Bayesian Optimization using Deep Gaussian Processes for non-stationary problems

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

EDF R&D
Séminaire APSPP - 21/03/2019

ONERA
THE FRENCH AEROSPACE LAB

Inria
INVENTEURS DU MONDE NUMÉRIQUE

Université de Lille
1 SCIENCES ET TECHNOLOGIES

Table of Contents

Introduction

Review on Bayesian Optimization
- BO framework
- Definitions
- Limits of GPs
- Review of methods for handling non stationarity

Deep Gaussian Processes
- Definition
- Coupling of BO and DGPs

Experimentations
- Analytic single-objective problem
- Analytic Multi-objective problem
- multi-objective aerospace problem

Conclusions

Context

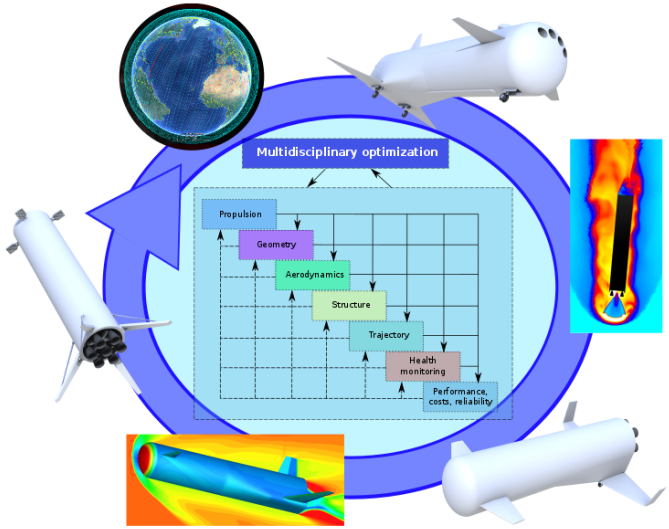
Resolution of an optimization problem characterized by :

- Black box and computationally expensive functions.
- Non-stationary functions.

Context

Resolution of an optimization problem characterized by :

- Black box and computationally expensive functions.
- Non-stationary functions.



Multi-disciplinary optimization of an aerospace vehicle

Context

Resolution of an optimization problem characterized by :

- Black box and computationally expensive functions.
- Non-stationary functions.

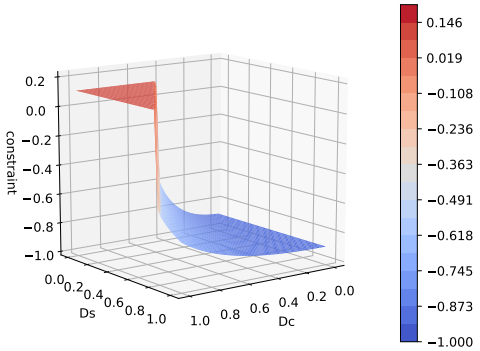
✗ Gradient based optimization approaches	✗ Classic evolutionary algorithms	✓ Bayesian Optimization

Context

Resolution of an optimization problem characterized by :

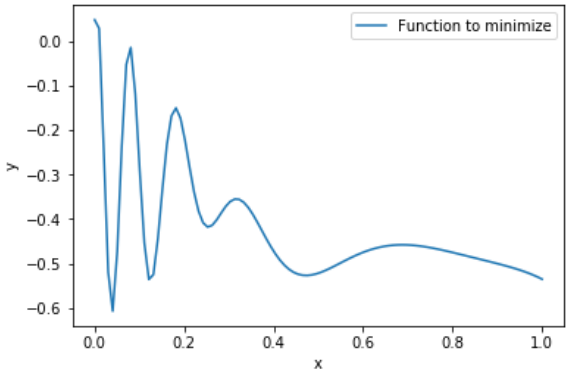
- Black box and computationally expensive functions.
- Non-stationary functions.

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 :

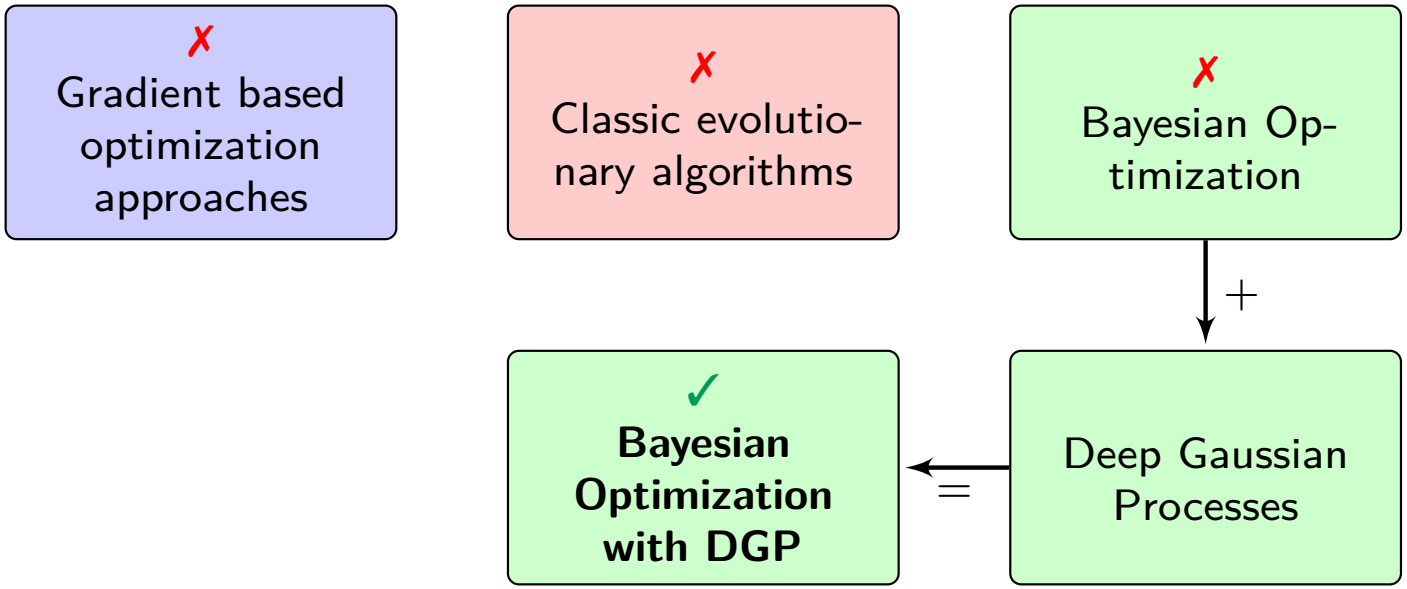
- Black box and computationally expensive functions.
- Non-stationary functions.

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

Context

Resolution of an optimization problem characterized by :

- Black box and computationally expensive functions.
- Non-stationary functions.



[Damianou and Lawrence, 2013]

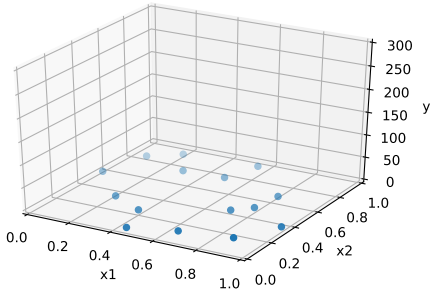
BO framework

Bayesian Optimization (BO) [Jones et al., 1998]

BO framework

Bayesian Optimization (BO) [Jones et al., 1998]

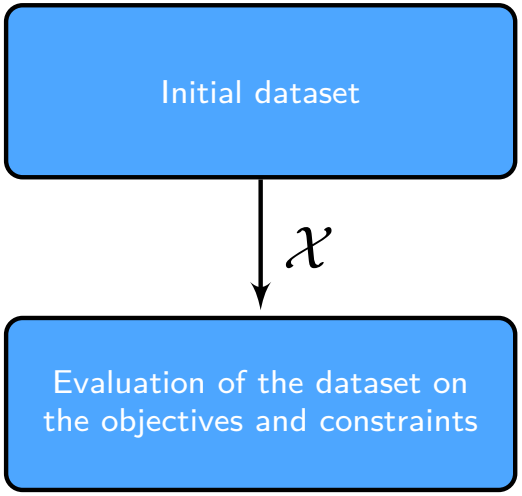
Initial dataset

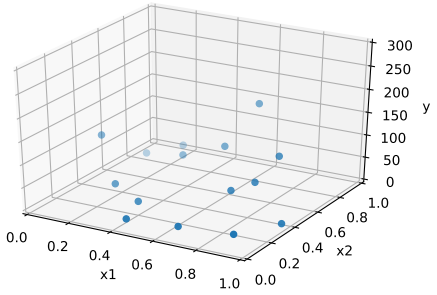


Design of Experiment depending on the dimension and the nature of the problem

BO framework

Bayesian Optimization (BO) [Jones et al., 1998]

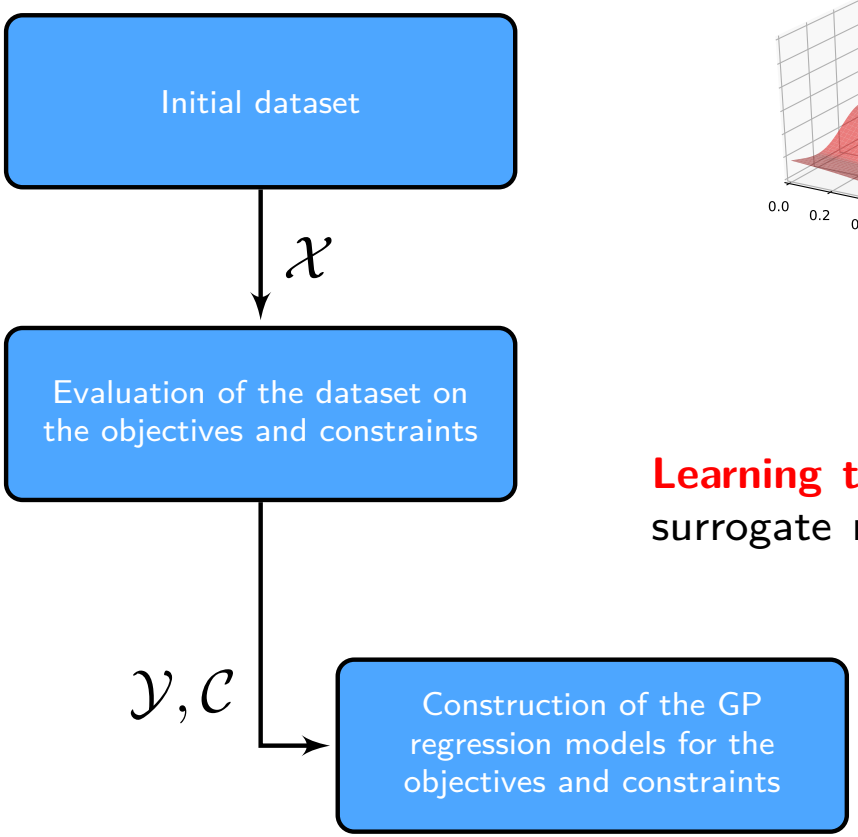


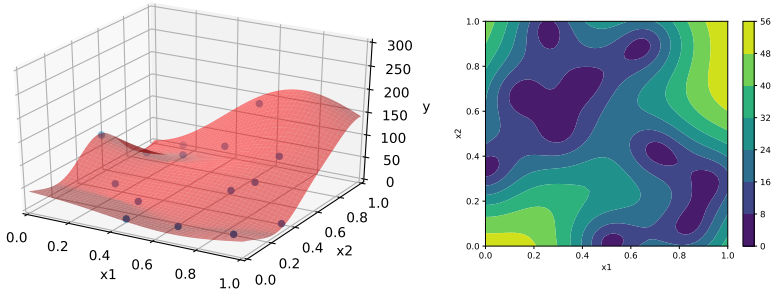


Calls the **expensive** black-box functions

BO framework

Bayesian Optimization (BO) [Jones et al., 1998]

