A Deep Gaussian Process based model for Multi-Objective optimization

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

The 13th International Conference on Multiple Objective Programming and Goal Programming. 2019

ONERA
THE FRENCH AEROSPACE LAB

Inria
INVENTEURS DU MONDE NUMÉRIQUE

Université de Lille
1 SCIENCES ET TECHNOLOGIES

Table of Contents

Context

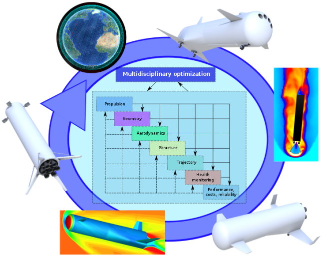
Resolution of a multi-objective optimization problem characterized by :

- Black box and computationally expensive functions,
- Correlated objectives.

Context

Resolution of a multi-objective optimization problem characterized by :

- Black box and computationally expensive functions,
- Correlated objectives.



Multi-disciplinary optimization of an aerospace vehicle

Context

Resolution of a multi-objective optimization problem characterized by :

- Black box and computationally expensive functions,
- Correlated objectives.

✗ Gradient based optimization approaches

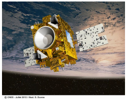
✗ Classic evolutionary algorithms

✓ Bayesian Optimization

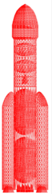
Context

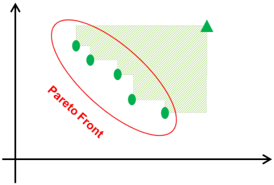
Resolution of a multi-objective optimization problem characterized by :

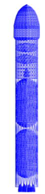
- Black box and computationally expensive functions,
- Correlated objectives.



Maximize f_1 : The payload value.





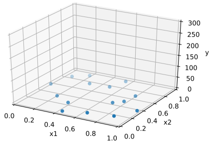


Minimize f_2 : The gross lift off-weight value.

MO-BO framework

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

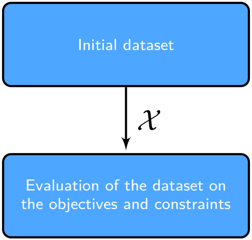
Initial dataset

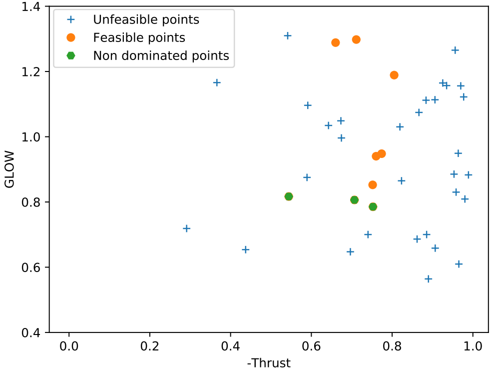


Design of Experiment depending on the dimension and the nature of the problem

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]

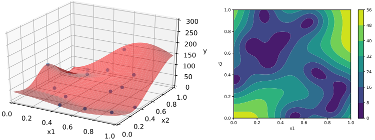
Initial dataset

$\downarrow \mathcal{X}$

Evaluation of the dataset on the objectives and constraints

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

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

$\mathcal{X}$

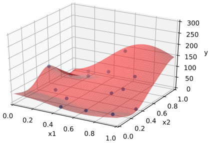
Evaluation of the dataset on the objectives and constraints

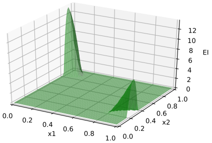
$\mathcal{Y}, \mathcal{G}$

Construction of the GP regression models for the objectives and constraints

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

Adding the most promising point(s)

