Reducing dimension in Bayesian Optimization

Rodolphe Le Riche[1], Adrien Spagnol[1,2], David Gaudrie[1,3], Sébastien Da Veiga[2], Victor Picheny[4]

[1] CNRS LIMOS at Mines Saint Etienne, France
[2] Safran Tech , [3] PSA , [4] Prowler.io

July 2020
LIMOS internal seminar

Foreword

This talk was given at the LIMOS on July the 9th 2020 and was mainly intended for an audience of non specialists of Gaussian processes (GPs). The first slides (up to slide 12) about GPs and Bayesian Optimization should probably be skipped by readers already aware about these topics.
However, the two research contributions on variable selection for optimization 1) by kernel methods and, 2) by penalized likelihood in the PCA space, may be of interest to some experts.

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 ...

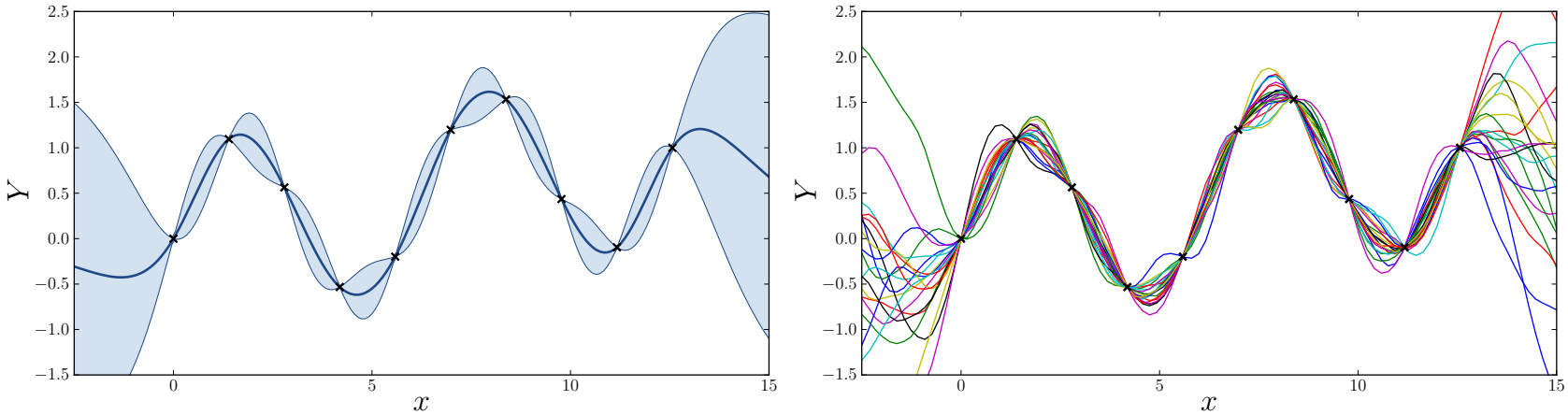
An exciting part of machine learning: algorithm design critical to performance, use expert knowledge.

Context: optimization of costly functions

To save calls to f, build a model of it based on previous evaluations and rely on it whenever possible $\longrightarrow$ metamodel / surrogate based optimization.

Gaussian process as metamodel : Bayesian Optimization.

Gaussian Process Regression (kriging)



$Y(x)|Y(\mathbb{X})=\mathbb{F}$ is $\mathcal{N}(m(.), c(.,.))$ with

$$m(x) = \mathbb{E}[Y(x)|Y(\mathbb{X})=\mathbb{F}] = k(x,\mathbb{X})k(\mathbb{X},\mathbb{X})^{-1}\mathbb{F}$$
$$c(x,x') = \mathrm{Cov}[Y(x), Y(x')|Y(\mathbb{X})=\mathbb{F}] = k(x,x') - k(x,\mathbb{X})k(\mathbb{X},\mathbb{X})^{-1}k(\mathbb{X},x')$$

$Y(x)$ is parameterized through $k(x,x';\theta)$.