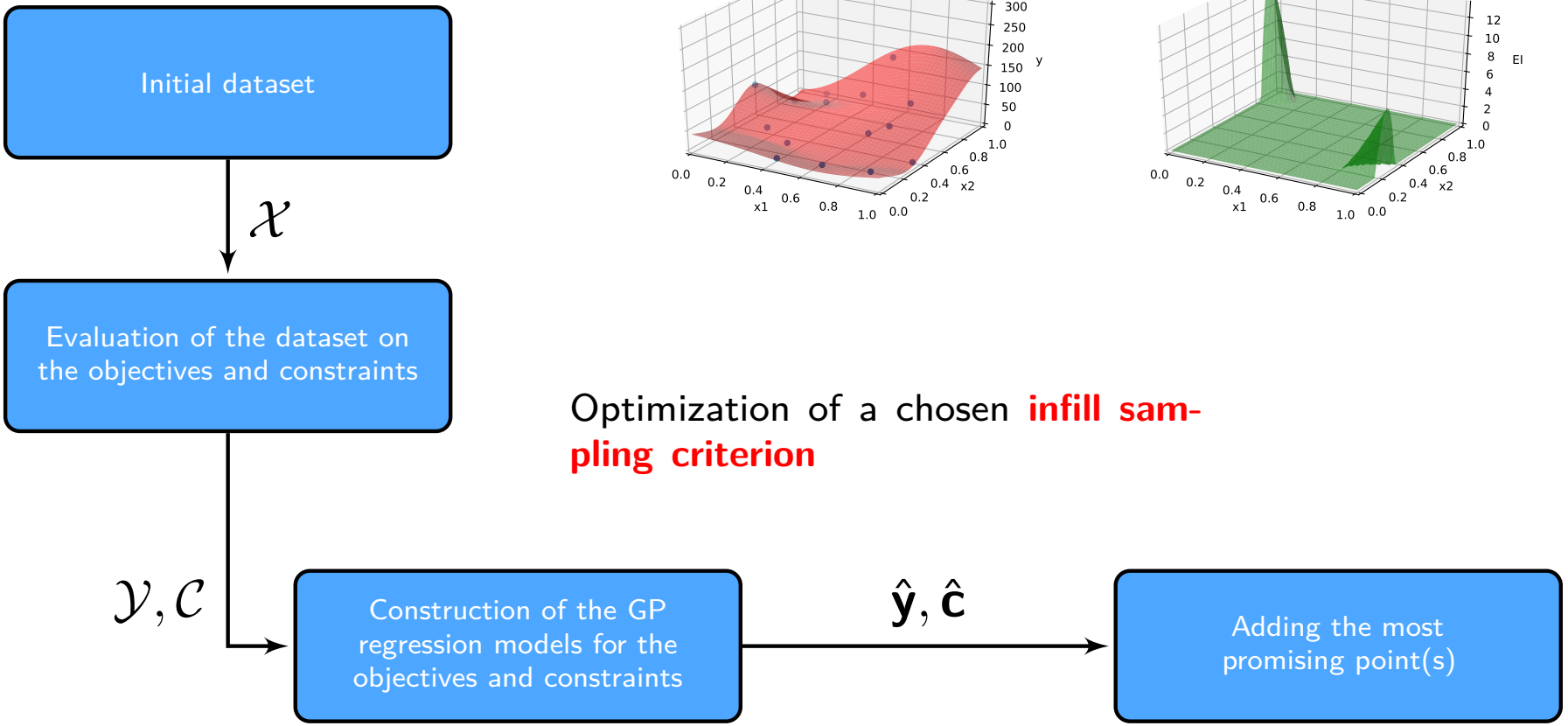


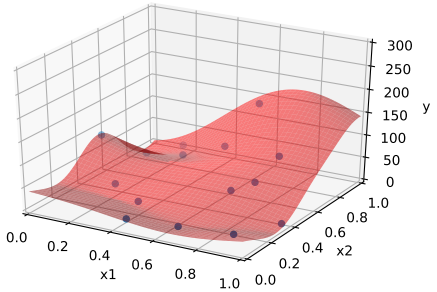


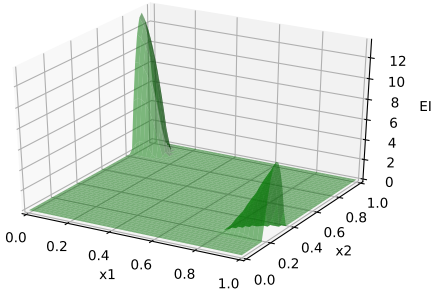
Learning the hyperparameters of each surrogate model

BO framework

Bayesian Optimization (BO) [Jones et al., 1998]



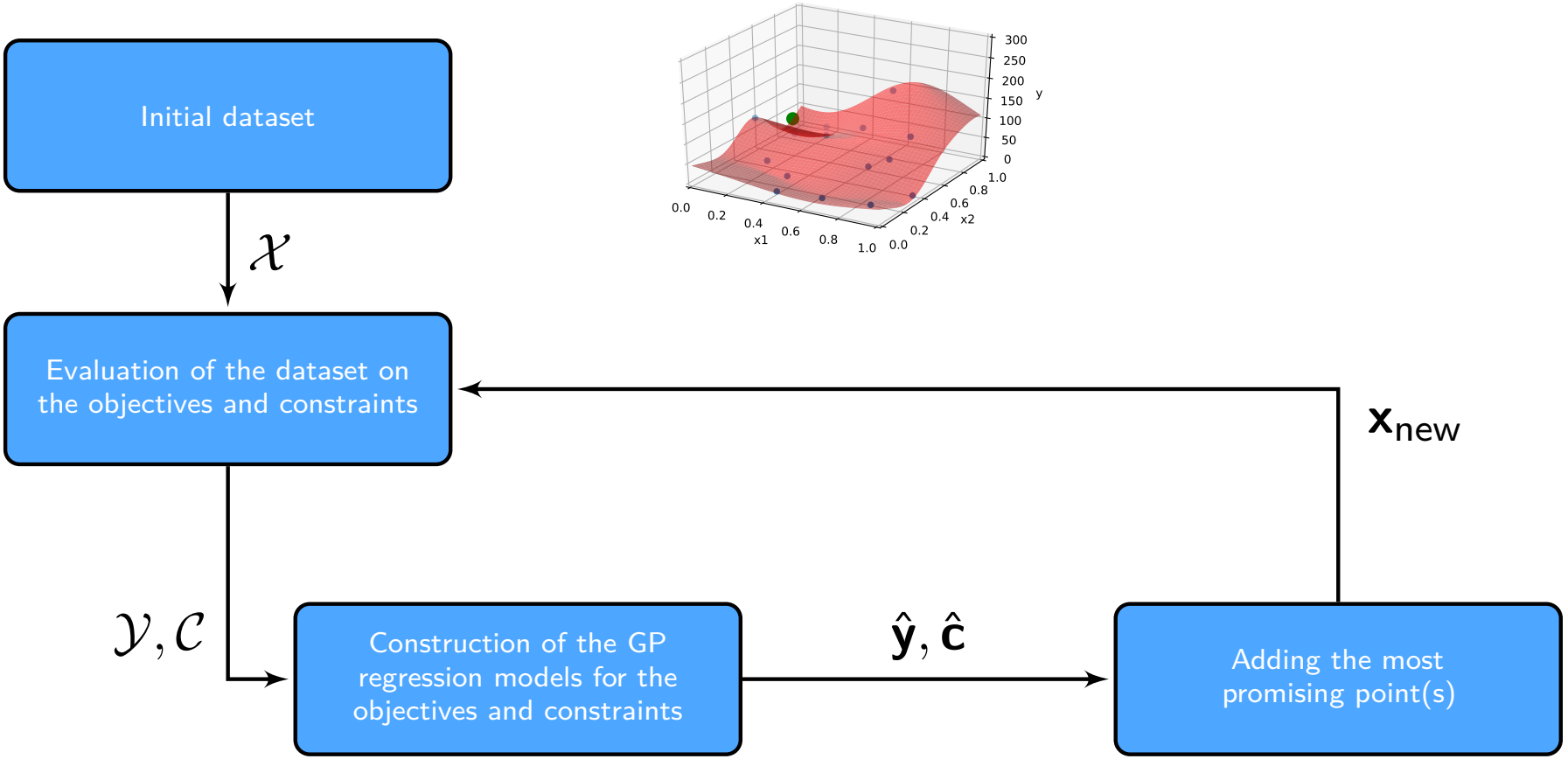




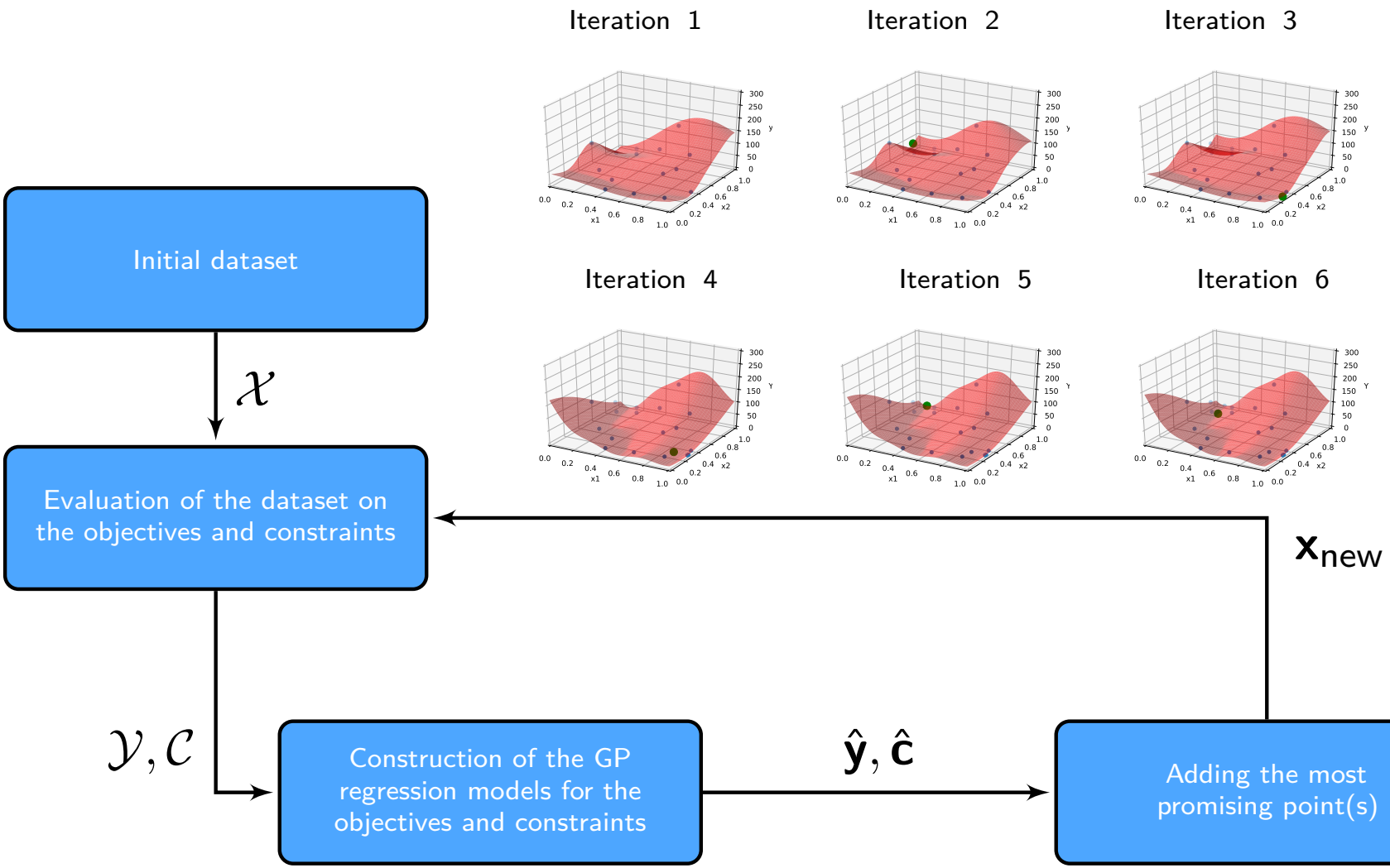
Optimization of a chosen **infill sampling criterion**