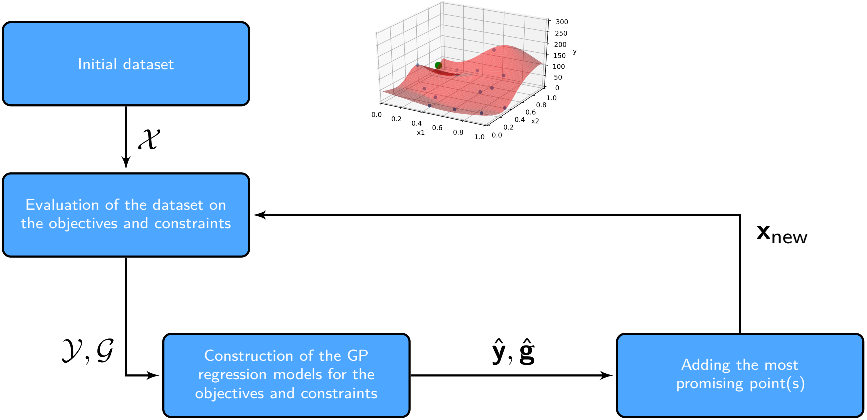




Optimization of a chosen **infill sampling criterion**

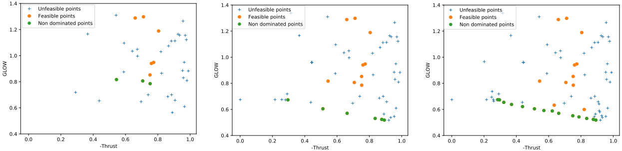
MO-BO framework

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



MO-BO framework

Initial DoE 30 added points 50 added points



Initial dataset

$\mathcal{X}$

Evaluation of the dataset on the objectives and constraints

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

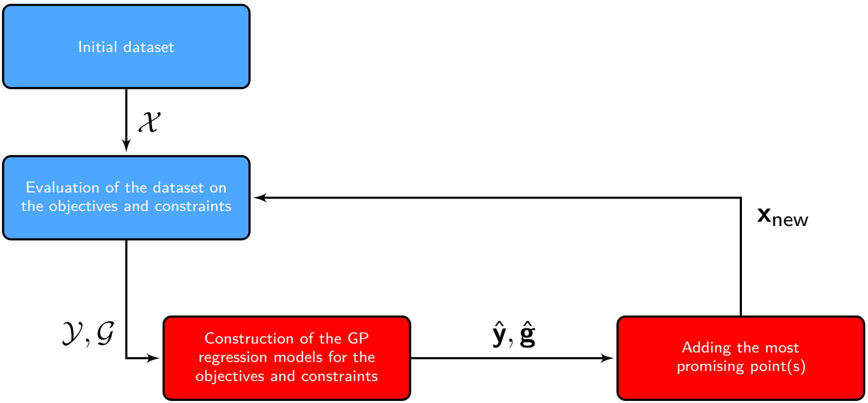
Construction of the GP regression models for the objectives and constraints

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

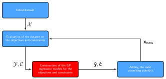
Adding the most promising point(s)

$\mathbf{x}_{\text{new}}$

MO-BO framework

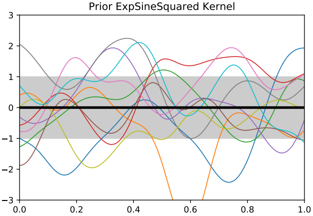


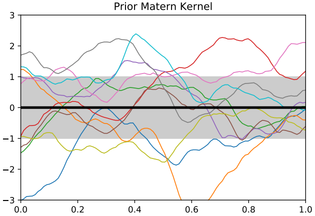
Gaussian Process Regression



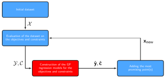
Gaussian process [Rasmussen, 2004]

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





Gaussian Process Regression



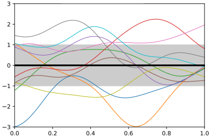
Gaussian process [Rasmussen, 2004]

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

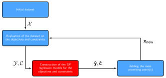
Automatic Relevance Determination (ARD) squared exponential kernel :

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

Prior Gaussian Process



Gaussian Process Regression



Gaussian process [Rasmussen, 2004]

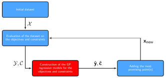
A Gaussian Process is used to describe a distribution over function. It is a collection of infinite random variables, **any finite number of which have a joint Gaussian distribution**.
It is defined by its mean function and covariance function (Kernel) : $f(.) \sim \mathcal{GP}(\mu(.), k^{\Theta}(.))$

