

Which Gaussian Process for Bayesian Optimization ?

Rodolphe Le Riche*,

David Gaudrie, Victor Picheny, Youssef Diouane,
Adrien Spagnol, Sébastien Da Veiga

* CNRS at LIMOS (Mines Saint Etienne, UCA) France

16-18 May 2022
2022 Optimization Days

Inspector BO



stands for

Bayesian Optimization

(old Lithuanian)



Context: optimization of costly functions

$$\min_{x \in \mathcal{S}} f(x)$$

\mathcal{S} : search space, continuous, discrete, mixed, others (graphs?).
Default $\mathcal{S} \in \mathbb{R}^d$ (hyper-rectangle). d is the dimension.

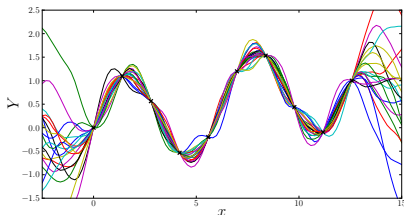
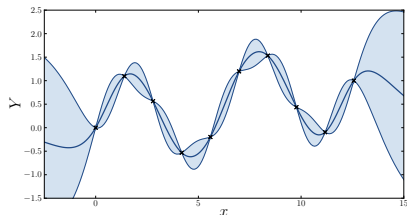
Costly: one call to f takes more CPU than the rest of the optimization algorithm. Examples: nonlinear partial differential equations (finite elements), training of a neural network, real experiment ...

To save calls to f , build a model of it based on previous evaluations and rely on it whenever possible \rightarrow metamodel / surrogate based optimization. **Gaussian process as metamodel : Bayesian Optimization**



n the head
of inspector
BO

Gaussian Process Regression (kriging)



$Y(x)$ is $\mathcal{N}(\mu(x), k(x, x'))$

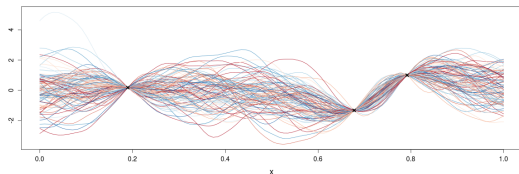
$Y(x) | Y(\mathbb{X}) = \mathbb{F}$ is also Gaussian, interpolating and depends on $k(., .)$ and $\mu(.)$ through parameters θ .

Ex: $k(x, x') = \sigma^2 \exp\left(-\sum_{i=1}^d \frac{(x_i - x'_i)^2}{2\theta_i^2}\right)$.

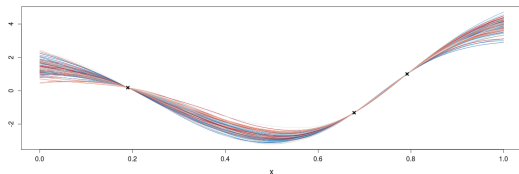
Learn the GP typically by max. (log) likelihood,
 $\theta^* = \arg \max_{\theta} LL(\theta; \mathbb{F})$.

Gaussian Process Regression (kriging)

θ 's as length scales, $k(x, x') = \sigma^2 \prod_{i=1}^d \text{correlation}_i \left(\frac{|x_i - x'_i|}{\theta_i} \right)$



$\theta = 0.1$



$\theta = 0.5$

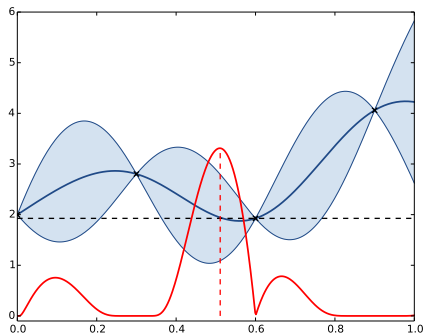
(Matérn kernel, $\sigma = 1$)

The Expected Improvement

Measure of progress: the improvement,

$$I(x) = \max(0, (\min(\mathbb{F}) - Y(x) \mid Y(\mathbb{X})=\mathbb{F})).$$

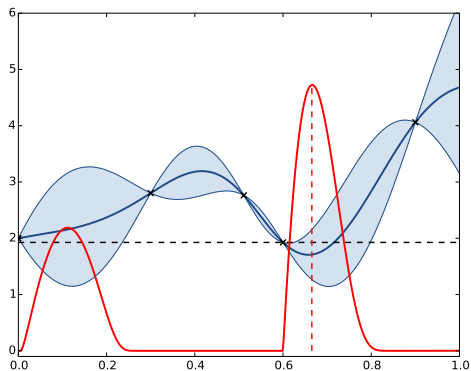
Acquisition criterion: $\mathbb{E}I(x)$, to maximize at each iteration