BO framework

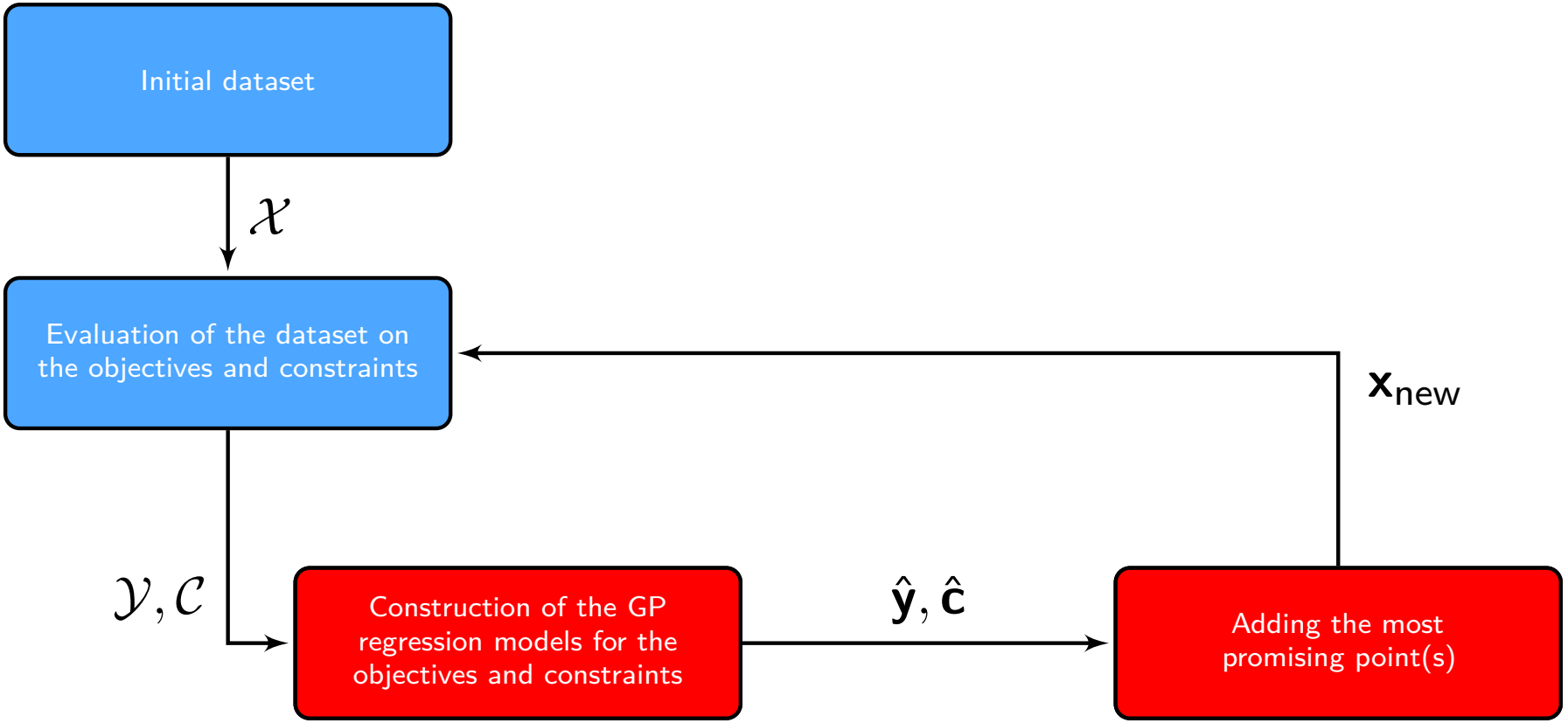
Bayesian Optimization (BO) [Jones et al., 1998]



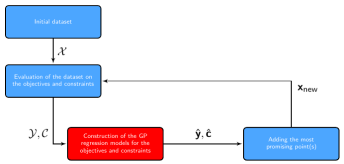
BO framework



BO framework

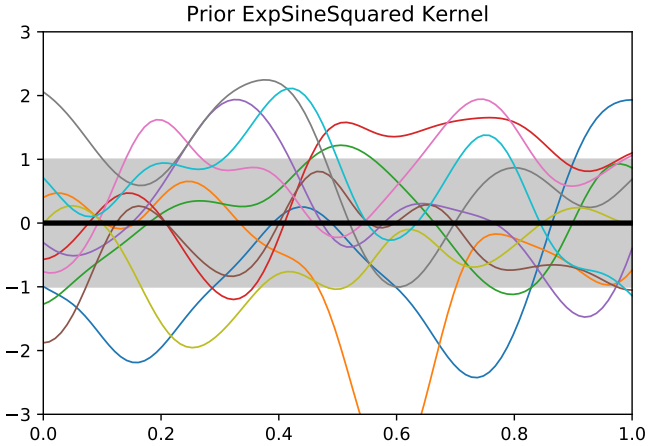


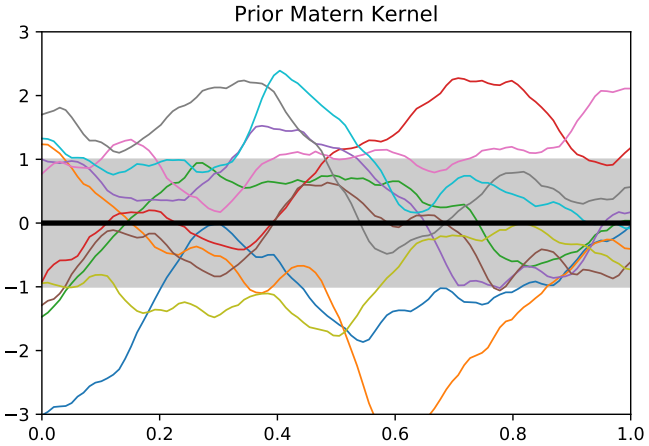
Gaussian Process Regression



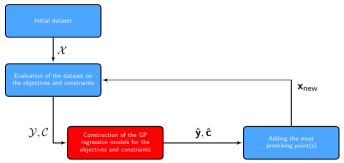
Gaussian process [Rasmussen, 2004]

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) : $f(.) \sim \mathcal{GP}(\mu(.), k^{\Theta}(.))$





Gaussian Process Regression

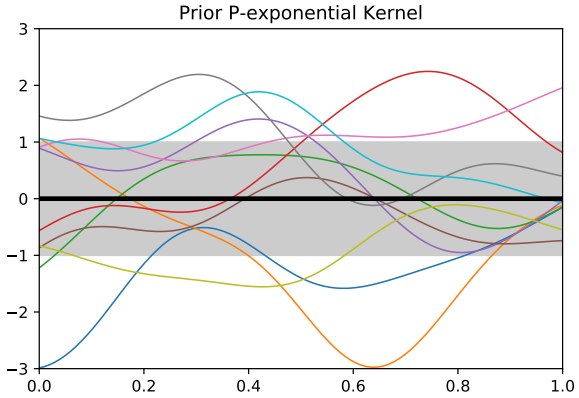


Gaussian process [Rasmussen, 2004]

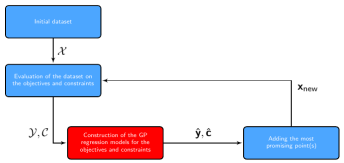
A Gaussian Process is used to describe a distribution over function. It is a collection of infinite random variables, **any finite number of which have a joint Gaussian distribution**.
It is defined by its mean function and covariance function (Kernel) : $f(.) \sim \mathcal{GP}(\mu(.), k^{\Theta}(.))$

Automatic Relevance Determination (ARD)p-exponential kernel :

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



Gaussian Process Regression

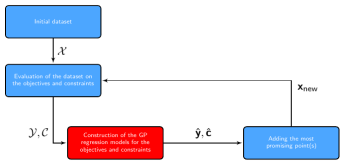


Gaussian process [Rasmussen, 2004]

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) : $f(.) \sim \mathcal{GP}(\mu(.), k^{\Theta}(.))$

Goal : predict y^* in $\mathbf{x}^*$ given $\mathcal{Y}$ a set of observations in $\mathcal{X}$

Gaussian Process Regression



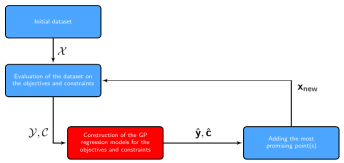
Gaussian process [Rasmussen, 2004]

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) : $f(.) \sim \mathcal{GP}(\mu(.), k^{\Theta}(.))$

Goal : predict y^* in $\mathbf{x}^*$ given $\mathcal{Y}$ a set of observations in $\mathcal{X}$

$$\xrightarrow{\text{GP property}} \begin{pmatrix} \mathcal{Y} \\ y^* \end{pmatrix} \sim \mathcal{N}\left(\mu, \begin{matrix} K^{\Theta}_{\mathcal{X},\mathcal{X}} & K^{\Theta}_{\mathcal{X},\mathbf{x}^*} \\ K^{\Theta}_{\mathbf{x}^*,\mathcal{X}} & K^{\Theta}_{\mathbf{x}^*,\mathbf{x}^*} \end{matrix}\right)$$

Gaussian Process Regression



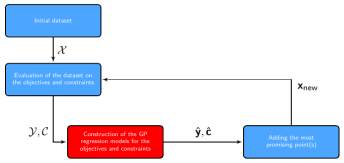
Gaussian process [Rasmussen, 2004]

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) : $f(.) \sim \mathcal{GP}(\mu(.), k^{\Theta}(.))$

Goal : predict y^* in $\mathbf{x}^*$ given $\mathcal{Y}$ a set of observations in $\mathcal{X}$

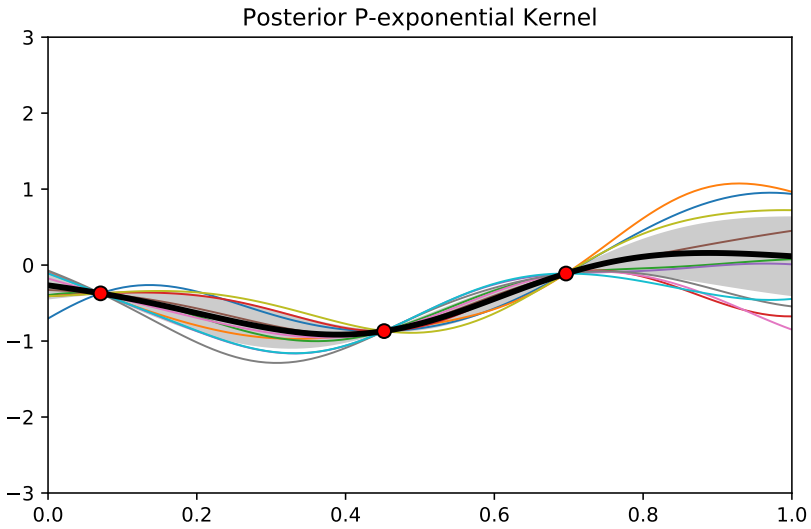
$$\xrightarrow{\text{GP property}} \begin{pmatrix} \mathcal{Y} \\ y^* \end{pmatrix} \sim \mathcal{N}\left(\mu, \begin{matrix} K^{\Theta}_{\mathcal{X},\mathcal{X}} & K^{\Theta}_{\mathcal{X},\mathbf{x}^*} \\ K^{\Theta}_{\mathbf{x}^*,\mathcal{X}} & K^{\Theta}_{\mathbf{x}^*,\mathbf{x}^*} \end{matrix}\right) \xrightarrow{\text{Bayes' theorem}} y^*|\mathbf{x}^*, \mathcal{X}, \mathcal{Y} \sim \mathcal{N}(\hat{y}^*, \hat{s}^*)$$

Gaussian Process Regression



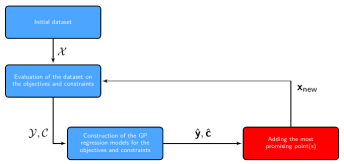
Gaussian process [Rasmussen, 2004]

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) : $f(.) \sim \mathcal{GP}(\mu(.), k^{\Theta}(.))$