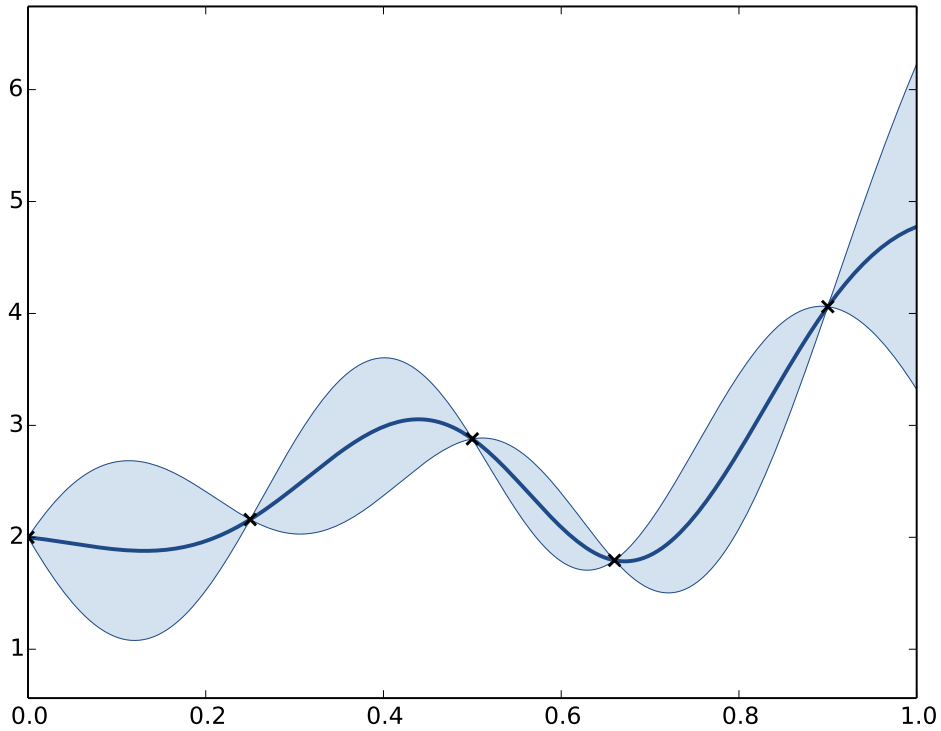
Ex: $k(x,x') = \sigma^2 \exp(-\sum_{i=1}^{d} \frac{(x_i - x_i')^2}{2\theta_i^2})$.

Bayesian Optimization

Global optimization methods are a trade-off between

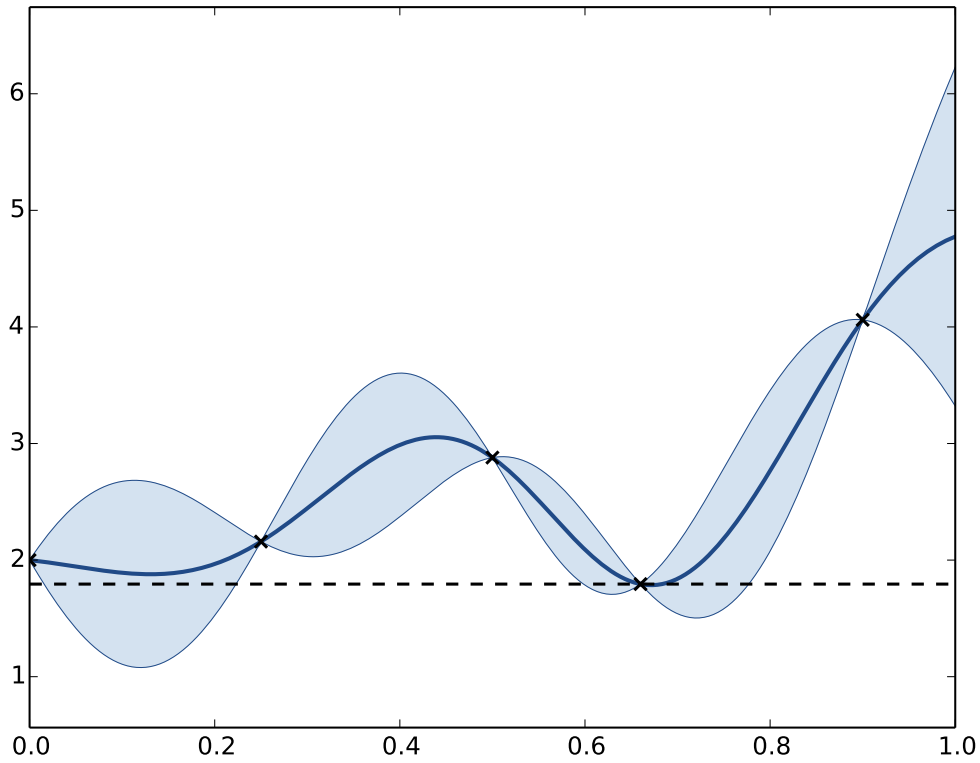
- Intensification in known good regions
- Exploration of new regions

How can kriging models be helpful?