Automatic Relevance Determination (ARD) squared exponential kernel :

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

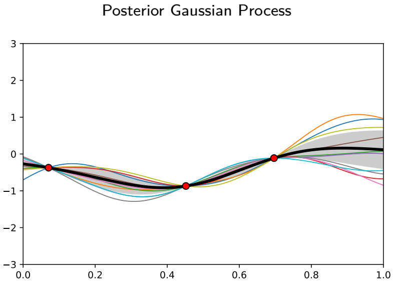
$$\text{Maximize w.r.t } \Theta \ : p(\mathbf{y}|\mathcal{X}) = \mathcal{N}(\mathbf{y}|0, \mathbf{K}^{\Theta}_{NN}\mathbf{I})$$

Gaussian Process Regression



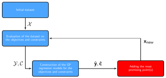
Gaussian process [Rasmussen, 2004]

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



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

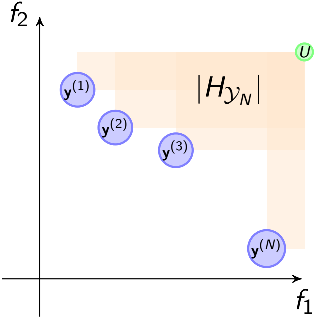
Infill Criteria



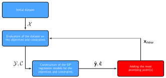
Hypervolume indicator

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

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



Infill Criteria

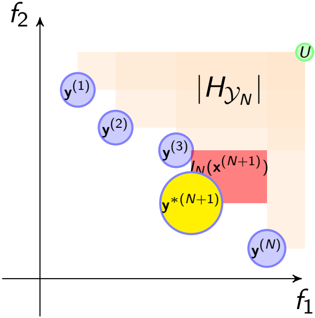


Hypervolume improvement

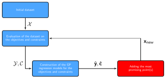
[Emmerich et al., 2006]

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

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



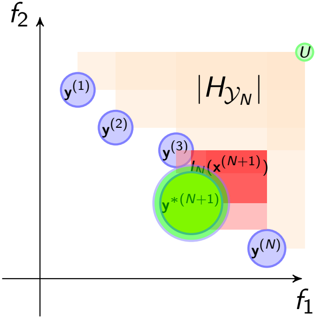
Infill Criteria



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

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



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

Multi-task GPs

- Classic MO-BO approaches use an independent GP for each objective $\rightarrow$ assumption of independency between the objectives.
- Multi-task GPs [Shah and Ghahramani, 2016] : exhibit correlation between functions by introducing a coregionalization matrix K^{coreg} :

$$\mathrm{Cov}\left(f_i(\mathbf{x}), f_j(\mathbf{x}')\right) = K^{coreg}_{ij} k(\mathbf{x}, \mathbf{x}')$$

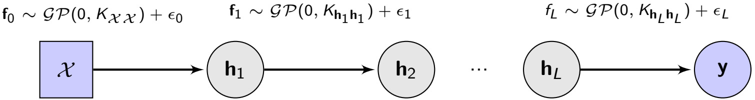
- Each function is marginally identically distributed up to a scaling factor.

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. They consider 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

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

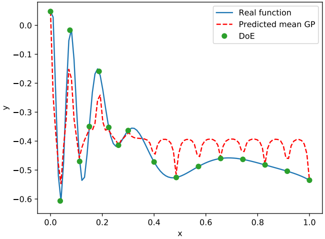
$$\mathcal{X} \longrightarrow \mathbf{h}_1 \longrightarrow \mathbf{h}_2 \quad \cdots \quad \mathbf{h}_L \longrightarrow \mathbf{y}$$

$\mathcal{X}$ A deterministic observed variable

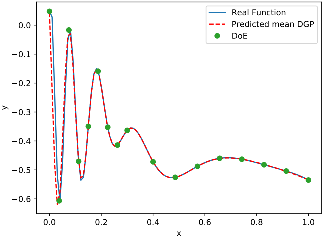
$\mathbf{h}_i$ A distribution with **Non**-observed instantiations

$\mathbf{y}$ A distribution with observed instantiations

Deep Gaussian Processes



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 [?]

