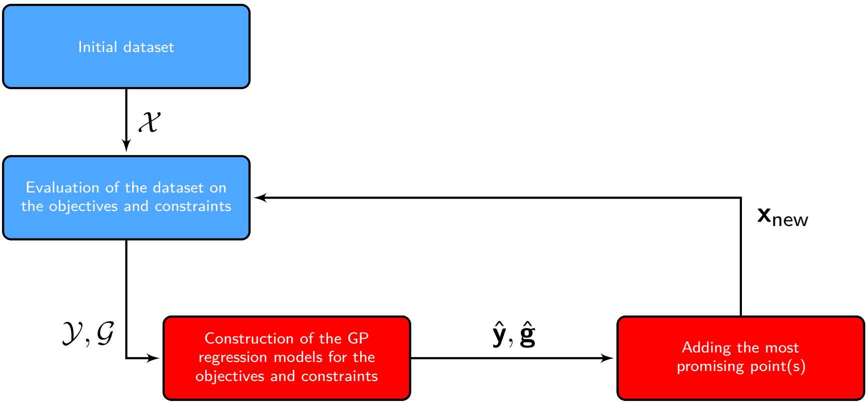
Multi-Objective Bayesian Optimization (MO-BO) [Emmerich et al., 2006]



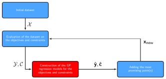
MO-BO framework



MO-BO framework

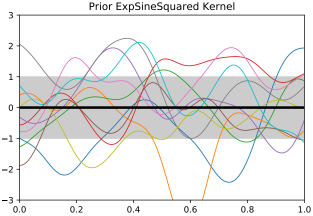


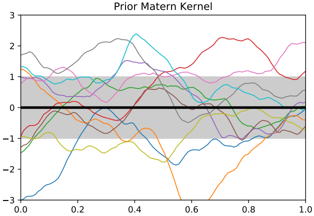
Gaussian Process Regression



Gaussian process [Rasmussen, 2004]

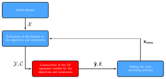
A Gaussian Process is used to describe a distribution over function. It is a collection of infinite random variables, **any finite number of which have a joint Gaussian distribution**.
It is defined by its mean function and covariance function (Kernel) : $f(.) \sim \mathcal{GP}(\mu(.), k^{\Theta}(.))$





ONERA
THE FRENCH AEROSPACE LAB

Gaussian Process Regression



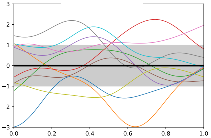
Gaussian process [Rasmussen, 2004]

A Gaussian Process is used to describe a distribution over function. It is a collection of infinite random variables, **any finite number of which have a joint Gaussian distribution**.
It is defined by its mean function and covariance function (Kernel) : $f(.) \sim \mathcal{GP}(\mu(.), k^{\Theta}(.))$

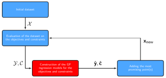
Automatic Relevance Determination (ARD) squared exponential kernel :

$$K^{\Theta}(\mathbf{x}, \mathbf{x'}) = \sigma^2 \exp\left(-\sum_{i=1}^{D} \theta_i . |x_i - x'_i|^2\right)$$

Prior Gaussian Process



Gaussian Process Regression



Gaussian process [Rasmussen, 2004]

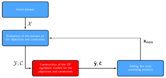
A Gaussian Process is used to describe a distribution over function. It is a collection of infinite random variables, **any finite number of which have a joint Gaussian distribution**.
It is defined by its mean function and covariance function (Kernel) : $f(.) \sim \mathcal{GP}(\mu(.), k^{\Theta}(.))$

Automatic Relevance Determination (ARD) squared exponential kernel :

$$K^{\Theta}(\mathbf{x},\mathbf{x'}) = \sigma^2 \exp\left(-\sum_{i=1}^{D} \theta_i . |x_i - x'_i|^2\right)$$

$$\text{Maximize w.r.t } \Theta \quad : p(\mathbf{y}|\mathcal{X}) = \mathcal{N}(\mathbf{y}|0, \mathbf{K}^{\Theta}_{NN}\mathbf{I})$$

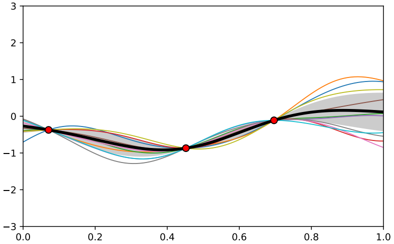
Gaussian Process Regression



Gaussian process [Rasmussen, 2004]