Expected Improvement

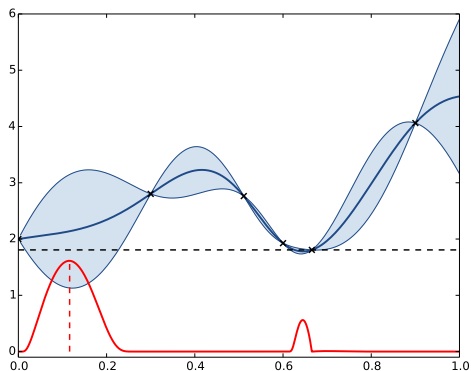
$$x^{t+1} = \arg \max_{x \in \mathcal{S}} \text{EI}(x)$$

Let's see how it works... iteration 1



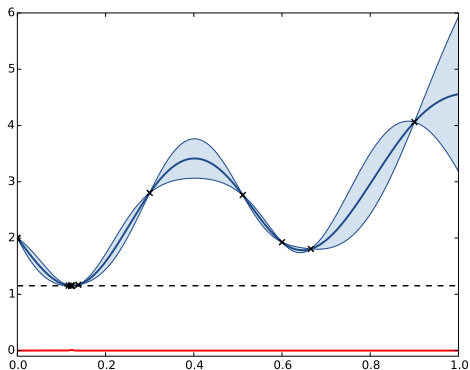
Expected Improvement

$$x^{t+1} = \arg \max_{x \in \mathcal{S}} \text{EI}(x) \dots \text{iteration 2}$$



Expected Improvement

$$x^{t+1} = \arg \max_{x \in \mathcal{S}} \text{EI}(x) \dots \text{iteration 5}$$



BO algorithm skeleton

[Mockus, 1975, Jones et al., 1998, Frazier, 2018]

- 1 make an initial design of experiments \mathbb{X} and calculate the associated \mathbb{F} , $t = \text{length}(\mathbb{F})$
- 2 build a Gaussian Proc. from (\mathbb{X}, \mathbb{F}) (max. log likelihood $\rightarrow \theta$)
- 3 $x^{t+1} = \arg \max_{x \in \mathcal{S}} \text{EI}(x)$
- 4 calculate $\mathbb{F}_{t+1} = f(\mathbb{X}_{t+1})$, increment t
- 5 stop ($t > t^{\max}, \dots$) or go to 2.

Note the 2 internal optimization problems, one in \mathcal{S} (d dimensions), one in the number of parameters of the GP (typically $\mathcal{O}(d)$).

BO's degrees of freedom

BO a mature algorithm?

Opensource implementations : Spearmint, DiceOptim, BayesOpt, SMAC, GPyOpt, GPflowOpt, RoBO, STK, Botorch, SMT, ...

But **still many open questions**. Of course: it is quite generic.

In [Le Riche and Picheny, 2021], we empirically studied

- Initial DoE size
- Trend function
- Kernel
- EI optimization
- Modifying the exploration/intensification tradeoff
- Non-linear transformations of the input and output

Testing BO with COCO I

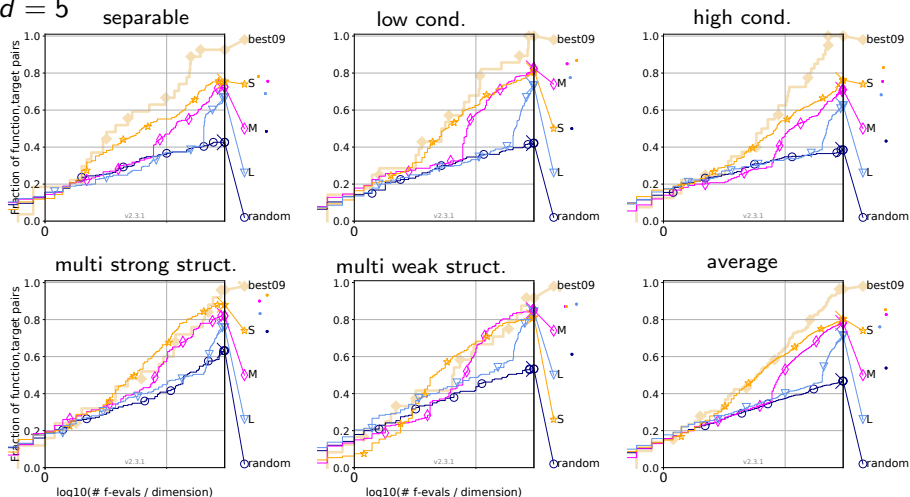
- COCO : COmparing Continuous Optimizers [Hansen et al., 2016] with 24 functions of the BBOB noiseless suite [Hansen et al., 2010].
- Functions structured in 5 groups: separable, low or moderate conditioning, unimodal with high conditioning, multimodal with adequate structure, multimodal with weak global structure.