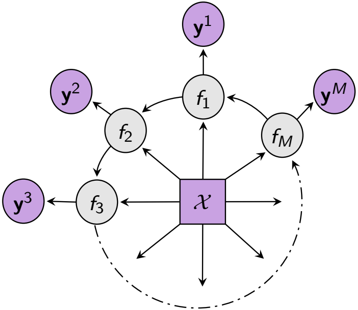


DGP approximation of a non-stationary 1-D function. The DGP model appropriately capture the two regions with different smoothness

MO-DGP model

MO-DGP model

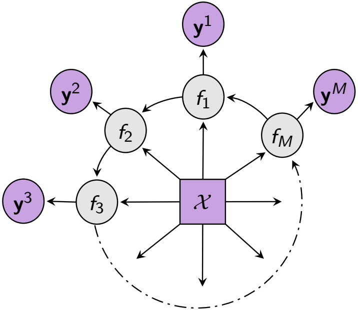
The MO-DGP model is a DGP network where each layer represents an objective. Moreover, a connection is made between the first and the last layer, creating a loop DGP. Propagating through the loop allows to take into account the different correlations between the objectives.



Specifications of the MO-DGP model

MO-DGP model

The MO-DGP model is a DGP network where each layer represents an objective. Moreover, a connection is made between the first and the last layer, creating a loop DGP. Propagating through the loop allows to take into account the different correlations between the objectives.



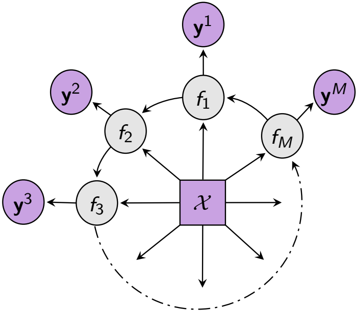
A multi-objective covariance function

$$k_l^{\rho}(\mathbf{x},\mathbf{x}')k_l^{f}\left(f_i^*(\mathbf{x}), f_{i+1}^*(\mathbf{x}')\right) + k_l^{\gamma}(\mathbf{x},\mathbf{x}')$$

Specifications of the MO-DGP model

MO-DGP model

The MO-DGP model is a DGP network where each layer represents an objective. Moreover, a connection is made between the first and the last layer, creating a loop DGP. Propagating through the loop allows to take into account the different correlations between the objectives.



The evidence lower bound

The evidence lower bound is derived using the sparse variational approximation of a GP inference [Salimbeni and Deisenroth, 2017] :

$$P\left(\mathbf{y}^1, \ldots, \mathbf{y}^M | \mathcal{X}\right) \geq ELBO$$

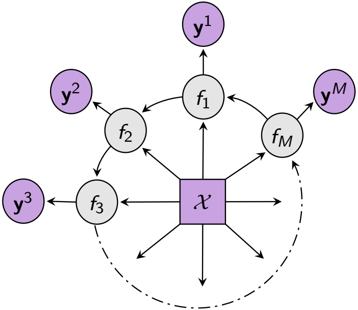
with

$$ELBO = \sum_{t=1}^{M} \sum_{i=1}^{N} \mathbb{E}_{q(f_t^{(i),t})} \left[\log p(y^{(i),t} | f_t^{(i),t})\right] - \sum_{l=1}^{M} KL\left[q(\mathbf{u}_l) || p(\mathbf{u}_l; Z_{l-1})\right]$$

Specifications of the MO-DGP model

MO-DGP model

The MO-DGP model is a DGP network where each layer represents an objective. Moreover, a connection is made between the first and the last layer, creating a loop DGP. Propagating through the loop allows to take into account the different correlations between the objectives.