(EGO figures from [Durrande and Le Riche, 2017])

In our example, the best observed value is 1.79

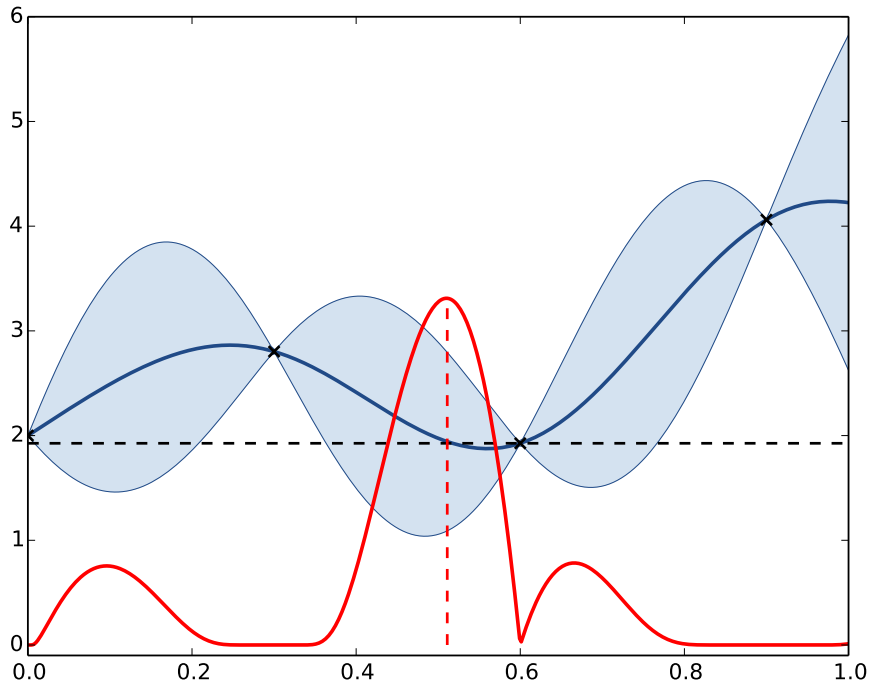


We need a criterion that uses the GP and seeks a compromise between exploration and intensification: the expected improvement (among other acquisition criteria).

The Expected Improvement

Measure of progress: the improvement,
$I(x) = \max(0, (\min(F) - Y(x) \mid Y(\mathbb{X})=\mathbb{F}))$.
Acquisition criterion: $\mathrm{EI}(x) = \int_{-\infty}^{+\infty} I(x)\ dy(x) = \cdots =$
$\sqrt{c(x,x)}\,[w(x)\mathrm{cdf}_{\mathcal{N}}(w(x)) + \mathrm{pdf}_{\mathcal{N}}(w(x))]$

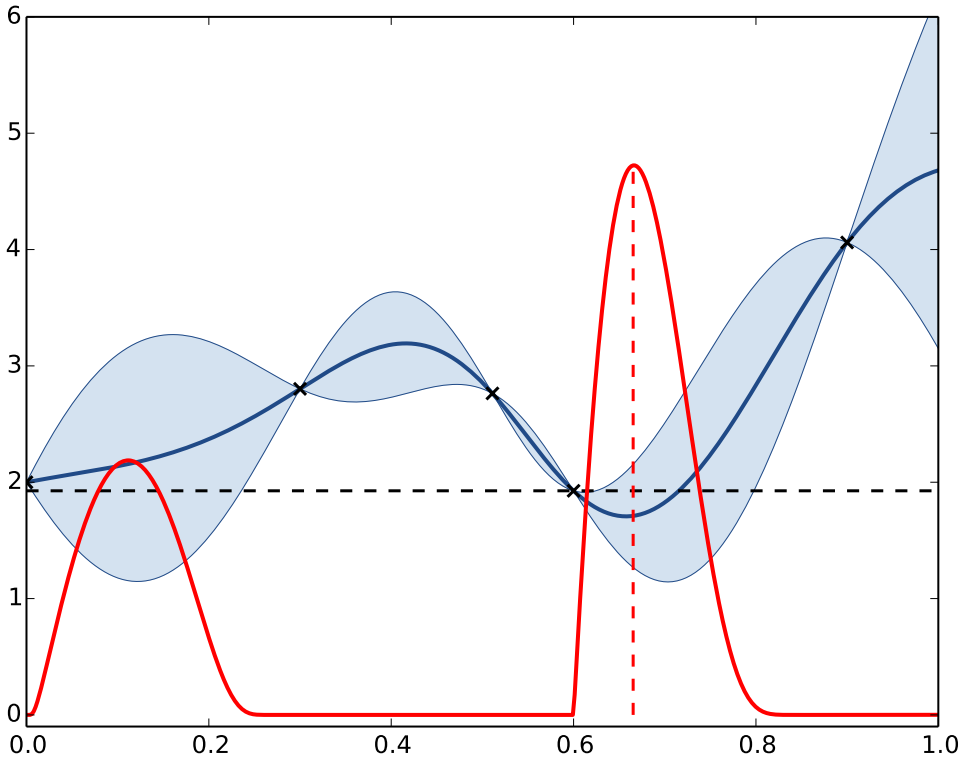
with $w(x) = \frac{\min(F) - m(x)}{\sqrt{(c(x,x))}}$.