Testing BO with COCO II

- For each version of the algo : 15 repetitions of runs of length $30 \times d$ ($=2,3,5,10$) \rightarrow 360 optimizations per dimension, 432000 maximizations solved, millions of covariance matrices inversions.
- Default algorithm: medium DoE size ($7.5 \times d$), Matérn 5/2 kernel, constant trend and multi-start BFGS for EI optimization.

Size of initial DoE

$d = 5$



Small DoE ($d + 4$) \geq Medium Doe ($7.5 \times d$) $>$ Large DoE ($20 \times d$)

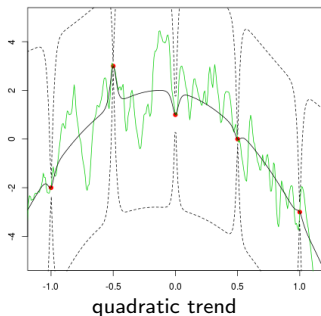
Effect of trend

$$Y(x) \sim \mathcal{N}(\mu(x), k(x, x'))$$

Compare trends $\mu(\cdot)$:
constant, linear and quadratic

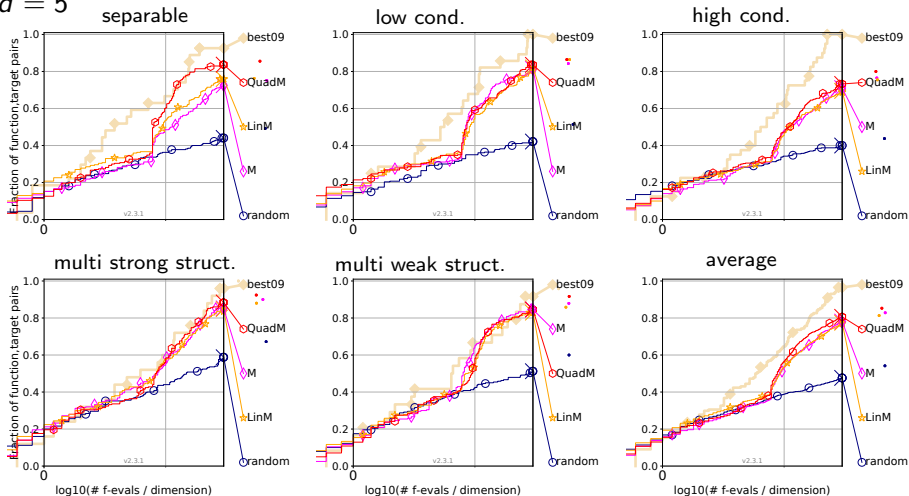
$$\mu(x) = \beta_0 + \sum_{i=1}^d \beta_i x_i + \sum_{i=1}^d \beta_{d+1+i} x_i^2$$

Default : constant



Effect of trend

$d = 5$



Quadratic trend never harms, and helps on separable functions (which includes the quadratic sphere and ellipse).

Other lessons from COCO tests

Other observations from [Le Riche and Picheny, 2021]:

- Kernel: Matérn 5/2 is a good default
- EI internal optimization: important, do it well
- Modifying the exploration/intensification tradeoff: sometimes (1/5) minimizing the kriging average is marginally better.
- Non-linear transformations of the input and output: not beneficial.

Most importantly: the effect of dimension ...