Optimization of the ELBO

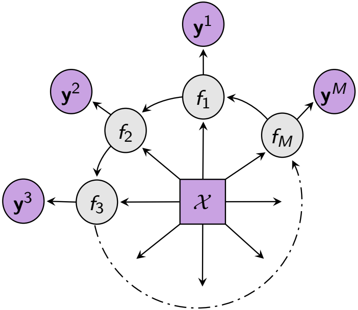
An optimization loop is considered :

- An optimization step with the ordinary gradient with respect to the deterministic parameters,
- An optimization step with the natural gradient with respect to the variational distributions.

specifications of the MO-DGP model

MO-DGP model

The MO-DGP model is a DGP network where each layer represents an objective. Moreover, a connection is made between the first and the last layer, creating a loop DGP. Propagating through the loop allows to take into account the different correlations between the objectives.



Complexity

The computational complexity of the model is :

$$\mathcal{O}(SMNK^2)$$

where : S is the number of samples used for the evaluation of the ELBO,
M is the number of objectives,
N is the number of training inputs,
K is the number of inducing inputs.

Analytic multi-objective problem

dtlz1a [Deb, 2001] is defined for $\mathbf{x} \in [0,1]^6$:

$$\left|\begin{array}{lll} \text{Min} & f_1(\mathbf{x}) & = -0.5x_1\left(1+g(x)\right) \\ \text{Min} & f_2(\mathbf{x}) & = -0.5(1-x_1)\left(1+g(x)\right) \\ \text{with} & g(\mathbf{x}) & = 100\left[5+\sum_{i=2}^{6}(x_i-0.5)^2+\cos\left(2\pi(x_i-0.5)\right)\right] \end{array}\right.$$

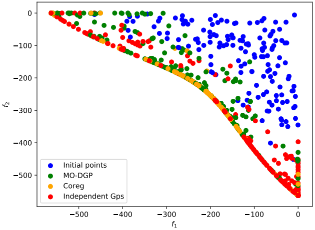
- Initial DoE : 30 initial points using a Latin Hypercube sampling,
- Added points : 60,
- EHVI criterion optimized with a parallel differential evolution algorithm.

Analytic multi-objective problem

dtlz1a [Deb, 2001] is defined for $\mathbf{x} \in [0,1]^6$:

$$\begin{array}{lll} \text{Min} & f_1(\mathbf{x}) & = -0.5x_1\,(1+g(x)) \\ \text{Min} & f_2(\mathbf{x}) & = -0.5(1-x_1)\,(1+g(x)) \\ \text{with} & g(\mathbf{x}) & = 100\left[5+\sum_{i=2}^{6}(x_i-0.5)^2+\cos\left(2\pi(x_i-0.5)\right)\right] \end{array}$$

- Initial DoE : 30 initial points using a Latin Hypercube sampling,
- Added points : 60,
- EHVI criterion optimized with a parallel differential evolution algorithm.



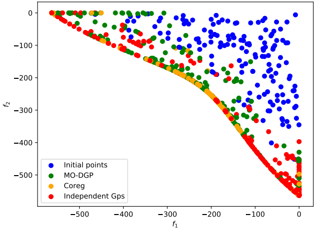
Approximated Pareto Fronts

Analytic multi-objective problem

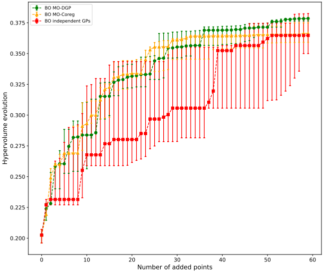
dtlz1a [Deb, 2001] is defined for $\mathbf{x} \in [0,1]^6$:

$$\begin{array}{|lll} \text{Min} & f_1(\mathbf{x}) & = -0.5x_1\,(1+g(x)) \\ \text{Min} & f_2(\mathbf{x}) & = -0.5(1-x_1)\,(1+g(x)) \\ \text{with} & g(\mathbf{x}) & = 100\left[5+\sum_{i=2}^{6}(x_i-0.5)^2+\cos\left(2\pi(x_i-0.5)\right)\right] \end{array}$$

- Initial DoE : 30 initial points using a Latin Hypercube sampling,
- Added points : 60,
- EHVI criterion optimized with a parallel differential evolution algorithm.