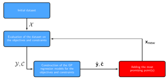
A Gaussian Process is used to describe a distribution over function. It is a collection of infinite random variables, **any finite number of which have a joint Gaussian distribution**.
It is defined by its mean function and covariance function (Kernel) : $f(.) \sim \mathcal{GP}(\mu(.), k^{\Theta}(.))$

Posterior Gaussian Process



$$y^* | \mathbf{x}^*, \mathcal{X}, \mathcal{Y} \sim \mathcal{N}(\hat{y}^*, \hat{s}^*)$$

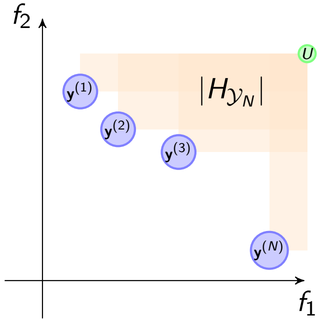
Infill Criteria



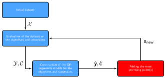
Hypervolume indicator

The hypervolume indicator expresses the hypervolume of the objective space dominated by the approximated Pareto set.

$$H_{\mathcal{Y}_N} = \left\{ \mathbf{y} \in \mathbb{B}; \exists i \in \{1, \dots, N\}, \mathbf{y}^{(i)} \prec \mathbf{y} \right\}$$



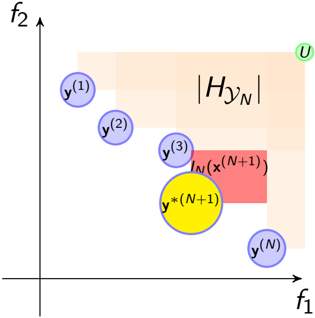
Infill Criteria



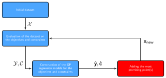
Hypervolume improvement
[Emmerich et al., 2006]

The hypervolume improvement is the improvement of the hypervolume by adding a candidate to the data set

$$I_N(\mathbf{x}^{(N+1)}) = |H\mathcal{Y}_{N+1}| - |H\mathcal{Y}_N|$$



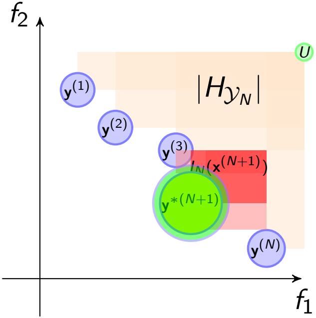
Infill Criteria



Expected Hypervolume improvement [Emmerich et al., 2006]

The expected hypervolume improvement is the mathematical expected improvement of the hypervolume by adding a candidate to the sample

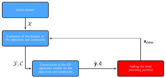
$$EHVI_N(\mathbf{x}) = \mathbb{E}(|H\mathcal{Y}_{N+1}| - |H\mathcal{Y}_N|)$$
$$= \int_{\mathbb{B}\setminus H\mathcal{Y}_N} \mathbb{P}(\mathbf{y}^{*(N+1)} \prec p)\mathrm{d}p$$



with $\mathbf{y}^{*(N+1)} = [y_1^{*(N+1)}, y_2^{*(N+1)}]$

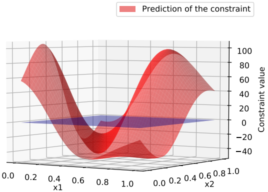
and $y_1^{*(N+1)} \sim \mathcal{N}\left(\hat{y}_1^{*(N+1)}, \hat{s}_1^{*(N+1)}\right)$ and $y_2^{*(N+1)} \sim \mathcal{N}\left(\hat{y}_2^{*(N+1)}, \hat{s}_2^{*(N+1)}\right)$

Infill Criteria



Probability of feasability

The probability of feasability is the probability that all the constraints are satisfied.
$P_f(\mathbf{x}) = \prod_{i=1}^{n_c} \mathbb{P}(g_i^*(\mathbf{x})) < 0$