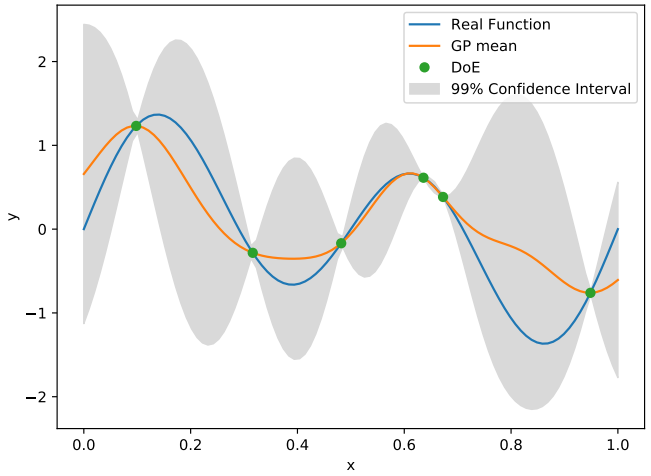


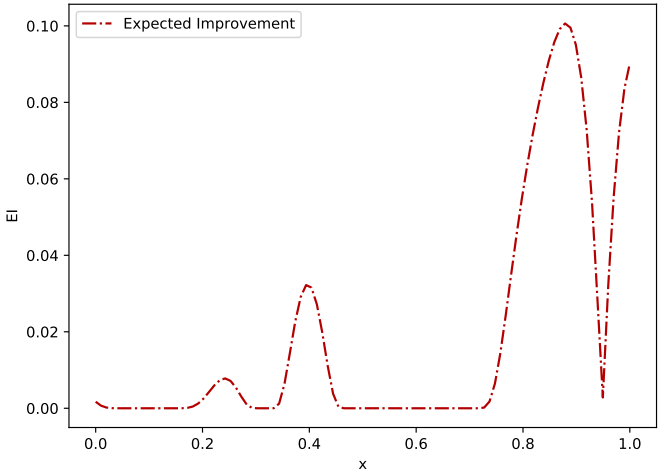
$$y^*|\mathbf{x}^*, \mathcal{X}, \mathcal{Y} \sim \mathcal{N}(\hat{y}^*, \hat{s}^*)$$

Infill Criteria

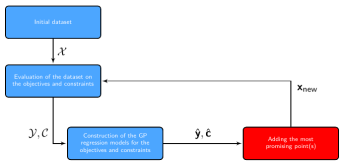


$$
\begin{aligned}
EI(\mathbf{x}) &= \int t.\mathrm{PDF}_{I(\mathbf{x})}(t)\mathrm{d}t \\
&= (y_{\min} - \hat{y}(\mathbf{x}))\Phi\left(\frac{y_{\min} - \hat{y}(\mathbf{x})}{\hat{s}(\mathbf{x})}\right) \\
&+ \hat{s}(\mathbf{x})\phi\left(\frac{y_{\min} - \hat{y}(\mathbf{x})}{\hat{s}(\mathbf{x})}\right).
\end{aligned}
$$

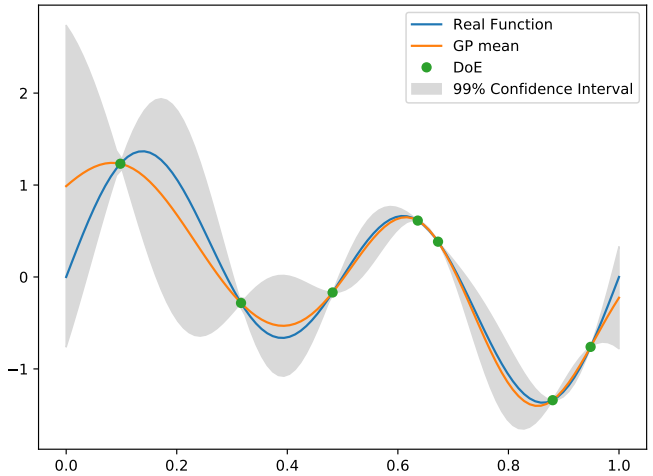


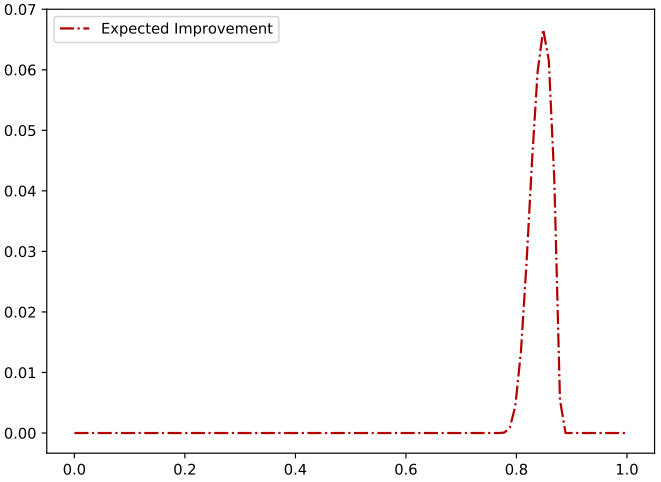


Infill Criteria



$$
\begin{aligned}
EI(\mathbf{x}) &= \int t.\mathrm{PDF}_{I(\mathbf{x})}(t)\mathrm{d}t \\
&= (y_{\min} - \hat{y}(\mathbf{x}))\Phi\left(\frac{y_{\min} - \hat{y}(\mathbf{x})}{\hat{s}(\mathbf{x})}\right) \\
&+ \hat{s}(\mathbf{x})\phi\left(\frac{y_{\min} - \hat{y}(\mathbf{x})}{\hat{s}(\mathbf{x})}\right).
\end{aligned}
$$

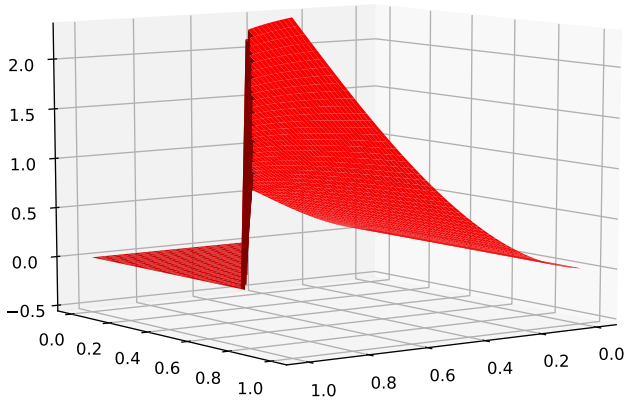




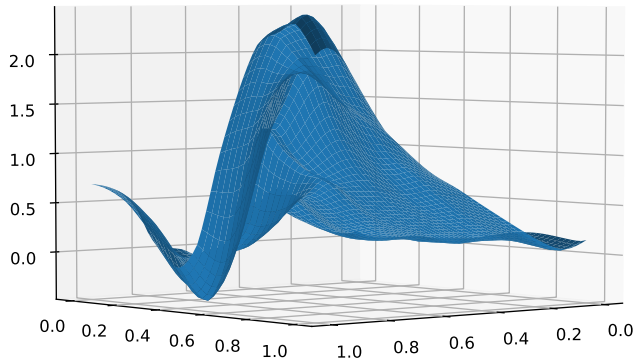
Limits of GPs

Handling non stationarity

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



2-D Function with an abrupt change

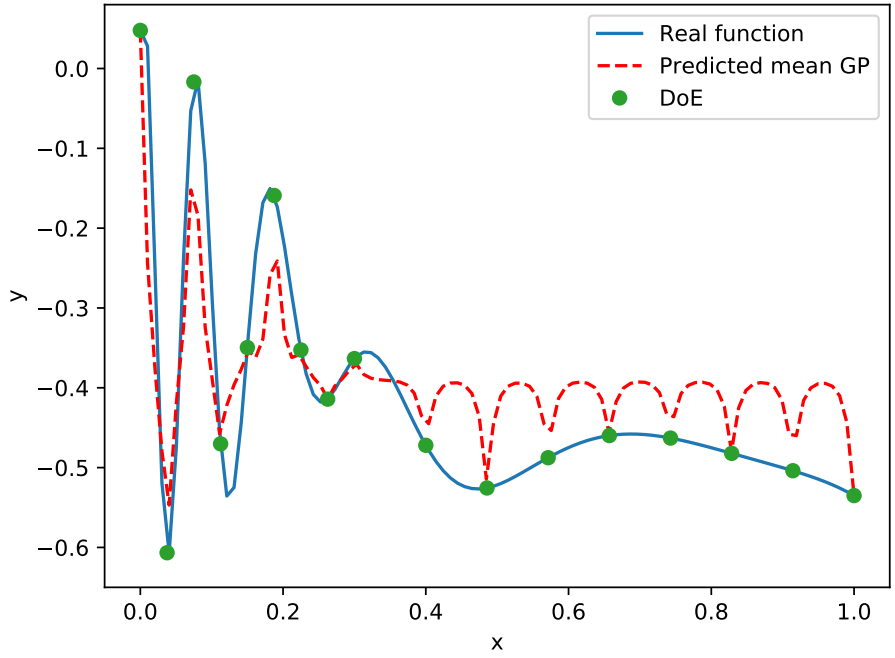


GP Approximation

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

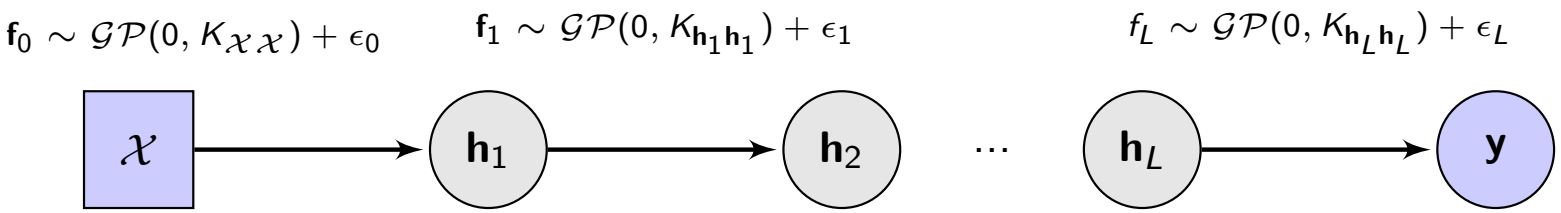
- Direct formulation of non-stationary covariance function :
  - Non stationary version of the squared exponential covariance function [Higdon et al., 1999],
  - Non stationary version of the Matérn covariance function [Paciorek and Schervish, 2006].
- Local stationary covariance function :
  - Moving window approach where the training and prediction region move along the input space [Haas, 1990],
  - Different stationary GPs in different subspaces of the input space [Rasmussen and Ghahramani, 2002].
- Non-linear mapping :
  - Piece-wise density function with parametrized knots to map the input space with a deformed space to model the non-stationary response by a stationary model [Xiong et al., 2007].

Deep Gaussian Processes

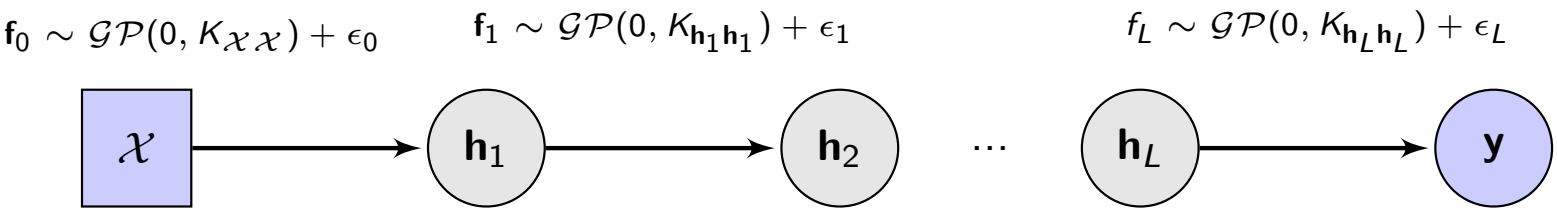
Deep Gaussian Processes [Damianou and Lawrence, 2013]

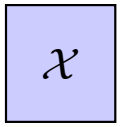
DGPs are a class of surrogate models based on the structure of neural networks, where each layer is a GP. It considers that the statistical relationship between the inputs and the response is expressed by a functional composition of GPs :