Expected Improvement

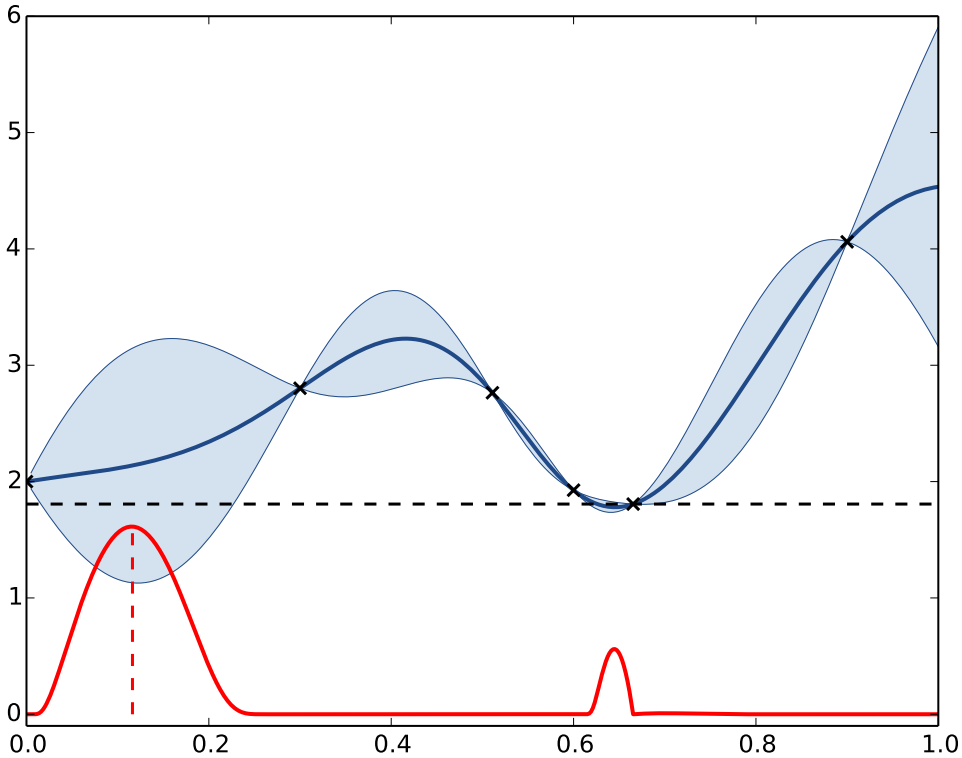
$$x^{t+1} = \arg\max_{x \in \mathcal{X}} \mathrm{EI}(x)$$