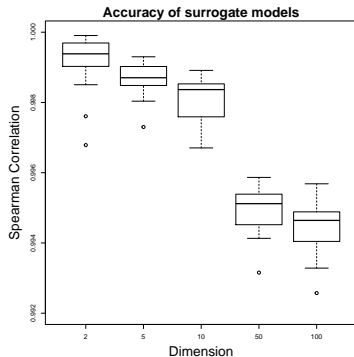
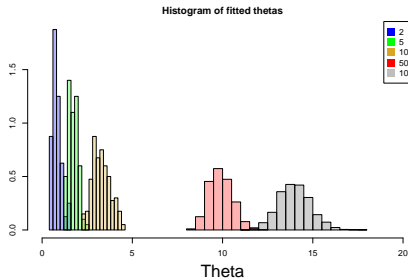
Effect of dimension

General curse of dimensionality :

- geometrical increase in number of points (N^d) to keep the distance between them constant
- a non-informative geometry of the points w.r.t. euclidean distance : by Central Limit Th. applied to $x \sim \mathcal{U}[-1, 1]^d$, as $d \nearrow$, the mass of the points is on a sphere of radius $\sqrt{d/3}$, inter-points distances tend to a constant $\sqrt{2d/3}$

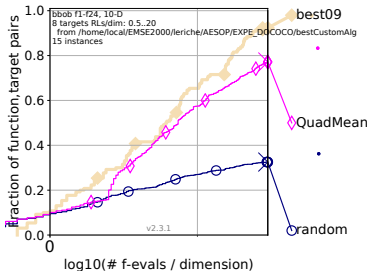
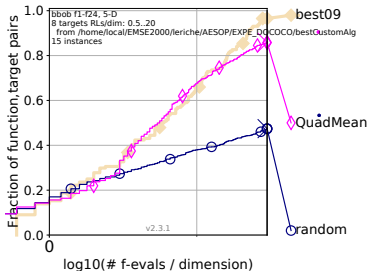
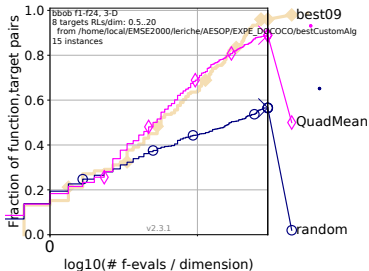
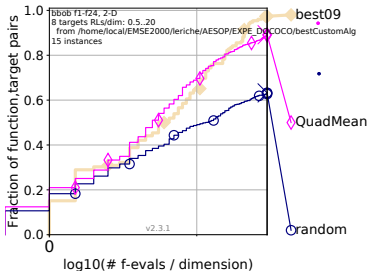
Effect of dimension on a Gaussian Process

the θ 's of max. log likelihood \nearrow in \sqrt{d} , marginal loss of accuracy



(sphere function, tensor product Matérn, $10 \times d$ points)

Effect of dimension on Bayesian Optimization



Bayesian optimization and dimension

Bayesian optimizers are competitive at low number of function evaluations but they lose this advantage with dimension.

Loss of GP accuracy? EI sample too often at boundary?

Recent efforts:

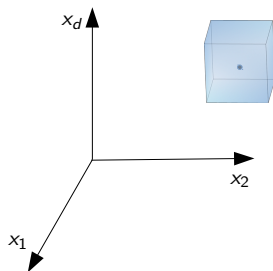
- search locally around good points (trust regions).
- search in low dimensional linear subspaces.

“search” has 2 ingredients :

build a metamodel + max. acquisition criterion (EI).

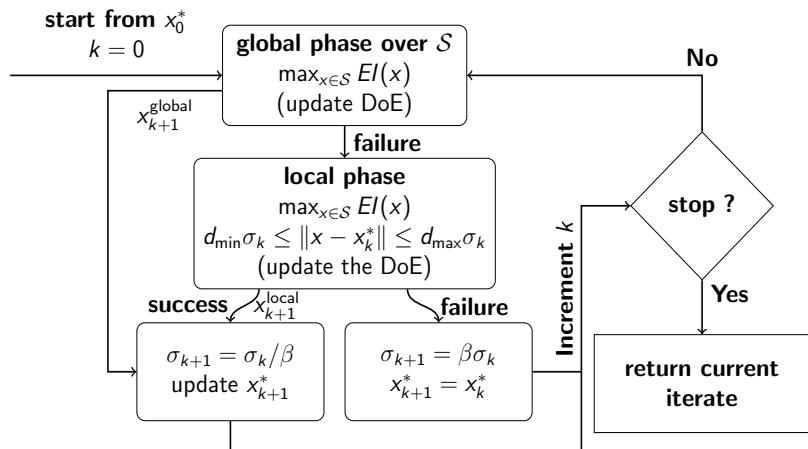
BO and trust regions

Principle: counteract the effect of increasing dimension (volume) by restricting the search to a smaller (controlled) trust region.