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



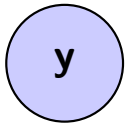
Deep Gaussian Processes





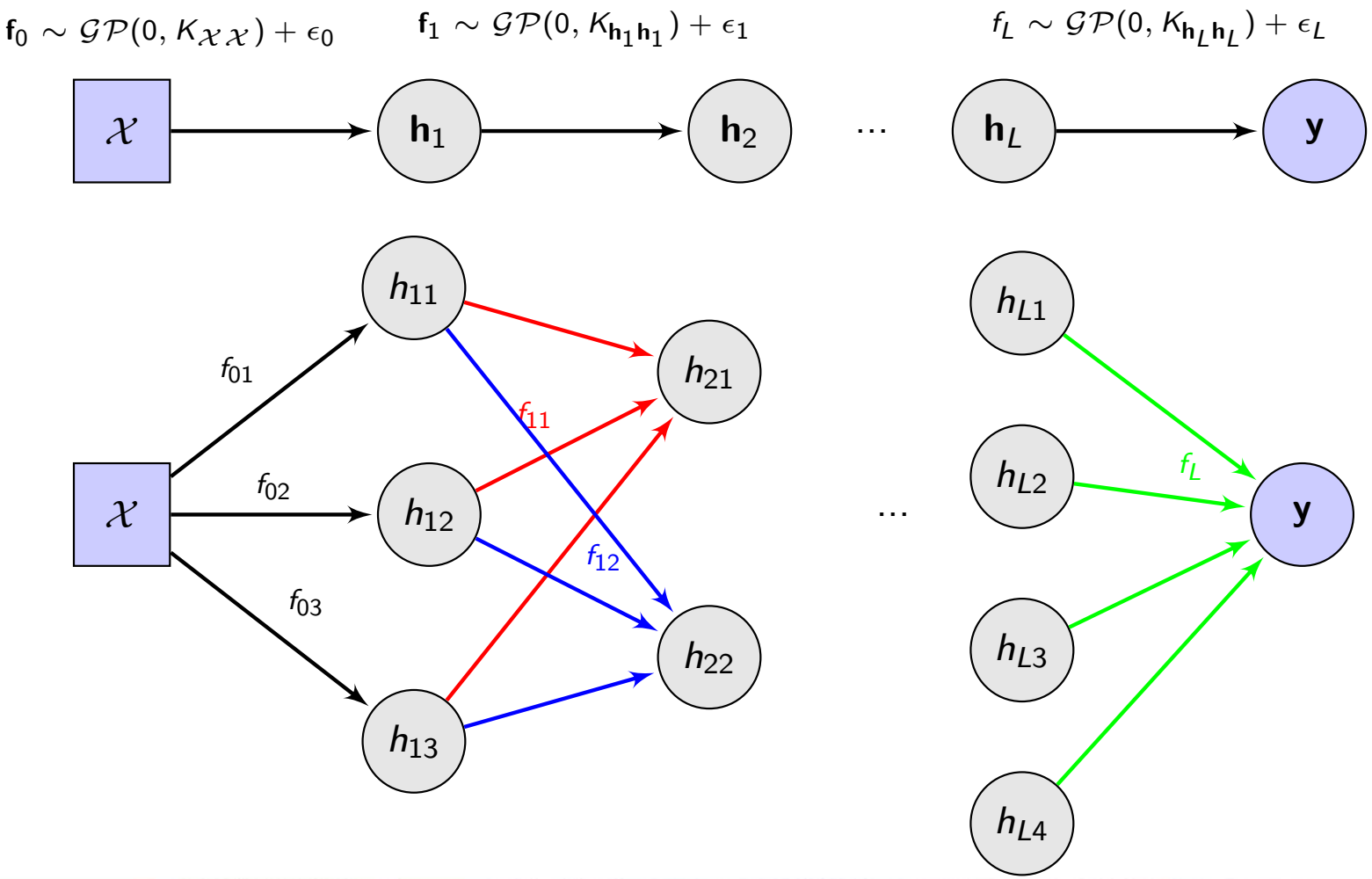
A deterministic observed variable

A distribution with **Non**-observed instantiations

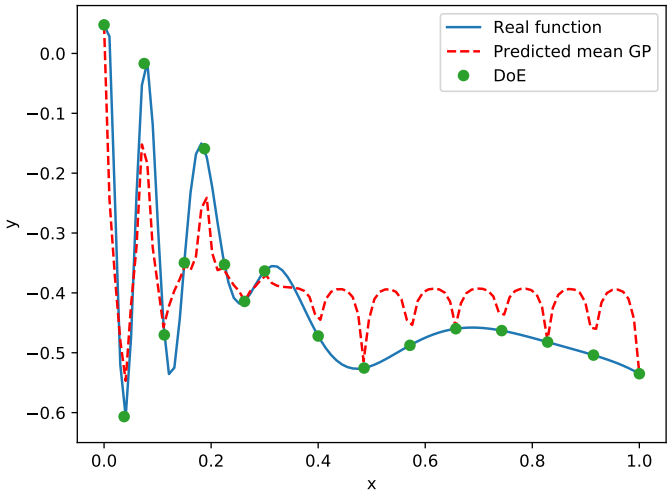


A distribution with observed instantiations

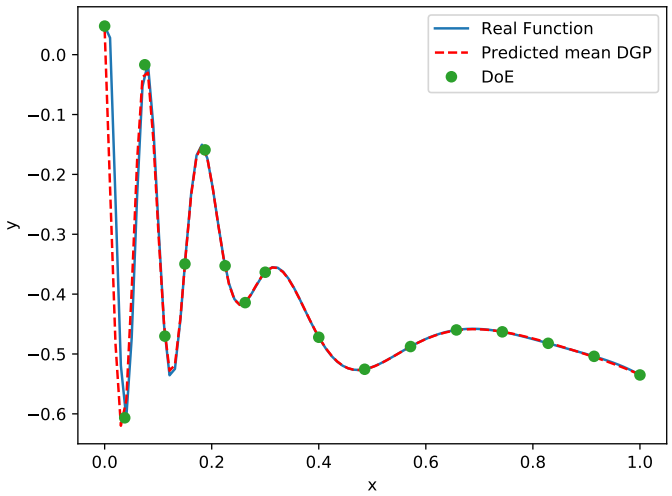
Deep Gaussian Processes



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.

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.

Training the model

Infill criteria

Configuration of the network

Coupling of BO and DGPs

Training the model | Infill criteria | Configuration of the network

$$p\left(\mathbf{y}|\mathcal{X}\right) = \int_{\{\mathbf{h}_l\}_1^L} p\left(\mathbf{y}, \{\mathbf{h}_l\}_1^L | \mathcal{X}\right) \mathrm{d}\{\mathbf{h}_l\}_1^L$$

← Analytically Intractable

Coupling of BO and DGPs

Training the model | Infill criteria | Configuration of the network

$$p(\mathbf{y}|\mathcal{X}) = \int_{\{\mathbf{h}_l\}_1^L} p\left(\mathbf{y}, \{\mathbf{h}_l\}_1^L | \mathcal{X}\right) \mathrm{d}\{\mathbf{h}_l\}_1^L$$

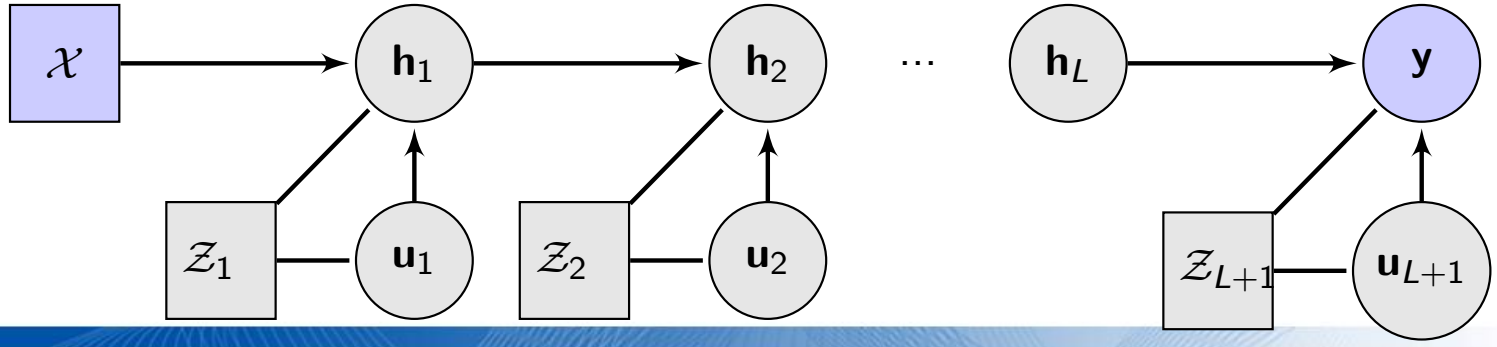
← Analytically Intractable

↓

Augmenting the latent space by inducing inputs-outputs $\{\mathcal{Z}_l\}_{l=1}^L, \{\mathbf{u}_l\}_{l=1}^L$

← Sparse Approximation

$\mathbf{f}_0 \sim \mathcal{GP}(0, K_{\mathcal{X}\mathcal{X}} + \sigma_0^2 \mathbf{I})$ $\mathbf{f}_1 \sim \mathcal{GP}(0, K_{\mathbf{h}_1\mathbf{h}_1} + \sigma_1^2 \mathbf{I})$ $\mathbf{f}_L \sim \mathcal{GP}(0, K_{\mathbf{h}_L\mathbf{h}_L} + \sigma_L^2 \mathbf{I})$



Coupling of BO and DGPs

Training the model | Infill criteria | Configuration of the network

$$p(\mathbf{y}|\mathcal{X}) = \int_{\{\mathbf{h}_l\}_1^L} p\left(\mathbf{y}, \{\mathbf{h}_l\}_1^L | \mathcal{X}\right) \mathrm{d}\{\mathbf{h}_l\}_1^L$$

← Analytically Intractable

↓

Augmenting the latent space by inducing inputs-outputs $\{\mathcal{Z}_l\}_{l=1}^L, \{\mathbf{u}_l\}_{l=1}^L$

↓

Variationnal approach : Obtaining a lower bound of the likelihood by approximating the latent distributions

← [Damianou and Lawrence, 2013]

$$\{q(\mathbf{u}_l) \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)\}_{l=1}^{l=L}$$

Coupling of BO and DGPs

Training the model | Infill criteria | Configuration of the network

$$p(\mathbf{y}|\mathcal{X}) = \int_{\{\mathbf{h}_l\}_1^L} p\left(\mathbf{y}, \{\mathbf{h}_l\}_1^L | \mathcal{X}\right) \mathrm{d}\{\mathbf{h}_l\}_1^L$$

Analytically Intractable

Augmenting the latent space by inducing inputs-outputs $\{\mathcal{Z}_l\}_{l=1}^L, \{\mathbf{u}_l\}_{l=1}^L$

Variationnal approach : Obtaining a lower bound of the likelihood by approximating the latent distributions

Optimization of the lower bound

$$\max \mathcal{L} = \sum_{i=1}^{N} \mathbb{E}_{q(\mathbf{h}_L^{(i)})} \left[\log p(y^{(i)}|\mathbf{h}_L^{(i)})\right] - \sum_{l=1}^{L} KL\left[q(\mathbf{u}_l||p(\mathbf{u}_l)\right]$$

According to :

-The Gaussian processes parameters : $\{\Theta\}_{l=1}^{l=L}$

-The inducing inputs $\{\mathcal{Z}_l\}_{l=1}^{l=L}$

-The gaussian approximations $\{q(\mathbf{u}_l) \sim \mathcal{N}(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)\}_{l=1}^{l=L}$