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



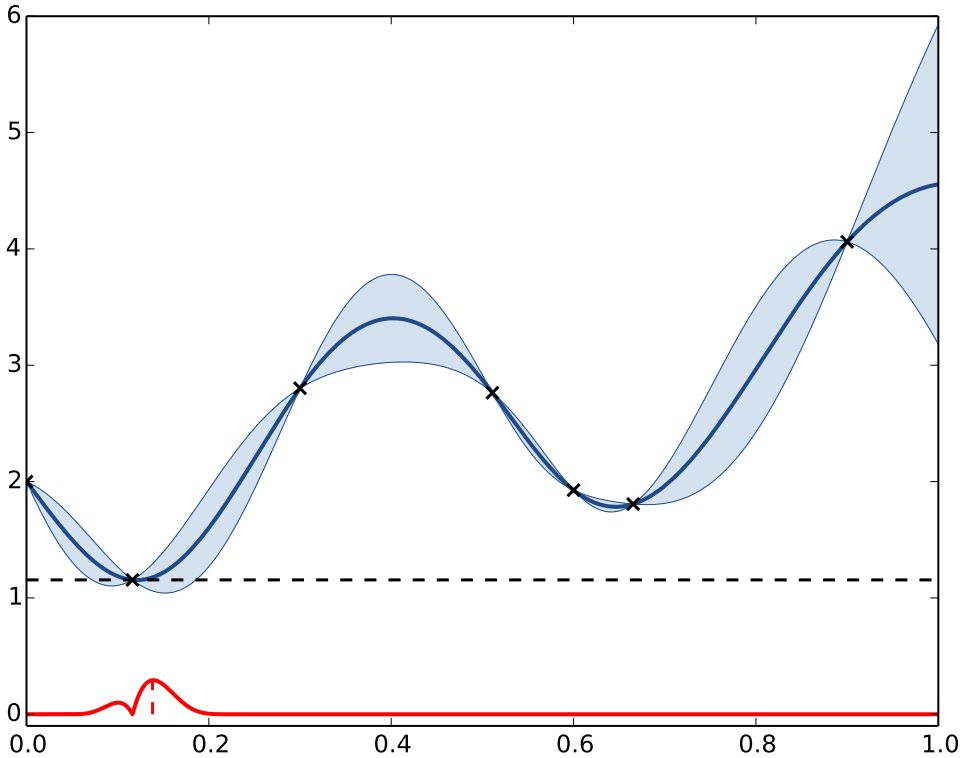
Expected Improvement

$x^{t+1} = \arg\max_{x \in \mathcal{X}} \mathrm{EI}(x)$... iteration 2



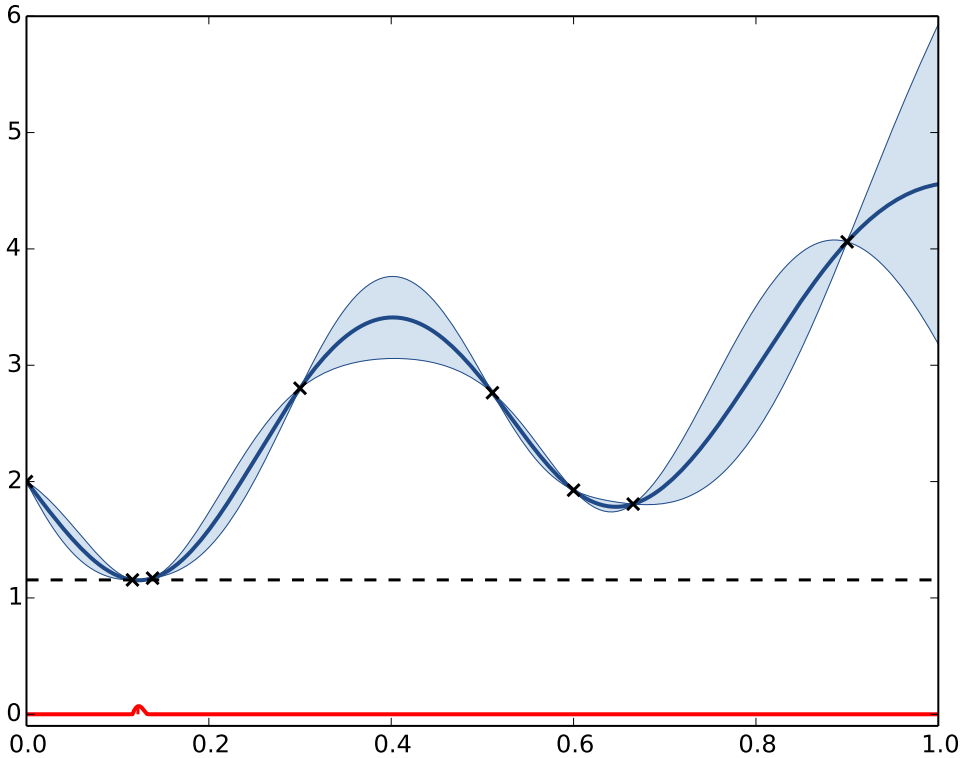
Expected Improvement

$x^{t+1} = \arg\max_{x \in \mathcal{X}} \mathrm{EI}(x)$... iteration 3



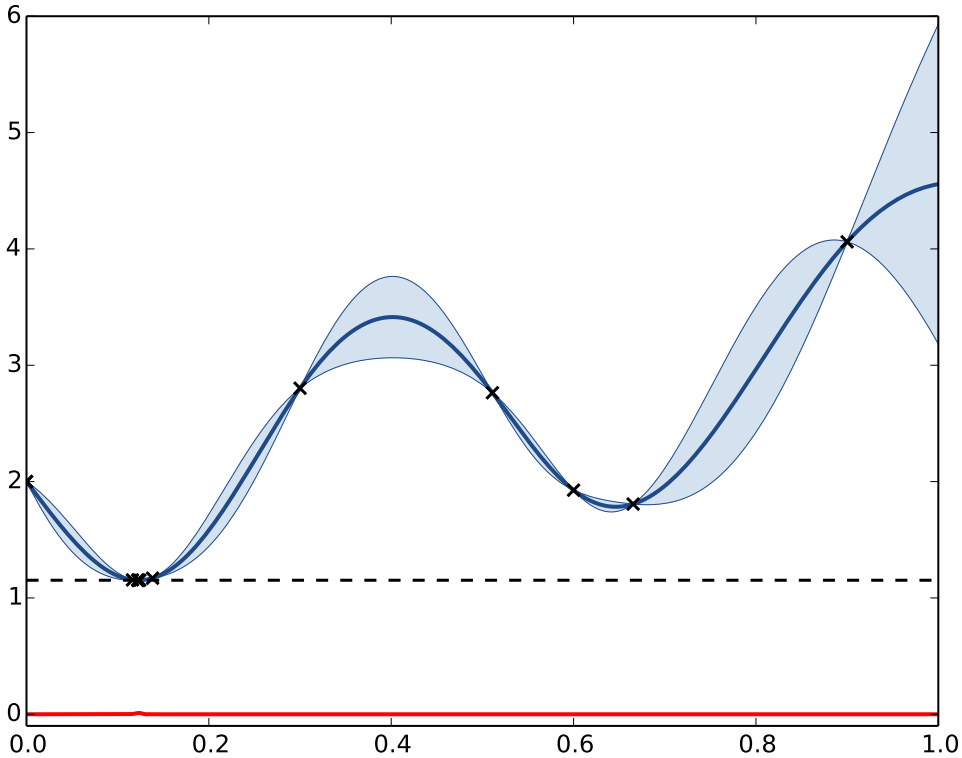
Expected Improvement

$x^{t+1} = \arg\max_{x \in \mathcal{X}} \mathrm{EI}(x)$... iteration 4



Expected Improvement

$x^{t+1} = \arg\max_{x \in \mathcal{X}} \mathrm{EI}(x)$... iteration 5



This algorithm is called **Efficient Global Optimization** (EGO, [Jones et al., 1998]), an instance of Bayesian Optimization (BO):

1. make an initial design of experiments X and calculate the associated F, $t = \texttt{length}(F)$