- TRIKE, Trust-Region Implementation in Kriging-based optimization with Expected Improvement, [Regis, 2016].
- TURBO, a TrUst-Region BO solver, [Eriksson et al., 2019].
- TREGO, a Trust-Region framework for EGO, [Diouane et al., 2021] : mix searches inside (local) and outside (global) the trust region.

TREGO algorithm



Parameters : $\sigma_0, \beta < 1$

Sufficient decrease condition for success of the local phase,

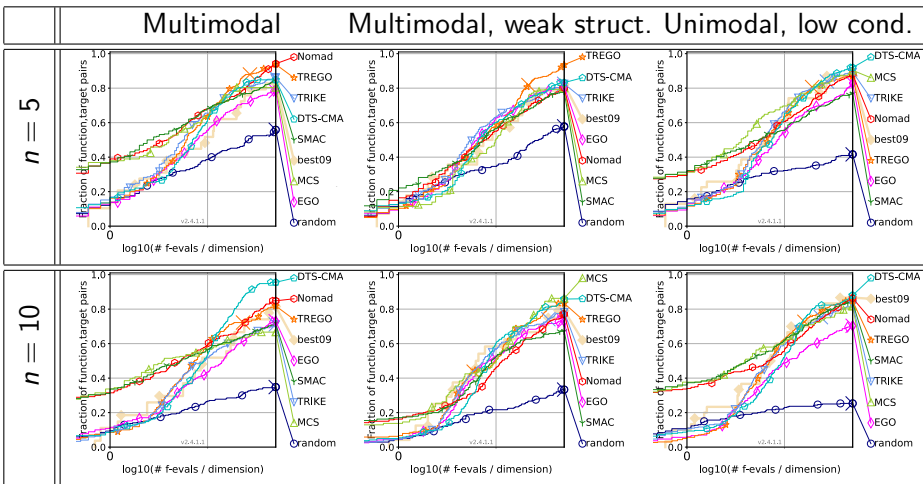
$$f(x_{k+1}^{\text{local}}) \leq f(x_k^*) - 10^{-4} \sigma_k^2$$

TREGO properties

From [Diouane et al., 2021],

- TREGO iterates converge to a local minimum : by assuming f is bounded below, Lipschitz continuous near the point of convergence, and by considering a subsequence of the local iterates. No assumption on GP or x_0^* .
- Empirical COCO tests:
 - more local than global steps (4 to 1) is beneficial
 - TREGO is robust to the values of σ_0 and β
 - A local GP was thought an asset for non stationary functions. But it is a drawback on badly conditioned functions. Not kept.

TREGO performance

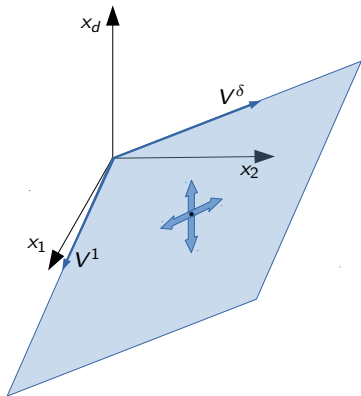


Trust regions solve BO's oversampling of the boundaries in high-dim. while helping on unimodal functions (not the natural target for BO).

Improving Bayesian Optimization in high dimension:

- search locally around good points (trust regions) \leftarrow TREGO
- search in low dimensional linear subspaces

$$\min_{x \in \mathcal{S} \subset \mathbb{R}^d} f(x) \Rightarrow \min_{\alpha \in \mathbb{R}^\delta} f(\text{Proj}_{\mathcal{S}}(V\alpha + b)) \quad , \quad \delta \ll d$$



Algorithm design:
choose V , b , $\text{Proj}_{\mathcal{S}}$

BO in a linear subspace

- Variable selection:

- $V = \begin{bmatrix} \dots & & & \\ & 0 & & \\ \dots & 1 & \dots & \\ & 0 & & \\ & \dots & & \end{bmatrix}$, $b =$ defaults for unselected variables.

- In [Spagnol et al., 2019], selection based on distance ($p(x_i), p(x_i | f(x) < T)$),
cf. Sébastien Da Veiga's talk at JOPT2022.

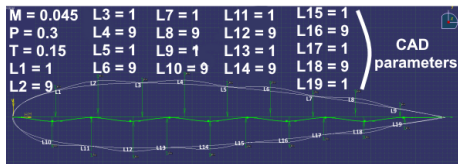
- (R)REMBO & improvements, Random EMbedding Bayesian Optimization, [Wang et al., 2016, Binois et al., 2020].
- Choice of V by Partial Least Squares, SEGOKPLS [Amine Bouhlel et al., 2018] (internal optim in high dimensions), EGORSE (EGO coupled with Random and Supervised Embeddings [Priem, 2020]).
- Choice of V by the active subspace method [Li et al., 2019].