Coupling of BO and DGPs

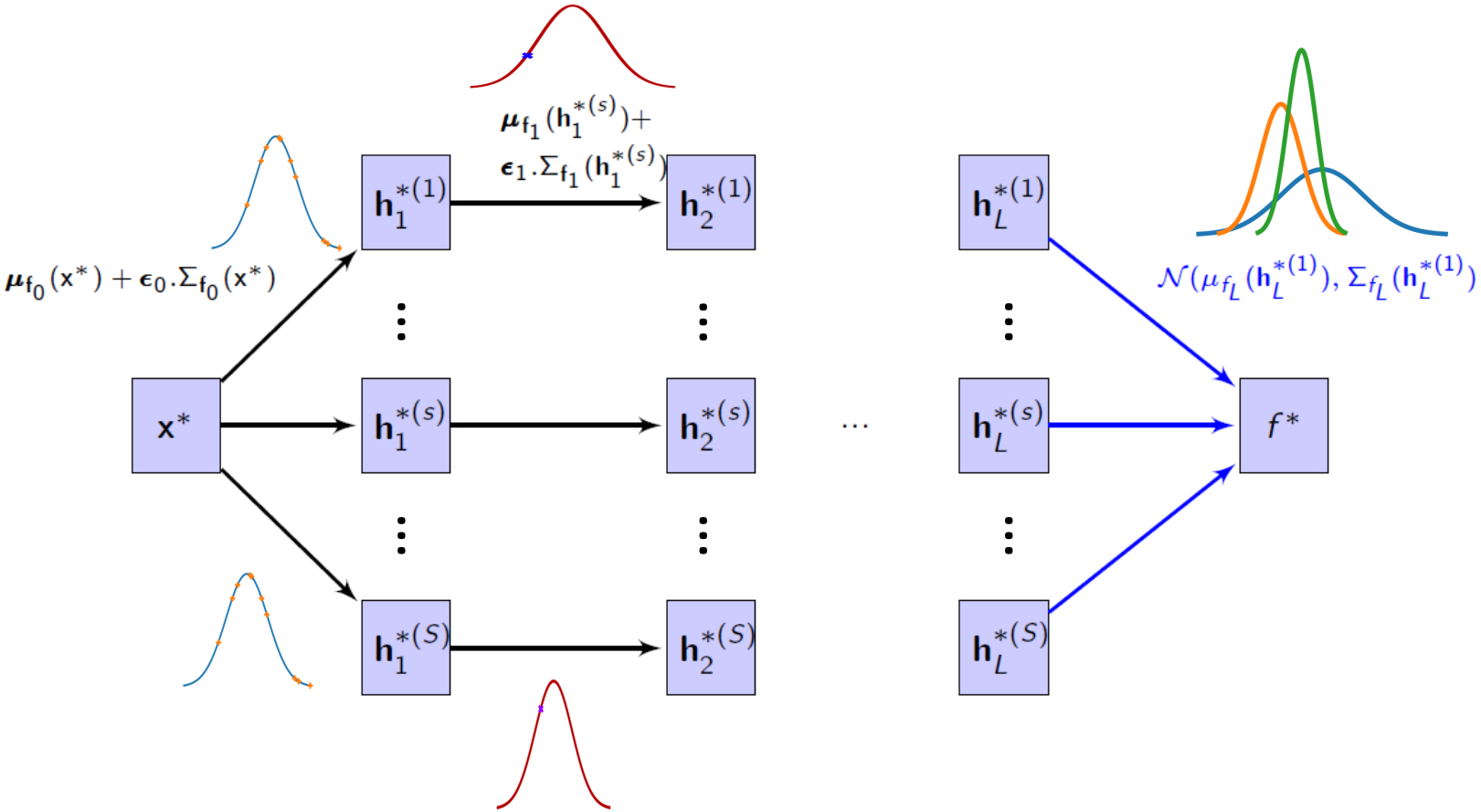
Training the model | **Infill criteria** | Configuration of the network

The Expected improvement is a widely used infill criterion. However, in DGPs the overall process prediction is no longer Gaussian. The direct application of the equation is not possible :

$$EI(\mathbf{x}) = (y_{\min} - \hat{y}(\mathbf{x}))\Phi\left(\frac{y_{\min} - \hat{y}(\mathbf{x})}{\hat{s}(\mathbf{x})}\right) + \hat{s}(\mathbf{x})\phi\left(\frac{y_{\min} - \hat{y}(\mathbf{x})}{\hat{s}(\mathbf{x})}\right)$$

an approximation of the prediction of DGPs by a Gaussian mixture is used :

Coupling of BO and DGPs



S Samples are drawn from the first layer, then, each sample is propagated through the whole network, with a single realization for each sample at each hidden layer. Until reaching the final layer where the mean and variance of the final GP are considered for each sample. Thus, the prediction is approximated by a Gaussian mixture of the S samples.

Coupling of BO and DGPs

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

The architecture of a DGP concerns :

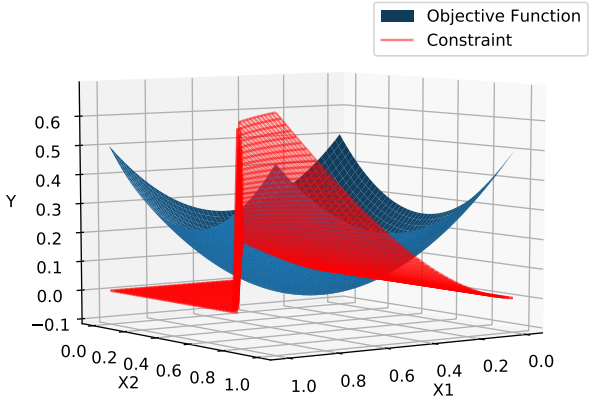
- The number of layers L
- the number of hidden units at each layer D_l
- the number of induced inputs M.
  The architecture of a DGP is directly related to the time complexity in its training and its power of representation. Hence, the architecture of the DGP has to consider the complexity of the problem in hand and the time budget available has to be made.

Experimentations

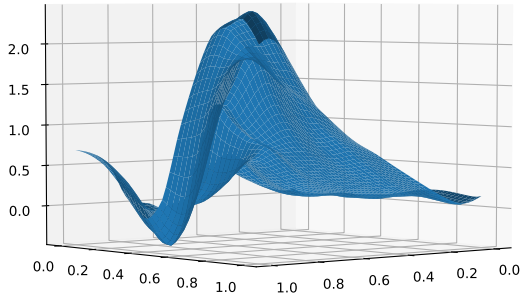
The objective of the experimentations is to :

- Confirm the interest of coupling BO and DGPs.
- Highlight the challenges arising from this coupling.

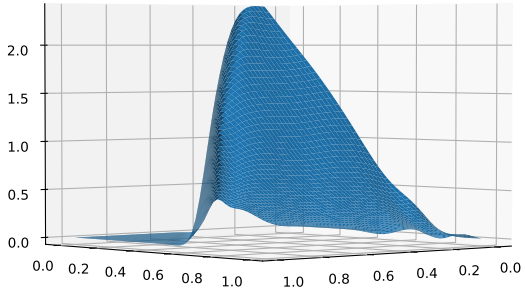
Constrained 2-D Problem



Constrained Function to minimize



Regular GP Approximation

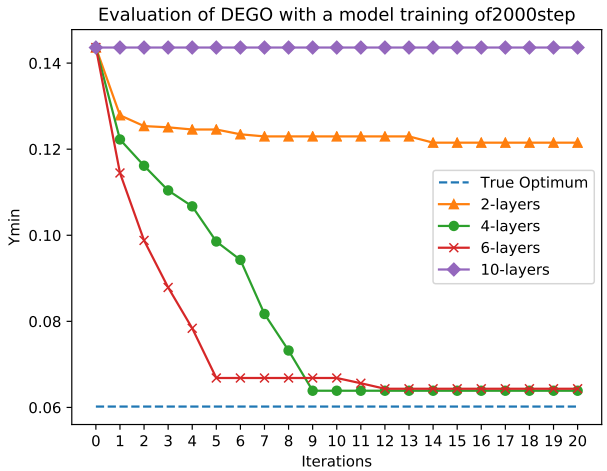


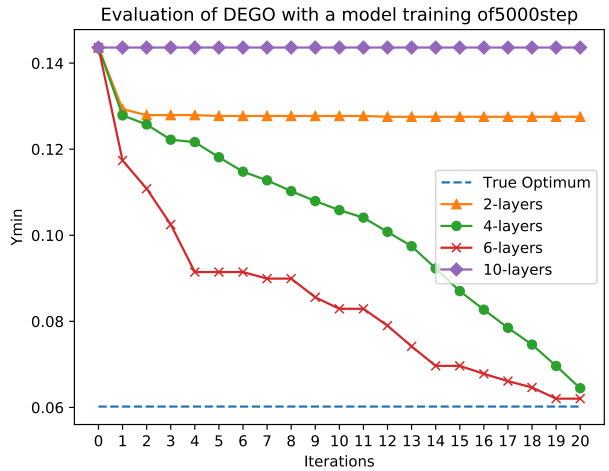
DGP Approximation

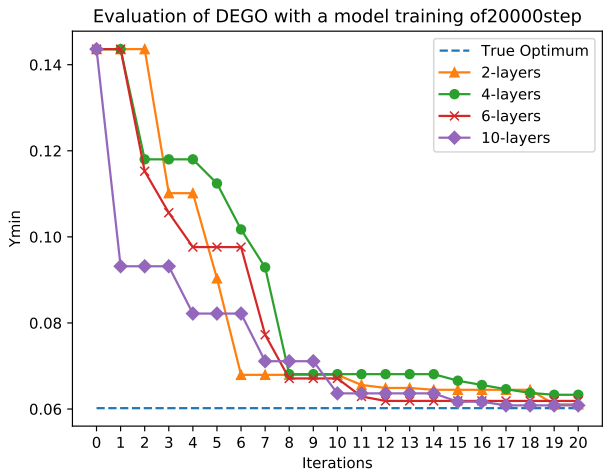
Constrained 2-D Problem

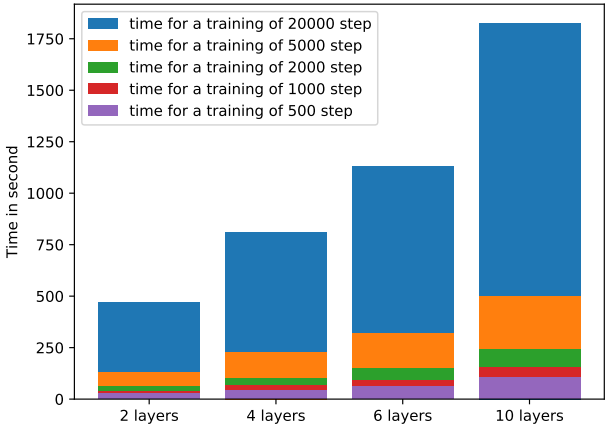
- Initial DoE : 15 points using Latin Hypercube Sampling
- Number of added points : 20
- EI criterion and Expected Violation criterion for the constraints with a threshold of 10^{-3} optimized with a differential evolution algorithm.