2. build a GP from (X, F) (max. likelihood $\rightarrow$ θ)
3. $x^{t+1} = \arg\max_x \mathrm{EI}(x)$ (with another optimizer, e.g. CMA-ES [Hansen and Ostermeier, 2001])
4. calculate $F_{t+1} = f(X_{t+1})$, increment t
5. stop ($t > t^{\max}$) or go to 2.

State-of-the-art for costly functions.

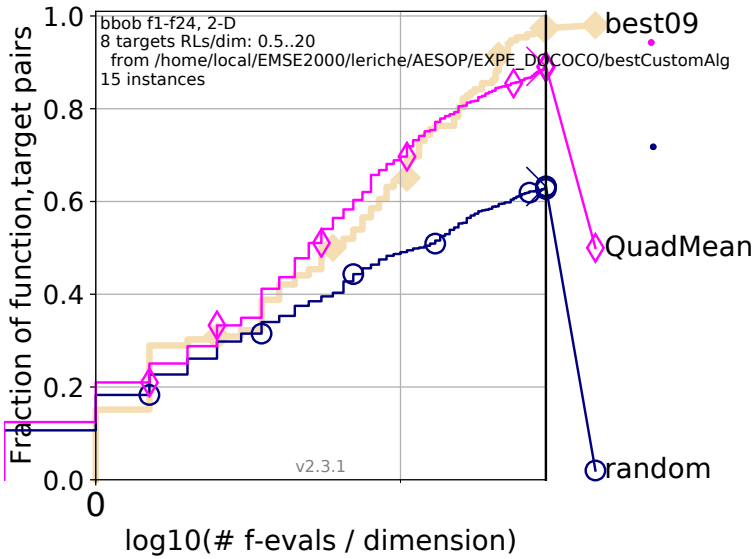
Bayesian optimization and COCO

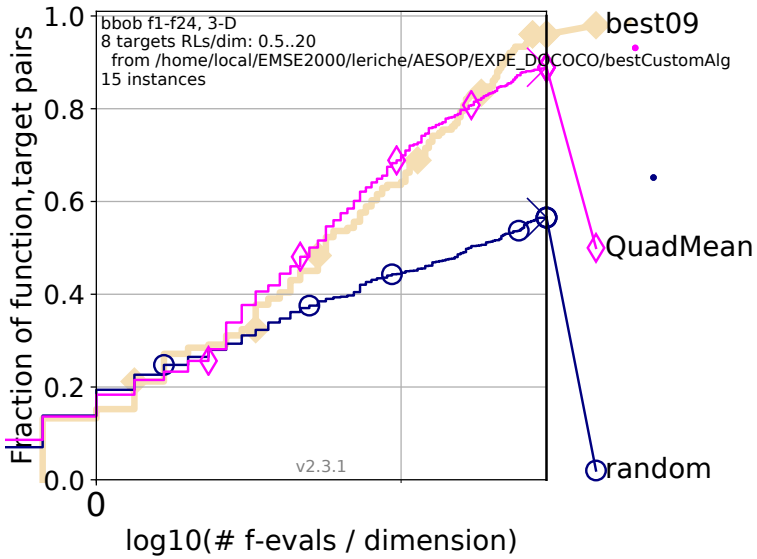
COCO : COmparing Continuous Optimizers [Hansen et al., 2016] with 24 functions of the BBOB noiseless suite [Hansen et al., 2010]. 15 repetitions of runs of length $30 \times d$ (=2,3,5,10) $\rightarrow$ 360 optimizations per dimension, 432000 maximizations solved, millions of covariance matrices inversions.