Costly shape optimization: airfoil

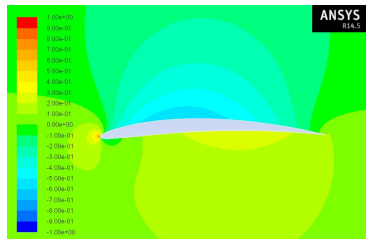
An example of linear embedding for Bayesian Optimization.

Minimize the drag of an airfoil, from [Gaudrie et al., 2020],

$$\min_{\phi \in \mathcal{S}} f(\phi) \quad , \quad \mathcal{S} \text{ "infinite" dimensional space of feasible shapes}$$



CAD shape ϕ generation,
not costly



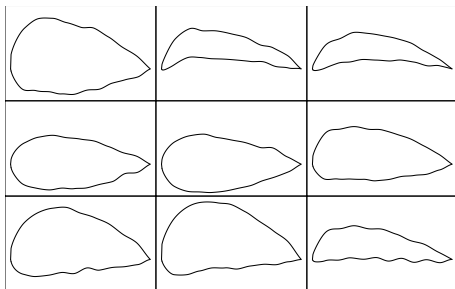
Contours of Pressure Coefficient

Feb 20, 2017
ANSYS Fluent 14.5 (2d, pbns, r6e)

Navier-Stokes resolution,
 $f(\phi)$ the drag, costly

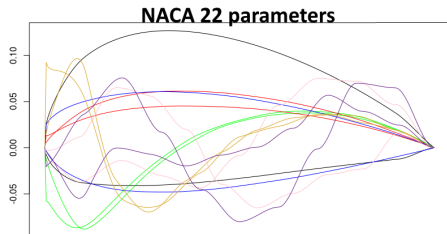
Eigenshape decomposition I

From a database of possible shapes $[\phi^{(1)}, \dots, \phi^{(5000)}]$,



...

extract a basis of most important shapes by principal component analysis, $\{V^1, \dots, V^\delta\}$

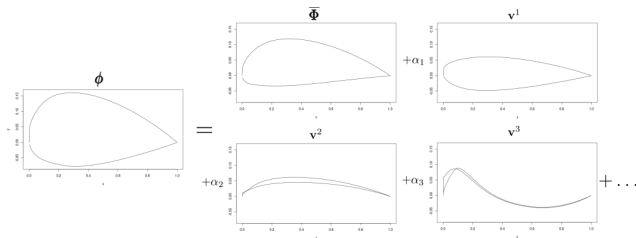


Eigenshape decomposition II

Shapes are now described with their eigencomponents α 's,

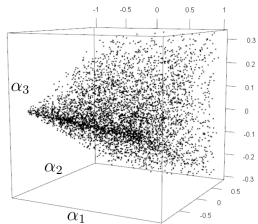
$$\phi \approx \bar{\phi} + \sum_{i=1}^{\delta} \alpha_i V^i$$

(general notation, $x = V\alpha + b$, $x \equiv \phi$, $\bar{\phi} \equiv b$)



$(\alpha_1, \dots, \alpha_\delta)$ make a specific manifold.

Cf. also [Raghavan et al., 2013, Li et al., 2018, Cinquegrana and Iuliano, 2018]



Further GP dimension control I

Build a GP to infer the drag from a shape, $Y(\alpha)$
+ control effects of dimension beyond the PCA.

Anisotropic kernel has 1 θ_i per dimension, **isotropic** has 1 for all dimensions.

$$\text{Expl: } k_{\text{ani}}(\alpha, \alpha') = \sigma^2 \exp\left(-\sum_{i=1}^d \frac{(\alpha_i - \alpha'_i)^2}{\theta_i^2}\right)$$
$$k_{\text{iso}}(\alpha, \alpha') = \sigma^2 \exp\left(-\frac{(\alpha - \alpha')^2}{\theta^2}\right)$$

Further GP dimension control II

- Likelihood that favors sparsity [Yi et al., 2011]:

$$\max_{\theta} \text{Log-Likelihood}(\theta; \mathbb{F}) - \lambda \|\theta^{-1}\|_1$$

⇒ active and non-active dimensions, α_a and $\alpha_{\bar{a}}$.