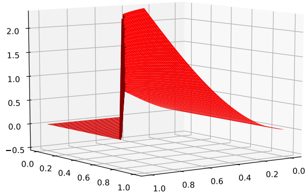


Prediction

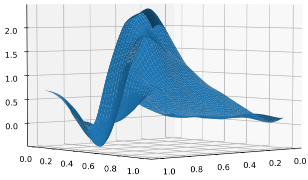
Limits of GPs

Handling non stationarity

- Abrupt change in the response
- Different behavior according to the input



2-D Function with an abrupt change

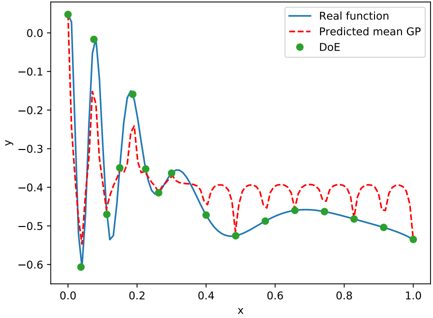


GP Approximation

Limits of GPs

Handling non stationarity

- Abrupt change in the response
- Different behavior according to the input



1-D Function with a different behaviour according to the input space

Review of methods of handling non stationarity

- Direct formulation of non-stationary covariance function [Higdon et al., 1999] [Paciorek and Schervish, 2006]
- Local stationary covariance function [Haas, 1990] [Rasmussen and Ghahramani, 2002]
- Non-linear mapping [Xiong et al., 2007]

- Curse of dimensionality,
- Scarce data.

Review of methods of handling non stationarity

- Direct formulation of non-stationary covariance function [Higdon et al., 1999] [Paciorek and Schervish, 2006]
- Local stationary covariance function [Haas, 1990] [Rasmussen and Ghahramani, 2002]
- Non-linear mapping [Xiong et al., 2007]

- Curse of dimensionality,
- Scarce data.

Deep Gaussian Processes

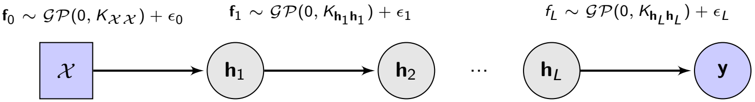
Deep Gaussian Processes [Damianou and Lawrence, 2013]

DGPs are a class of surrogate models based on the structure of neural networks, where each layer is a GP. It considers that the statistical relationship between the inputs and the response is expressed by a functional composition of GPs :

$$y = f_L(\mathbf{f}_{L-1}(\dots \mathbf{f}_1(\mathbf{f}_0(\mathbf{x}) + \epsilon_0) + \epsilon_1) \dots) + \epsilon_{L-1}) + \epsilon_L$$

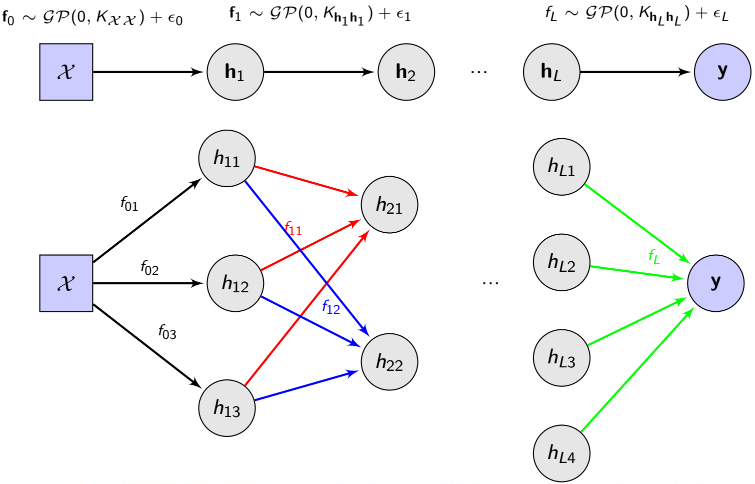
$$\mathcal{X} \xrightarrow{\mathbf{f}_0 \sim \mathcal{GP}(0, K_{\mathcal{X}\mathcal{X}}) + \epsilon_0} \mathbf{h}_1 \xrightarrow{\mathbf{f}_1 \sim \mathcal{GP}(0, K_{\mathbf{h}_1\mathbf{h}_1}) + \epsilon_1} \mathbf{h}_2 \quad \cdots \quad \mathbf{h}_L \xrightarrow{f_L \sim \mathcal{GP}(0, K_{\mathbf{h}_L\mathbf{h}_L}) + \epsilon_L} \mathbf{y}$$