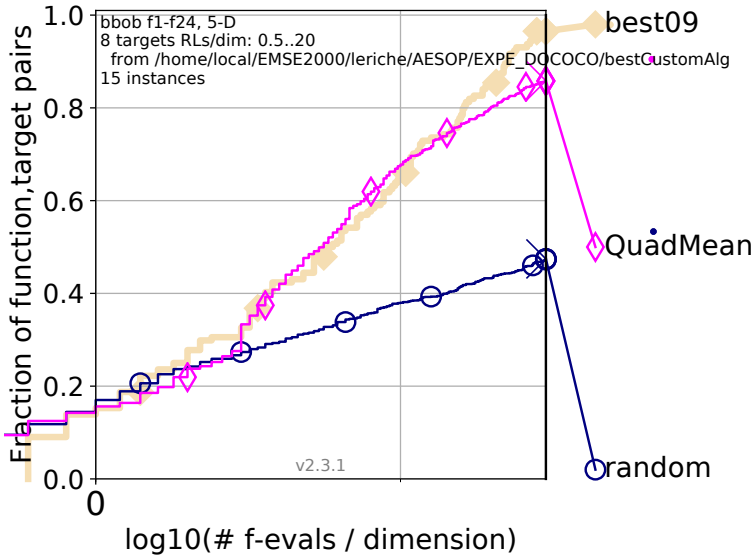
QuadMean : Bayesian Optimizer with quadratic trend optimized every 5 iterations.

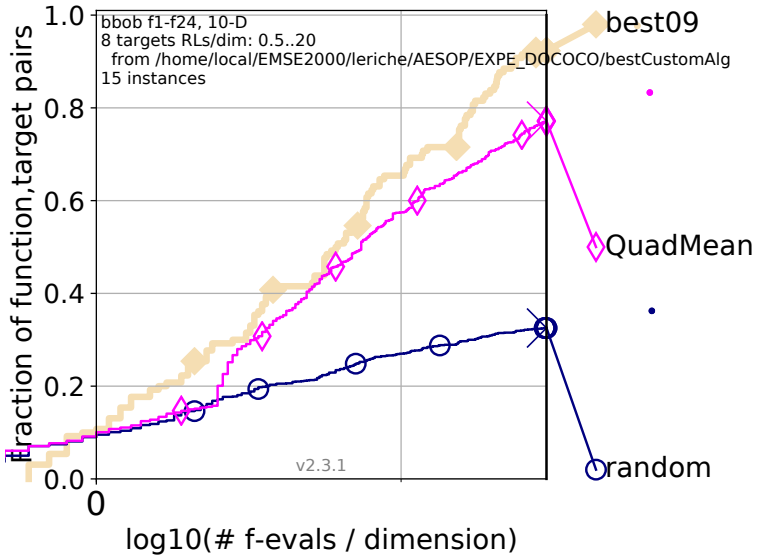
best09 : utopic algorithm made of the best (for each cost and dimension) of the 32 algorithms competing at BBOB 2009.

Bayesian optimization and COCO









Bayesian optimization and dimension

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

Intuitively logical since they attempt to build a model of the function throughout the search space $\mathcal{S}$.

2 research efforts:

1. Reduce dimension by selecting variables. Joined work with Adrien Spagnol and Sébastien Da Veiga [Spagnol et al., 2019].

Bayesian optimization and dimension

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

Intuitively logical since they attempt to build a model of the function throughout the search space $\mathcal{S}$.

2 research efforts:

1. Reduce dimension by selecting variables. Joined work with Adrien Spagnol and Sébastien Da Veiga [Spagnol et al., 2019].
2. Gaussian process and optimization in reduced dimension for shapes. Joined work with David Gaudrie and Victor Picheny [Gaudrie et al., 2020].