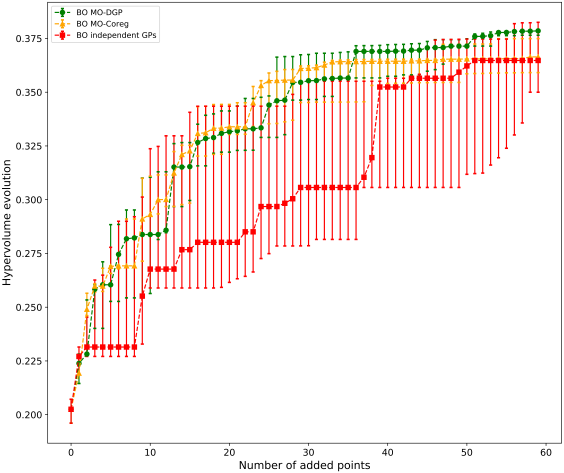


Approximated Pareto Fronts



Hypervolume evolution according to the added points

Analytic multi-objective problem



Hypervolume evolution according to the number of added points

Analytic multi-objective problem

Kursawe [Kursawe, 1990] is defined for $\mathbf{x} \in [-5, 5]^3$:

$$\left| \begin{array}{lll} \text{Min} & f_1(\mathbf{x}) & = \sum_{i=1}^{2} \left[-10 \exp\left(-0.5 \sqrt{x_i - 2 + x_{i+1}^2} \right) \right] \\ \text{Min} & f_2(\mathbf{x}) & = \sum_{i=1}^{3} \left[|x_i|^{0.8} + 5 \sin(x_i^3) \right] \end{array} \right.$$

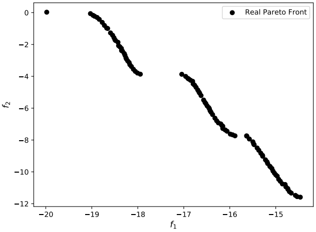
- Initial DoE : 15 initial points using a Latin Hypercube sampling,
- Added points : 20,
- EHVI criterion optimized with a parallel differential evolution algorithm.

Analytic multi-objective problem

Kursawe [Kursawe, 1990] is defined for $\mathbf{x} \in [-5, 5]^3$:

$$\left|\begin{array}{lll} \text{Min} & f_1(\mathbf{x}) & = \sum_{i=1}^{2} \left[-10 \exp\left(-0.5\sqrt{x_i - 2 + x_{i+1}^2}\right)\right] \\ \text{Min} & f_2(\mathbf{x}) & = \sum_{i=1}^{3} \left[|x_i|^{0.8} + 5\sin(x_i^3)\right] \end{array}\right.$$

- Initial DoE : 15 initial points using a Latin Hypercube sampling,
- Added points : 20,
- EHVI criterion optimized with a parallel differential evolution algorithm.



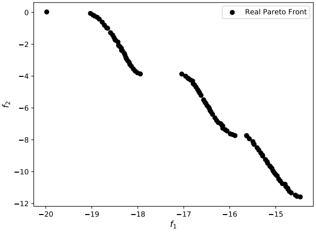
Exact Pareto Front

Analytic multi-objective problem

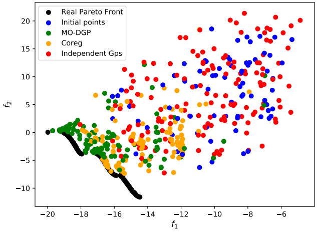
Kursawe [Kursawe, 1990] is defined for $\mathbf{x} \in [-5, 5]^3$:

$$\left| \begin{array}{lll} \text{Min} & f_1(\mathbf{x}) & = \sum_{i=1}^{2} \left[-10 \exp\left(-0.5 \sqrt{x_i - 2 + x_{i+1}^2} \right) \right] \\ \text{Min} & f_2(\mathbf{x}) & = \sum_{i=1}^{3} \left[|x_i|^{0.8} + 5 \sin(x_i^3) \right] \end{array} \right.$$

- Initial DoE : 15 initial points using a Latin Hypercube sampling,
- Added points : 20,
- EHVI criterion optimized with a parallel differential evolution algorithm.