Deep Gaussian Processes

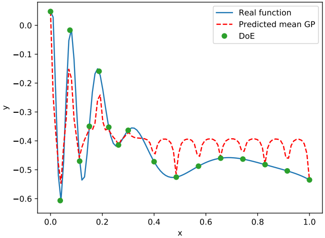


- $\mathcal{X}$ — A deterministic observed variable
- $\mathbf{h}_i$ — A distribution with **Non**-observed instantiations
- $\mathbf{y}$ — A distribution with observed instantiations

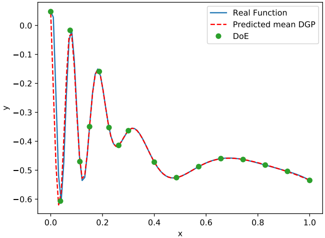
Deep Gaussian Processes



Deep Gaussian Processes



GP approximation of a non-stationary 1-D function. The GP model can not capture the stability of the region $[0.4, 1]$ and continues to oscillate



DGP approximation of a non-stationary 1-D function. The DGP model appropriately capture the two regions with different smoothness

Coupling of BO and DGPs

BO and DGPs

DGPs allow a flexible way of Bayesian kernel construction through input warping and dimensionality expansion to better fit data. Hence, its use for BO of non-stationary problems can be interesting. However, the integration of DGPs into the framework of BO is not direct [Hebbal et al., 2018].

Coupling of BO and DGPs

BO and DGPs

DGPs allow a flexible way of Bayesian kernel construction through input warping and dimensionality expansion to better fit data. Hence, its use for BO of non-stationary problems can be interesting. However, the integration of DGPs into the framework of BO is not direct [Hebbal et al., 2018].

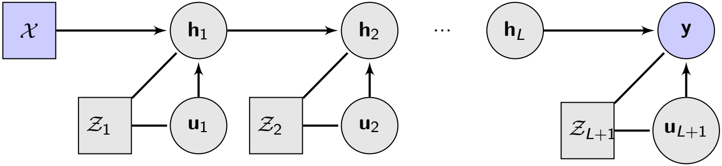
Training the model	Infill criteria	Configuration of the network

Coupling of BO and DGPs

Training the model | Infill criteria | Configuration of the network

The marginal likelihood of DGPs is not analytically tractable.
Multiple training approaches of DGPs have been developed based on variational inference and the introduction of inducing variables.

$\mathbf{f}_0 \sim \mathcal{GP}(0, K_{\mathcal{X}\mathcal{X}}) + \epsilon_0$ $\quad$ $\mathbf{f}_1 \sim \mathcal{GP}(0, K_{\mathbf{h}_1\mathbf{h}_1}) + \epsilon_1$ $\quad$ $f_L \sim \mathcal{GP}(0, K_{\mathbf{h}_L\mathbf{h}_L}) + \epsilon_L$



ONERA
THE FRENCH AEROSPACE LAB

Coupling of BO and DGPs

Training the model | Infill criteria | Configuration of the network

The marginal likelihood of DGPs is not analytically tractable.
Multiple training approaches of DGPs have been developed based on variational inference and the introduction of inducing variables.

- Assumption of independence between layers in the approximation of an analytic tractable evidence lower bound :
  $\mathcal{L} < p(\mathbf{y}|\mathcal{X})$
  [Damianou and Lawrence, 2013] [Dai et al., 2015] [Bui et al., 2016] :
- No assumption made. A sampling approach to approximate the evidence lower bound $\mathcal{L}$
  [Salimbeni and Deisenroth, 2017]

Coupling of BO and DGPs

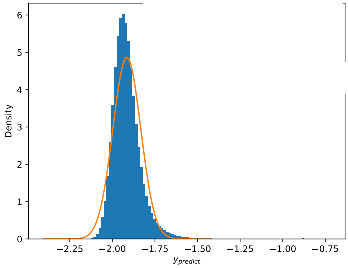
Training the model | **Infill criteria** | Configuration of the network

- When the posterior distribution is Gaussian, the Expected Hypervolume Improvement is analytically tractable for the two and three objectives problem.
- in DGPs the overall process prediction is *a priori* no longer Gaussian.
- A direct sampling approach on the value of the EHVI is computationally expensive.