Constrained 2-D Problem



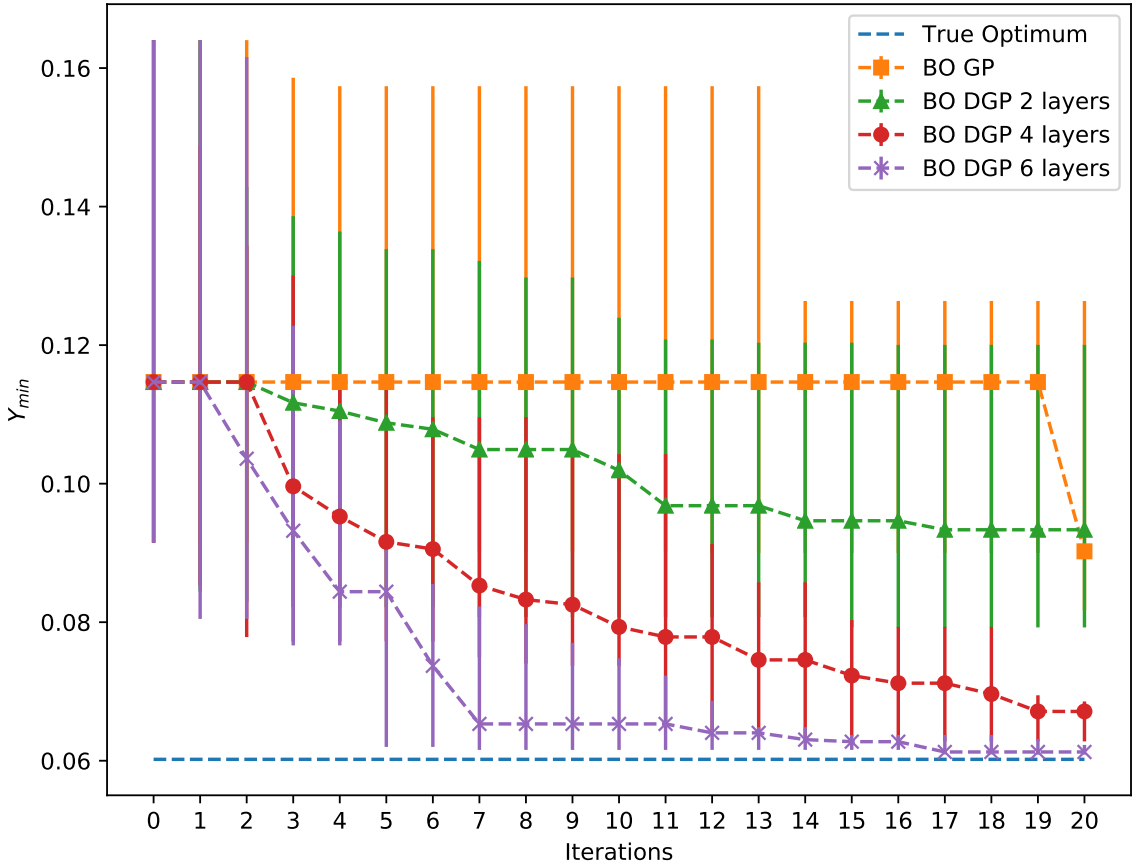






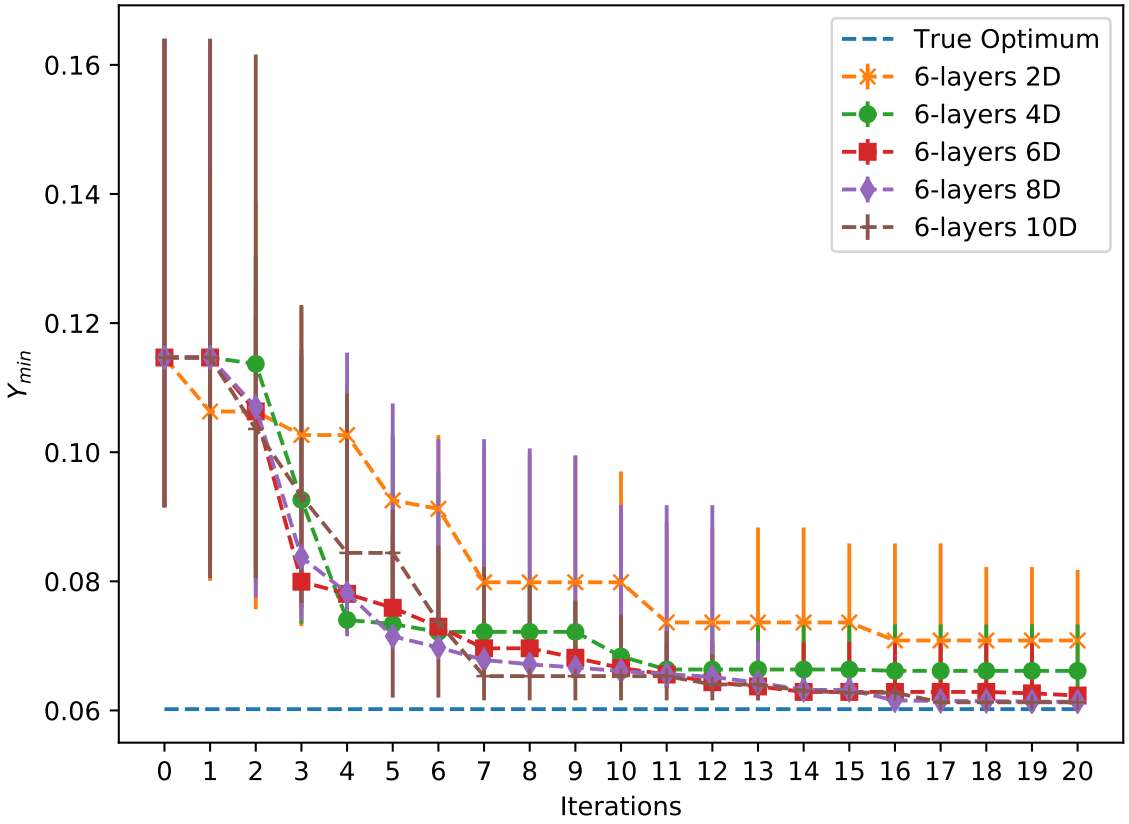
Plots of convergence of BO for different number of training steps for different architecture of DGP models (2,4,6,10 layers)

Constrained 2-D Problem



Plot of convergence of BO using different architecture of DGPs with 2000 training steps for 10 different initial DoE and a standard GP. The markers indicate the median of the minimum obtained while the errorbars indicate the first and the third quartiles.

Constrained 2-D Problem

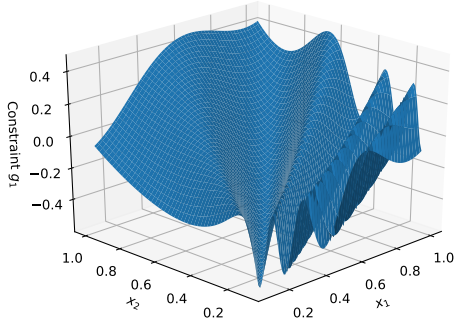


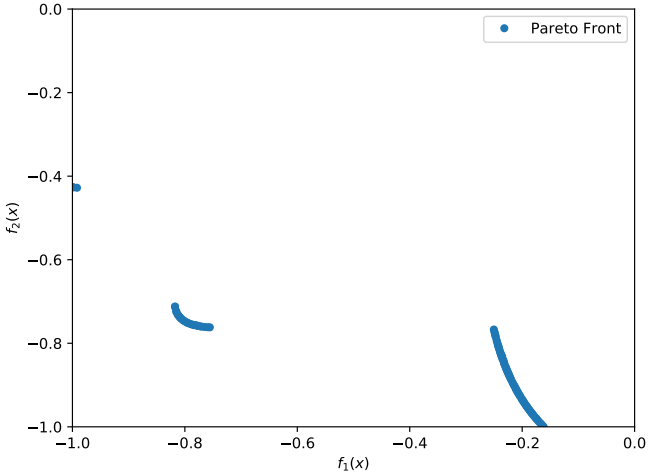
Plot of convergence of BO using different number of units (2,4,6,8,10) for a 6 layer DGP with 2000 training steps for 10 different initial DoE. The markers indicate the median of the minimum obtained while the errorbars indicate the first and the third quartiles.

Multi-objective optimization

Modified TNK Problem

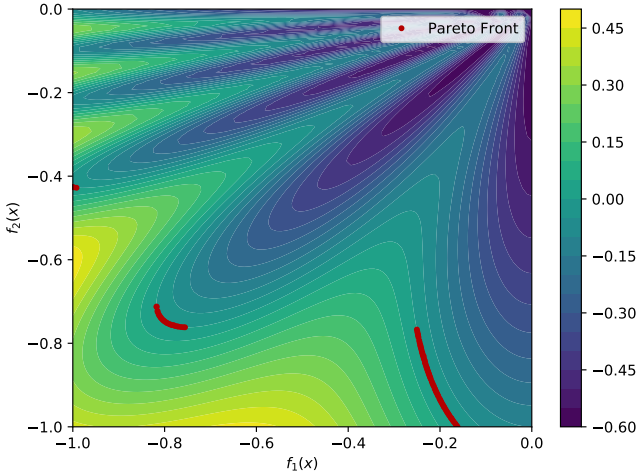
$$
\begin{array}{lll}
\text{Min} & f_1(\mathbf{x}) & = -x_1 \\
\text{Min} & f_2(\mathbf{x}) & = -x_2 \\
\text{s.t} & g_1(\mathbf{x}) & = 0.5(x_1^2 + x_2^2) - 0.2\cos(20\arctan(0.3\frac{x_1}{x_2})) - 0.4 \\
\text{with} & & 0 < x_1 \leq 1 \\
\text{and} & & 0 < x_2 \leq 1
\end{array}
$$





Exact Pareto front

Constraint fun

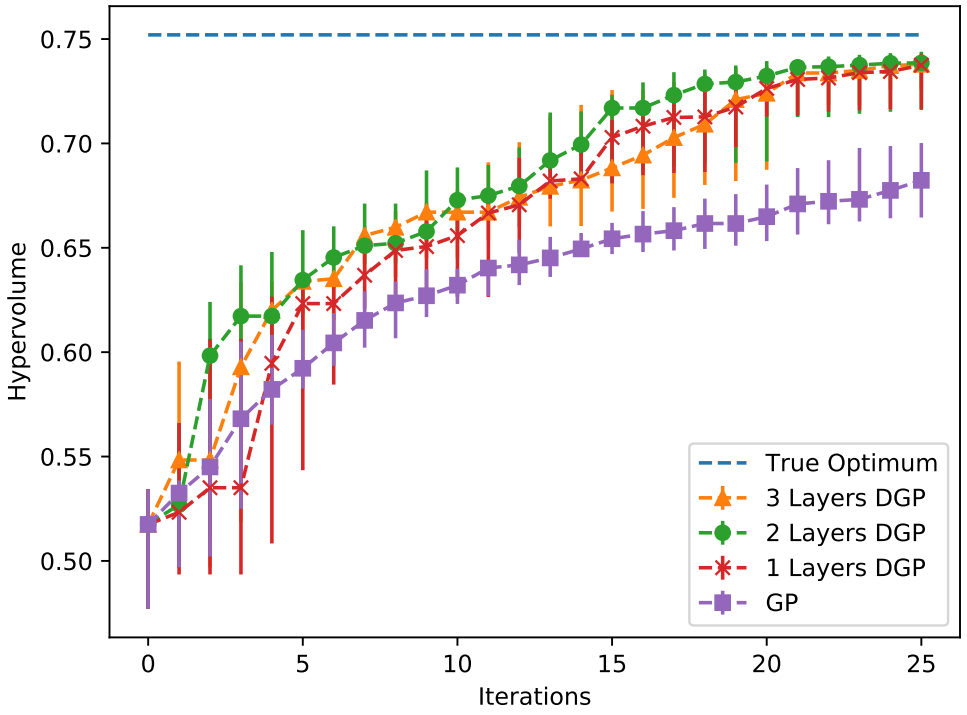


Exact Pareto front with constraint contour plot

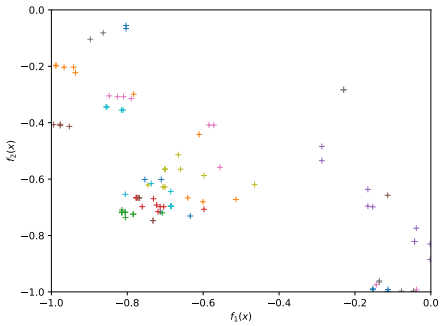
Multi-objective optimization

- 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.

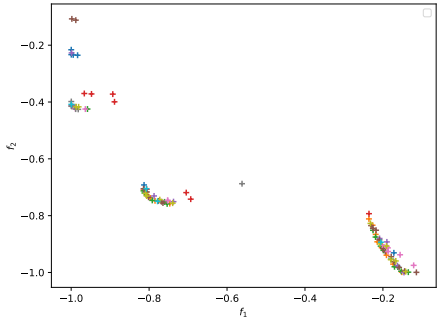
Multi-objective optimization



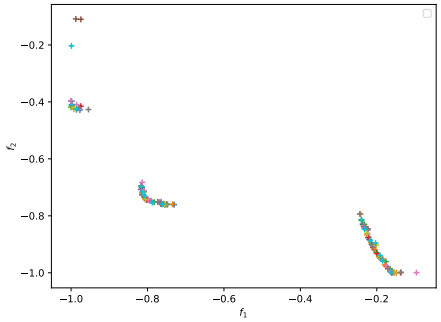
Convergence plot of BO with different architecture of DGPs and a regular GP. The markers indicate the median of the minimum obtained while the errorbars indicate the first and the third quartiles.



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).