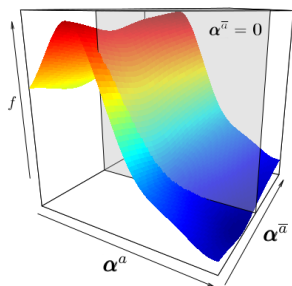
- GP as the sum of an anisotropic and isotropic GPs [Allard et al., 2016]:

$$k(\alpha, \alpha') = k_{\text{ani}}(\alpha_a, \alpha'_a) + k_{\text{iso}}(\alpha_{\bar{a}}, \alpha'_{\bar{a}})$$

Expl NACA22 :

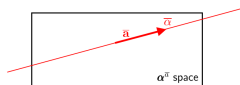
$$\alpha_a = (\alpha_1, \alpha_2, \alpha_3), \quad \delta_a = 3, \quad \delta = 20$$

⇒ 21 to 6 kernel parameters



- Optimize in the reduced dimensional space:

$$\alpha^{(t+1)*} = \arg \max_{[\alpha_a, \bar{\alpha}]} EI$$



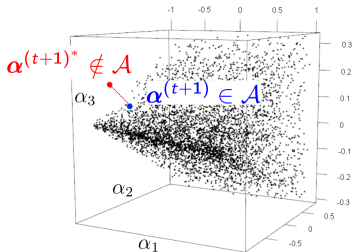
$\bar{\alpha}$ coordinate along a random line in non-active space, $\delta_a + 1$ dimensions.

- Proj_S** : projection of $V\alpha^{(t+1)*} + \bar{\phi}$ onto the closest CAD shape, $\phi_{CAD}^{(t+1)}$ with components α^{t+1} .

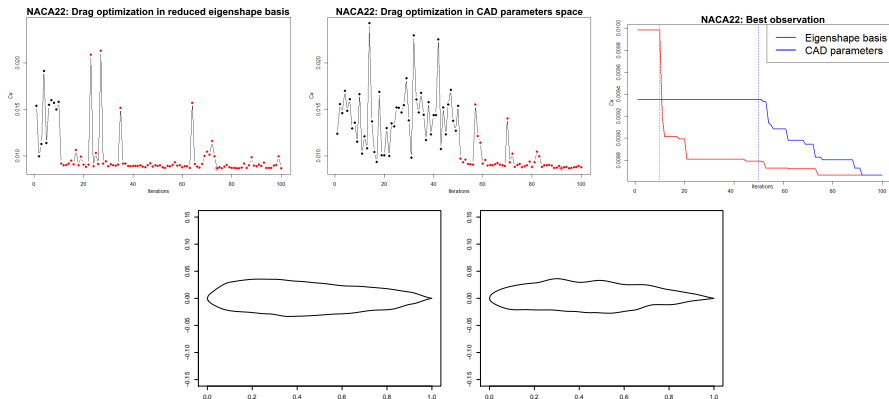
-

Calculate $f(\phi_{CAD}^{(t+1)})$.

Replication : update GP with both $\alpha^{(t+1)*}$ and $\alpha^{(t+1)}$



Example: NACA 22 airfoil drag minimization



- Faster decrease of the objective function in the reduced eigenshape basis (left) compared with the standard approach (right, CAD parameter space).
- Smoother airfoils are obtained because a shape basis is considered instead of a combination of local parameters.

Conclusions

- Building a metamodel (Gaussian Process) for optimization is different from building it for prediction : small initial DoE, quadratic trend. Ability to rank points early in the search may be a key.
- The integration of trust regions in BO algorithms expands the family of Derivative Free Optimization algorithms, creating a convergence of methods.
- State-of-the-art BO is competitive for multimodal functions in up to 10 dimensions. Much research on-going to go beyond.
- Main motivation for studying BO: an integrated mathematical framework for global optimization, with possibility to look at non-continuous spaces [Cuesta-Ramirez et al., 2022], parallel implementations [Janusevskis et al., 2012], problems with uncertainties [Pelamatti et al., 2022] ...

References I



Allard, D., Senoussi, R., and Porcu, E. (2016).
Anisotropy models for spatial data.
Mathematical Geosciences, 48(3):305–328.



Amine Bouhlel, M., Bartoli, N., Regis, R. G., Otsmane, A., and Morlier, J. (2018).
Efficient global optimization for high-dimensional constrained problems by using the kriging models combined with the partial least squares method.
Engineering Optimization, 50(12):2038–2053.