Coupling of BO and DGPs

Training the model | **Infill criteria** | Configuration of the network

Approximation by a Gaussian distribution of the prediction (by sampling) is made in order to use the analytic expression of the EHVI



Coupling of BO and DGPs

Training the model | Infill criteria | **Configuration of the network**

The architecture of a DGP concerns :

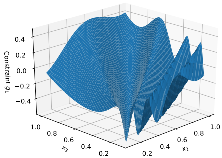
- The number of layers L
- the number of hidden units at each layer D_l
- the number of induced inputs M.

The architecture of a DGP is directly related to the time complexity in its training and its power of representation. Hence, the architecture of the DGP has to consider the complexity of the problem in hand and the time budget available has to be made.

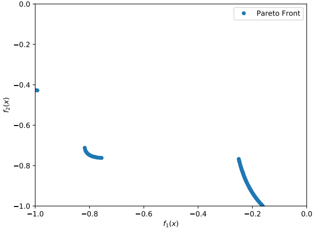
Analytic multi-objective problem

Modified TNK Problem [Deb et al., 2001]

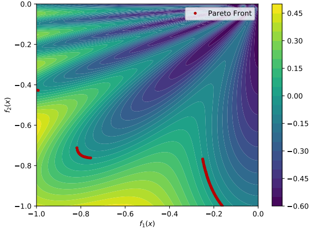
$$\begin{array}{rlll} \text{Min} & f_1(\mathbf{x}) & = -x_1 \\ \text{Min} & f_2(\mathbf{x}) & = -x_2 \\ \text{s.t} & g_1(\mathbf{x}) & = 0.5(x_1^2 + x_2^2) - 0.2\cos(20\arctan(0.3\frac{x_1}{x_2})) - 0.4 \\ \text{with} & & 0 < x_1 \leq 1 \\ \text{and} & & 0 < x_2 \leq 1 \end{array}$$



Constraint fun



Exact Pareto front

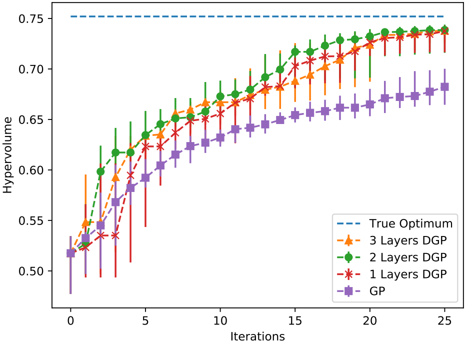


Exact Pareto front with constraint contour plot

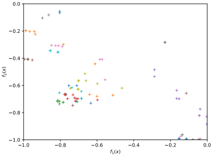
Analytic multi-objective problem

- Initial DoE : 20 points using a Latin Hypercube sampling
- Number of added points : 25
- EHVI criterion [Emmerich et al., 2006] with Probability of Feasability to handle the constraint, optimized with a differential evolution algorithm.

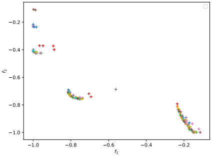
Analytic multi-objective problem



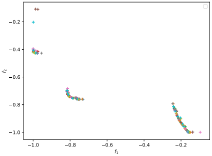
Convergence plot of BO with different architecture of DGPs and a regular GP. The markers indicate the median of the minimum obtained while the errorbars indicate the first and the third quartiles.



NSGA-II Pareto Fronts



BO-GP Pareto Fronts



DGP 3HL Pareto Fronts

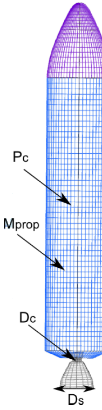
Problem : Solid-propellant booster

Problem : Optimization of a set of objectives for a booster engine.
Objectives : Minimization of the Gross Lift-off Weight ($GLOW$) and maximization of the change in velocity (ΔV).