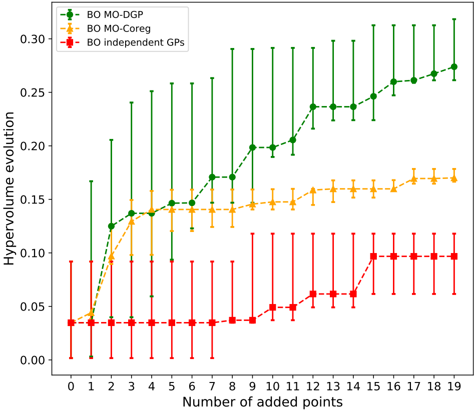


Exact Pareto Front



Solutions of the different MO-BO approaches.

Analytic multi-objective problem



Hypervolume evolution according to the number of added points

Analytic multi-objective problem

Kursawe 10D is defined for $\mathbf{x} \in [-5, 5]^10$:

$$\left| \begin{array}{lll} \text{Min} & f_1(\mathbf{x}) & = \sum_{i=1}^{9} \left[-10 \exp \left(-0.5 \sqrt{x_i - 2 + x_{i+1}^2} \right) \right] \\ \text{Min} & f_2(\mathbf{x}) & = \sum_{i=1}^{10} \left[|x_i|^{0.8} + 5 \sin(x_i^3) \right] \end{array} \right.$$

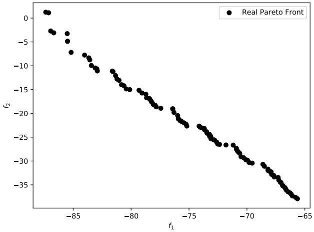
- Initial DoE : 50 initial points using a Latin Hypercube sampling,
- Added points : 85,
- EHVI criterion optimized with a parallel differential evolution algorithm.

Analytic multi-objective problem

Kursawe 10D is defined for $\mathbf{x} \in [-5, 5]^10$:

$$\left|\begin{array}{lll} \text{Min} & f_1(\mathbf{x}) & = \sum_{i=1}^{9} \left[-10 \exp\left(-0.5 \sqrt{x_i - 2 + x_{i+1}^2}\right)\right] \\ \text{Min} & f_2(\mathbf{x}) & = \sum_{i=1}^{10} \left[|x_i|^{0.8} + 5 \sin(x_i^3)\right] \end{array}\right.$$

- Initial DoE : 50 initial points using a Latin Hypercube sampling,
- Added points : 85,
- EHVI criterion optimized with a parallel differential evolution algorithm.



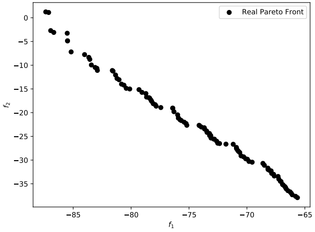
Exact Pareto Front

Analytic multi-objective problem

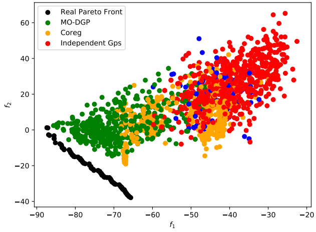
Kursawe 10D is defined for $\mathbf{x} \in [-5, 5]^{1}0$:

$$\left|\begin{array}{lll} \text{Min} & f_1(\mathbf{x}) & = \sum_{i=1}^{9} \left[-10 \exp\left(-0.5 \sqrt{x_i - 2 + x_{i+1}^2} \right) \right] \\ \text{Min} & f_2(\mathbf{x}) & = \sum_{i=1}^{10} \left[|x_i|^{0.8} + 5 \sin(x_i^3) \right] \end{array}\right.$$

- Initial DoE : 50 initial points using a Latin Hypercube sampling,
- Added points : 85,
- EHVI criterion optimized with a parallel differential evolution algorithm.