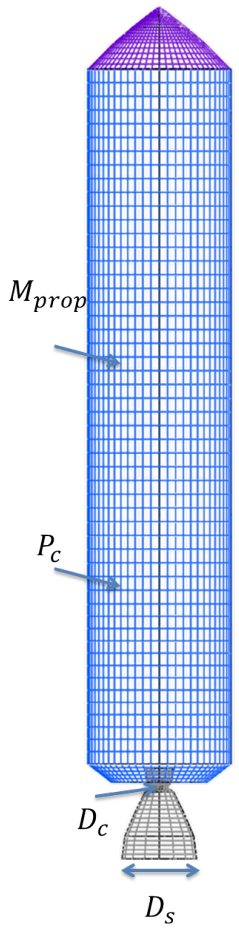
Minimize : $[GLOW(\mathbf{X}), -\Delta V(\mathbf{X})]$

According to : $\mathbf{X} = [\mathbf{M}_{prop}, \mathbf{P}_c, \mathbf{D}_c, \mathbf{D}_s]$

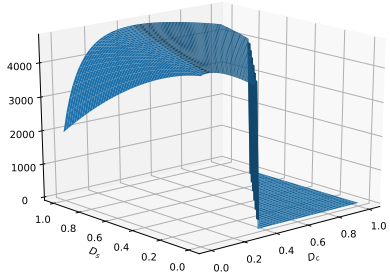
subject to : $\begin{cases} 1 \text{ Structural constraint} \\ 6 \text{ Geometrical constraints} \\ 1 \text{ Jet breakaway constraints} \end{cases}$

Design variables :

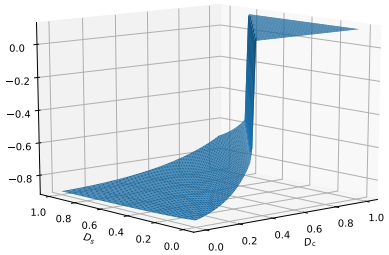
- Propellant mass : $5t \leq \mathbf{M}_{prop} \leq 15t$
- Combustion chamber pressure : $5Bar \leq \mathbf{P}_c \leq 100Bar$
- Throat nozzle diameter : $0.2m \leq \mathbf{D}_c \leq 1m$
- Nozzle exit diameter : $0.5m \leq \mathbf{D}_s \leq 1.2m$



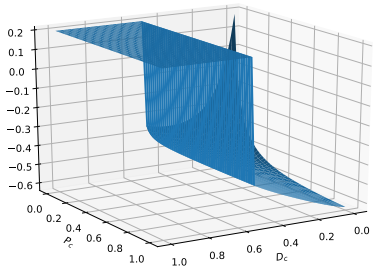
Problem : Solid-propellant booster



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

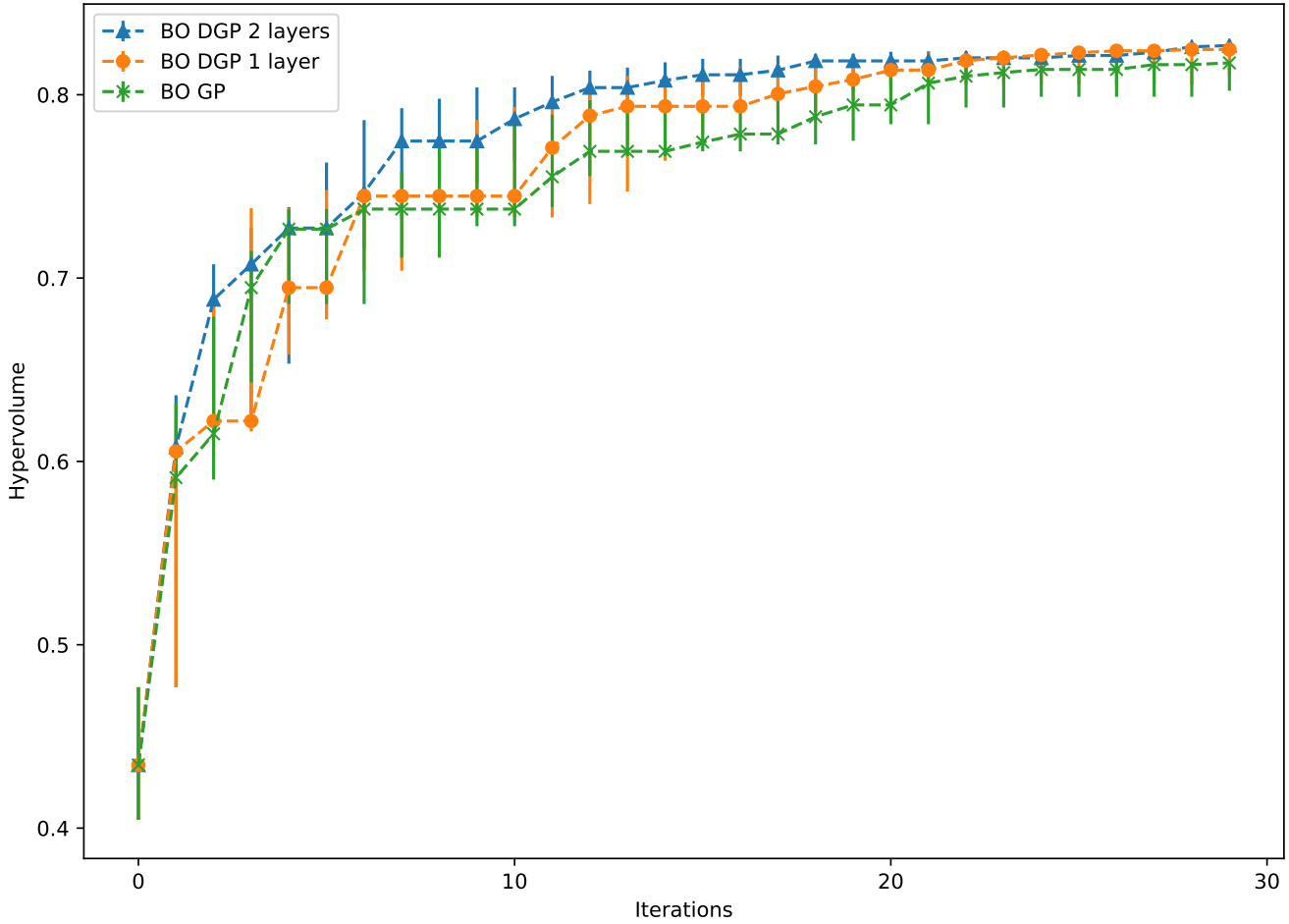


A sectional view of a constraint according to the diameters of the nozzle.



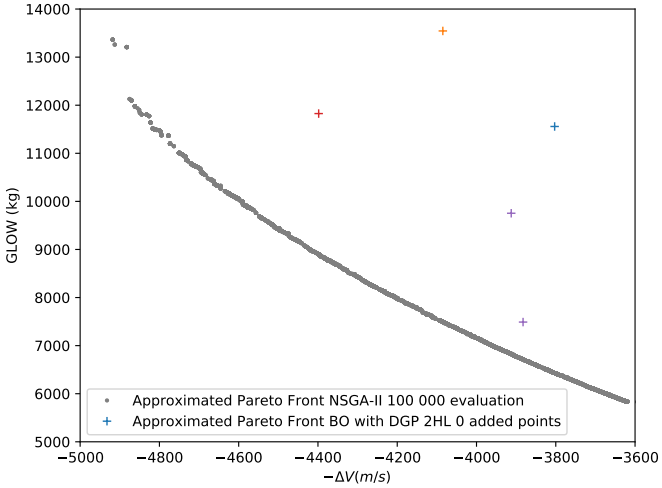
A sectional view of a constraint according to the throat nozzle diameter and the combustion chamber pressure.

Problem : Solid-propellant booster

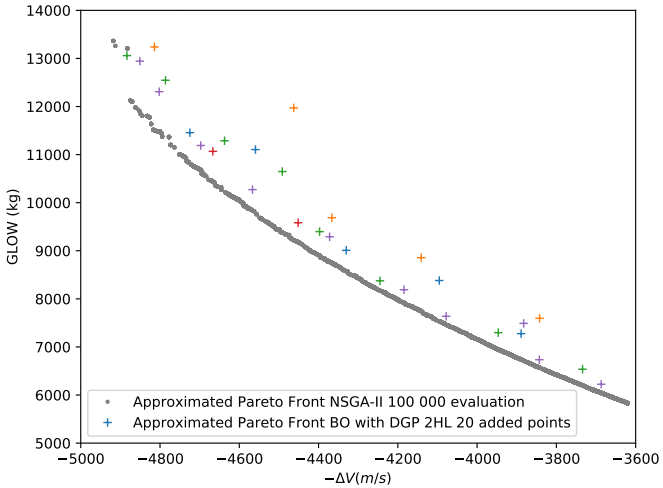


Convergence plot of BO with different architecture of DGPs and a regular GP.

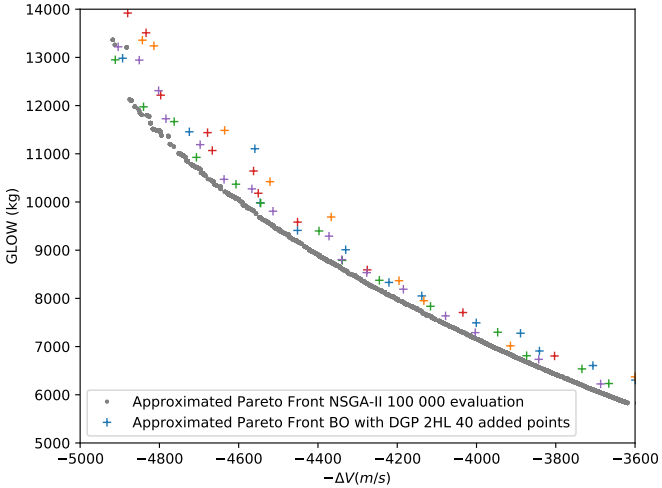
Problem : Solid-propellant booster



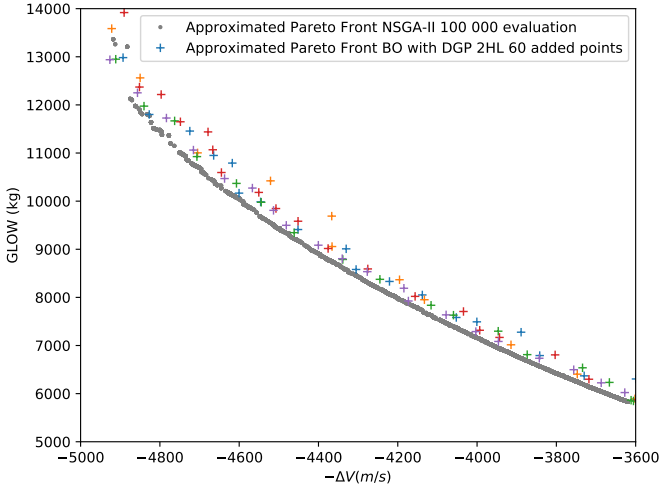
Initial DoE 40 initial points (0 Added points)



20 Added points



40 Added points



60 Added points

Conclusions

- Highlights of the challenges arising from the adaptation of DGPs to BO.
- Propositions to make the coupling of DGP and EGO possible.
- Numerical experimentations demonstrating the interest of this coupling and the challenges highlighted.

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

- Highlights of the challenges arising from the adaptation of DGPs to BO.
- Propositions to make the coupling of DGP and EGO possible.
- Numerical experimentations demonstrating the interest of this coupling and the challenges highlighted.

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

- Highlights of the challenges arising from the adaptation of DGPs to BO.
- Propositions to make the coupling of DGP and EGO possible.
- Numerical experimentations demonstrating the interest of this coupling and the challenges highlighted.

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

- Highlights of the challenges arising from the adaptation of DGPs to BO.
- Propositions to make the coupling of DGP and EGO possible.
- Numerical experimentations demonstrating the interest of this coupling and the challenges highlighted.

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.

List of communications and publications

- Hebbal, A., Brevault, L., Balesdent, M., Taibi, E. G., and Melab, N. Multi-Disciplinary Design Multi-Objective Optimization of Aerospace Vehicles using Surrogate Models à International Workshop on Optimization and Learning : Challenges and Applications (February 2018).
- Hebbal, A., Brevault, L., Balesdent, M., Taibi, E. G., and Melab, N. (2018, July). Efficient Global Optimization using Deep Gaussian Processes. In 2018 IEEE Congress on Evolutionary Computation (CEC) (pp. 1-8). IEEE
- Hebbal, A., Brevault, L., Balesdent, M., Talbi, E., and Melab, N., Multi-objective optimization using Deep Gaussian Processes : Application to Aerospace Vehicle Design, the AIAA 2019 Science and Technology Forum and Exposition (SciTech 2019).

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.

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.

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

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.

Bibliography II

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.