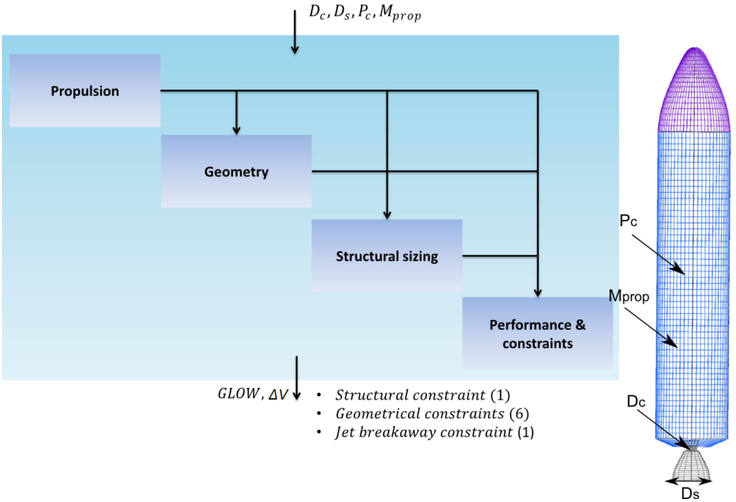
$$\begin{array}{ll} \text{Minimize :} & [GLOW(\mathbf{X}), -\Delta V(\mathbf{X})] \\ \text{According to :} & \mathbf{X} = [M_{prop}, P_c, D_c, D_s] \\ \text{subject to :} & \begin{cases} 1 \text{ Structural constraint} \\ 6 \text{ Geometrical constraints} \\ 1 \text{ Jet breakaway constraints} \end{cases} \end{array}$$

Design variables :

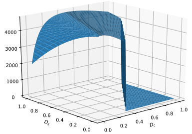
- Propellant mass : $5t \leq M_{prop} \leq 15t$
- Combustion chamber pressure : $5Bar \leq P_c \leq 100Bar$
- Throat nozzle diameter : $0.2m \leq D_c \leq 1m$
- Nozzle exit diameter : $0.5m \leq D_s \leq 1.2m$



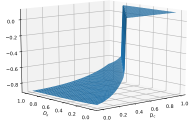
Problem : Solid-propellant booster



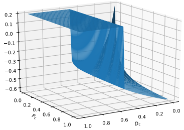
Problem : Solid-propellant booster



A sectional view of the change in velocity according to the diameters of the nozzle.

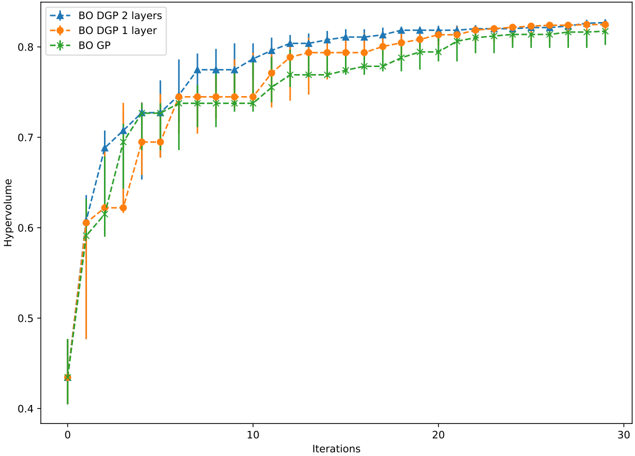


A sectional view of a constraint according to the diameters of the nozzle.



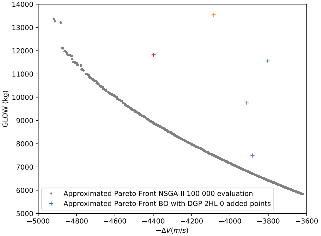
A sectional view of a constraint according to the throat nozzle diameter and the combustion chamber pressure.

Problem : Solid-propellant booster

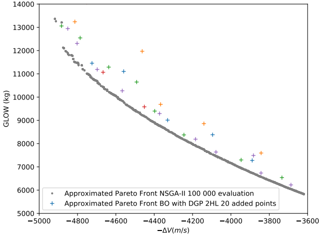


Convergence plot of BO with different architecture of DGPs and a regular GP.

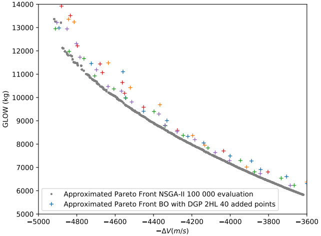
Problem : Solid-propellant booster



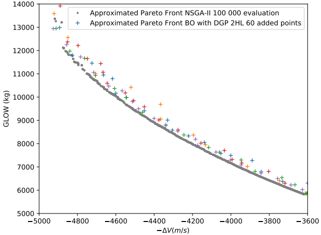
Initial DoE 40 initial points (0 Added points)



20 Added points



40 Added points



60 Added points

ONERA
THE FRENCH AEROSPACE LAB

Conclusions

- Propositions to make the coupling MO-BO and DGPs possible.
- Experimentations on analytical and representative physical problem confirming the interest of using DGPs over GPs in MO-BO.
- Highlights the challenges arising from this coupling.