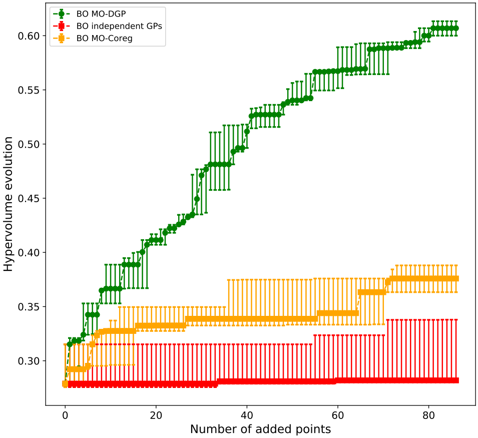


Exact Pareto Front



Solutions of the different MO-BO approaches.

Analytic multi-objective problem



Hypervolume evolution according to the added points

Summary

Comparison of the **average** hypervolume obtained by different approaches and for the same number of evaluations :

Algorithms	dtlz1a (90 eval)	Kursawe 3D (35 eval)	Kursawe 10D (140 eval)
MO-DGP	**0.3791**	**0.2976**	**0.6084**
MO-Coreg	0.3659	0.19429	0.3846
MO-GP	0.3621	0.13361	0.30764

Comparison of the hypervolume **standard deviation** obtained by different approaches and for the same number of evaluations :

Algorithms	dtlz1a (90 eval)	Kursawe 3D (35 eval)	Kursawe 10D (140 eval)
MO-DGP	**0.00407**	0.05213	**0.01127**
MO-Coreg	0.00983	**0.05041**	0.04088
MO-GP	0.02197	0.06034	0.03268

Conclusions

- Proposition of a Deep Gaussian Process multi-objective model, taking into account the correlations between objectives,
- Experimentations on analytical functions confirm the efficiency of the proposed model compared to coregionalization GP approach and independent GPs.
- MO-DGP is a more efficient model but the complexity to train the model is more important that regular GPs. Hence, it is more interesting for computationally expensive problems.

Future works :

- Derivation of an EHVI taking into account the correlations exhibited by the MO-DGP model,
- Application of the model to problems with over three objectives,
- Application to a real multi-objective aerospace problem

Conclusions

- Proposition of a Deep Gaussian Process multi-objective model, taking into account the correlations between objectives,
- Experimentations on analytical functions confirm the efficiency of the proposed model compared to coregionalization GP approach and independent GPs.
- MO-DGP is a more efficient model but the complexity to train the model is more important that regular GPs. Hence, it is more interesting for computationally expensive problems.

Future works :

- Derivation of an EHVI taking into account the correlations exhibited by the MO-DGP model,
- Application of the model to problems with over three objectives,
- Application to a real multi-objective aerospace problem

Conclusions

- Proposition of a Deep Gaussian Process multi-objective model, taking into account the correlations between objectives,
- Experimentations on analytical functions confirm the efficiency of the proposed model compared to coregionalization GP approach and independent GPs.
- MO-DGP is a more efficient model but the complexity to train the model is more important that regular GPs. Hence, it is more interesting for computationally expensive problems.

Future works :

- Derivation of an EHVI taking into account the correlations exhibited by the MO-DGP model,
- Application of the model to problems with over three objectives,
- Application to a real multi-objective aerospace problem

Conclusions

- Proposition of a Deep Gaussian Process multi-objective model, taking into account the correlations between objectives,
- Experimentations on analytical functions confirm the efficiency of the proposed model compared to coregionalization GP approach and independent GPs.
- MO-DGP is a more efficient model but the complexity to train the model is more important that regular GPs. Hence, it is more interesting for computationally expensive problems.

Future works :

- Derivation of an EHVI taking into account the correlations exhibited by the MO-DGP model,
- Application of the model to problems with over three objectives,
- Application to a real multi-objective aerospace problem

Conclusions

Thank you for your attention !

Bibliography I

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

Deb, K. (2001).
Multi-objective optimization using evolutionary algorithms, volume 16.
John Wiley & Sons.

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.

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.

Kursawe, F. (1990).
A variant of evolution strategies for vector optimization.
In *International Conference on Parallel Problem Solving from Nature*, pages 193–197. Springer.

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

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

Bibliography II

Shah, A. and Ghahramani, Z. (2016).
Pareto frontier learning with expensive correlated objectives.
In *International Conference on Machine Learning*, pages 1919–1927.