Binois, M., Ginsbourger, D., and Roustant, O. (2020).
On the choice of the low-dimensional domain for global optimization via random embeddings.
Journal of global optimization, 76(1):69–90.



Cinquegrana, D. and Iuliano, E. (2018).
Investigation of adaptive design variables bounds in dimensionality reduction for aerodynamic shape optimization.
Computers & Fluids, 174:89–109.

References II



Cuesta-Ramirez, J., Le Riche, R., Roustant, O., Perrin, G., Durantin, C., and Gliere, A. (2022).

A comparison of mixed-variables bayesian optimization approaches.

Advanced MOdeling and Simulation in engineering sciences.

to appear, available as preprint [arXiv:2111.01533](https://arxiv.org/abs/2111.01533).



Diouane, Y., Picheny, V., and Le Riche, R. (2021).

TREGO: a Trust-Region framework for Efficient Global Optimization.

arXiv.

preprint [arXiv:2101.06808](https://arxiv.org/abs/2101.06808).



Eriksson, D., Pearce, M., Gardner, J., Turner, R. D., and Poloczek, M. (2019).

Scalable global optimization via local bayesian optimization.

In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 5496–5507. Curran Associates, Inc.








Frazier, P. I. (2018).

A tutorial on Bayesian optimization.

arXiv preprint [arXiv:1807.02811](https://arxiv.org/abs/1807.02811).

References III

-  Gaudrie, D., Le Riche, R., Picheny, V., Enaux, B., and Herbert, V. (2020). Modeling and optimization with gaussian processes in reduced eigenbases. *Structural and Multidisciplinary Optimization*, 61:2343–2361.
-  Hansen, N., Auger, A., Mersmann, O., Tusar, T., and Brockhoff, D. (2016). Coco: A platform for comparing continuous optimizers in a black-box setting. *arXiv preprint arXiv:1603.08785*.
-  Hansen, N., Auger, A., Ros, R., Finck, S., and Pošík, P. (2010). Comparing results of 31 algorithms from the black-box optimization benchmarking bbob-2009. In *Proceedings of the 12th annual conference companion on Genetic and evolutionary computation*, pages 1689–1696. ACM.
-  Janusevskis, J., Le Riche, R., Ginsbourger, D., and Girdziusas, R. (2012). Expected improvements for the asynchronous parallel global optimization of expensive functions: Potentials and challenges. In *Learning and Intelligent Optimization*, pages 413–418. Springer.
-  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.

References IV



Le Riche, R. and Picheny, V. (2021).
Revisiting bayesian optimization in the light of the coco benchmark.
Structural and Multidisciplinary Optimization, 64(5):3063–3087.



Li, J., Bouhlel, M. A., and Martins, J. (2018).
A data-based approach for fast airfoil analysis and optimization.
In *2018 AIAA/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference*, page 1383.



Li, J., Cai, J., and Qu, K. (2019).
Surrogate-based aerodynamic shape optimization with the active subspace method.
Structural and Multidisciplinary Optimization, 59(2):403–419.



Mockus, J. (1975).
On Bayesian methods for seeking the extremum.
In *Optimization Techniques IFIP Technical Conference*, pages 400–404. Springer.



Pelamatti, J., Le Riche, R., Helbert, C., and Blanchet-Scalliet, C. (2022).
Coupling and selecting constraints in bayesian optimization under uncertainties.

References V



Priem, R. (2020).

High dimensional constrained optimization applied to aircraft design.

PhD thesis, Univ. de Toulouse - ISAE.

(in French).



Raghavan, B., Breitkopf, P., Tourbier, Y., and Villon, P. (2013).

Towards a space reduction approach for efficient structural shape optimization.

Structural and Multidisciplinary Optimization, 48(5):987–1000.



Regis, R. G. (2016).

Trust regions in Kriging-based optimization with expected improvement.

48:1037–1059.



Spagnol, A., Le Riche, R., and Da Veiga, S. (2019).

Global sensitivity analysis for optimization with variable selection.

SIAM/ASA Journal on Uncertainty Quantification, 7(2):417–443.



Wang, Z., Hutter, F., Zoghi, M., Matheson, D., and de Freitas, N. (2016).

Bayesian optimization in a billion dimensions via random embeddings.

Journal of Artificial Intelligence Research, 55:361–387.

References VI



Yi, G., Shi, J., and Choi, T. (2011).

Penalized Gaussian process regression and classification for high-dimensional nonlinear data.

Biometrics, 67(4):1285–1294.