Future works :

- An adaptive framework to set the configuration of the DGP according to the dynamic of BO.
- Investigation on the time reduction in the training of DGPs.
- Dependent multi-output DGPs for multi-objective optimization.

Conclusions

- Propositions to make the coupling MO-BO and DGPs possible.
- Experimentations on analytical and representative physical problem confirming the interest of using DGPs over GPs in MO-BO.
- Highlights the challenges arising from this coupling.

Future works :

- An adaptive framework to set the configuration of the DGP according to the dynamic of BO.
- Investigation on the time reduction in the training of DGPs.
- Dependent multi-output DGPs for multi-objective optimization.

Conclusions

- Propositions to make the coupling MO-BO and DGPs possible.
- Experimentations on analytical and representative physical problem confirming the interest of using DGPs over GPs in MO-BO.
- Highlights the challenges arising from this coupling.

Future works :

- An adaptive framework to set the configuration of the DGP according to the dynamic of BO.
- Investigation on the time reduction in the training of DGPs.
- Dependent multi-output DGPs for multi-objective optimization.

Conclusions

- Propositions to make the coupling MO-BO and DGPs possible.
- Experimentations on analytical and representative physical problem confirming the interest of using DGPs over GPs in MO-BO.
- Highlights the challenges arising from this coupling.

Future works :

- An adaptive framework to set the configuration of the DGP according to the dynamic of BO.
- Investigation on the time reduction in the training of DGPs.
- Dependent multi-output DGPs for multi-objective optimization.

Conclusions

Thank you for your attention !

Bibliography I

Bui, T., Hernández-Lobato, D., Hernandez-Lobato, J., Li, Y., and Turner, R. (2016).
Deep gaussian processes for regression using approximate expectation propagation.
In *International Conference on Machine Learning*, pages 1472–1481.

Dai, Z., Damianou, A., González, J., and Lawrence, N. (2015).
Variational auto-encoded deep gaussian processes.
arXiv preprint arXiv :1511.06455.

Damianou, A. and Lawrence, N. (2013).
Deep gaussian processes.
In *Artificial Intelligence and Statistics*, pages 207–215.

Deb, K., Pratap, A., and Meyarivan, T. (2001).
Constrained test problems for multi-objective evolutionary optimization.
In *evolutionary multi-criterion optimization*, pages 284–298. Springer.

Emmerich, M. T., Giannakoglou, K. C., and Naujoks, B. (2006).
Single-and multiobjective evolutionary optimization assisted by gaussian random field metamodels.
IEEE Transactions on Evolutionary Computation, 10(4) :421–439.

Haas, T. C. (1990).
Kriging and automated variogram modeling within a moving window.
Atmospheric Environment. Part A. General Topics, 24(7) :1759–1769.

Hebbal, A., Brevault, L., Balesdent, M., Taibi, E.-G., and Melab, N. (2018).
Efficient global optimization using deep gaussian processes.
In *2018 IEEE Congress on Evolutionary Computation (CEC)*, pages 1–8. IEEE.

Bibliography II

Higdon, D., Swall, J., and Kern, J. (1999).
Non-stationary spatial modeling.
Bayesian statistics, 6(1) :761–768.

Paciorek, C. J. and Schervish, M. J. (2006).
Spatial modelling using a new class of nonstationary covariance functions.
Environmetrics, 17(5) :483–506.

Rasmussen, C. E. (2004).
Gaussian processes in machine learning.
In *Advanced lectures on machine learning*, pages 63–71. Springer.

Rasmussen, C. E. and Ghahramani, Z. (2002).
Infinite mixtures of gaussian process experts.
In *Advances in neural information processing systems*, pages 881–888.

Salimbeni, H. and Deisenroth, M. (2017).
Doubly stochastic variational inference for deep gaussian processes.
arXiv preprint arXiv :1705.08933.

Xiong, Y., Chen, W., Apley, D., and Ding, X. (2007).
A non-stationary covariance-based kriging method for metamodelling in engineering design.
International Journal for Numerical Methods in Engineering, 71(6) :733–756.