Bayesian optimization and dimension

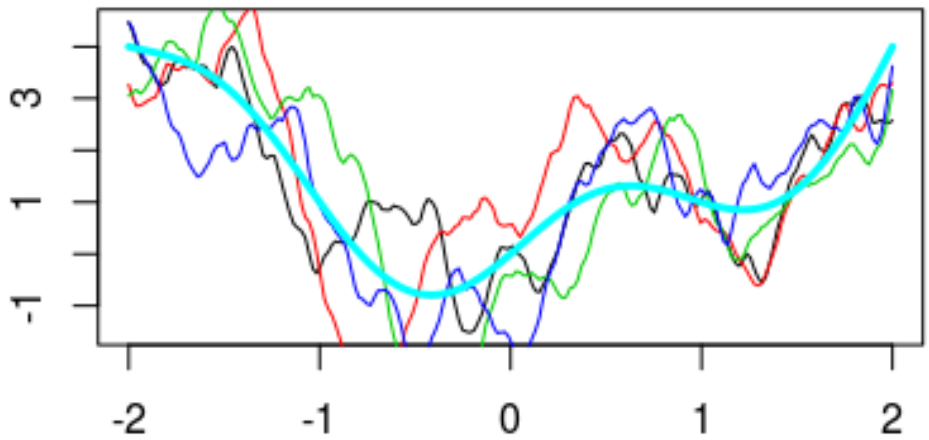
1. Reduce dimension by selecting variables. Joined work with Adrien Spagnol and Sébastien Da Veiga [Spagnol et al., 2019].

Kernel based sensitivity indices for optimization

Global sensitivity analysis: quantify the importance of a given set of variables for the function f.
Classically, the part of the function variance attributed to the set of variables

Sobol indices [Sobol, 1993]

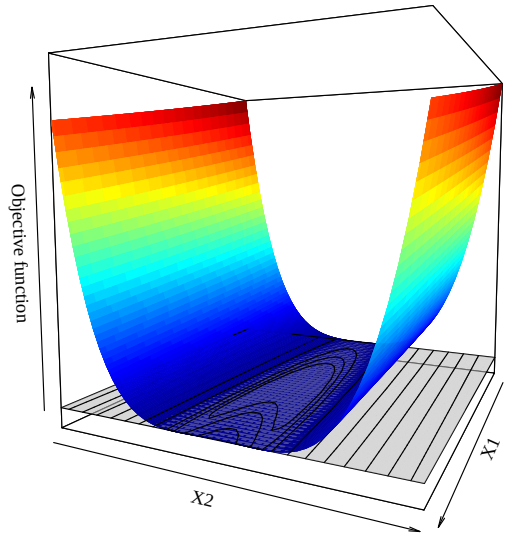
$$S_i = \frac{\mathrm{Var}(\mathbb{E}(Y \mid X_i))}{\mathrm{Var}(Y)}$$

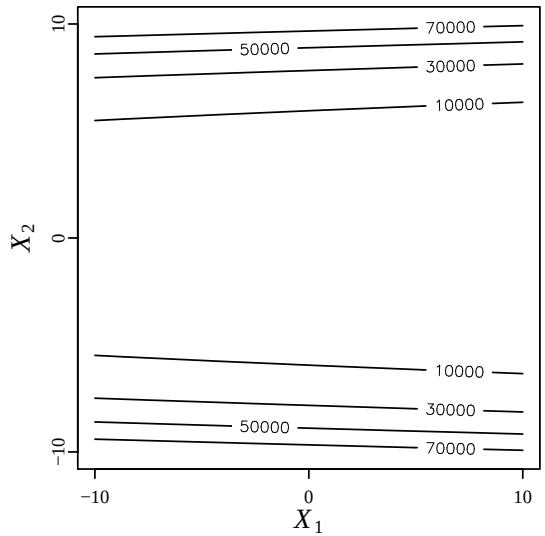


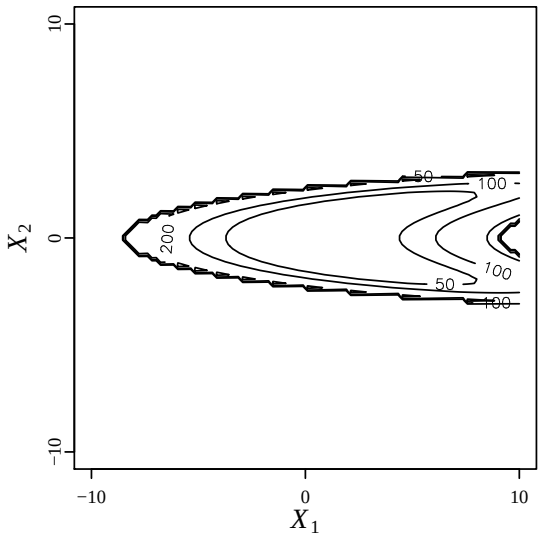
But optimization is focused on low regions of f (as opposed to all the fluctuations).

A goal-oriented index for optimization

Natural to use sublevel sets in optimization:
$\mathcal{D}_q = \{x \in \mathcal{S} \mid f(x) \leq q\}$

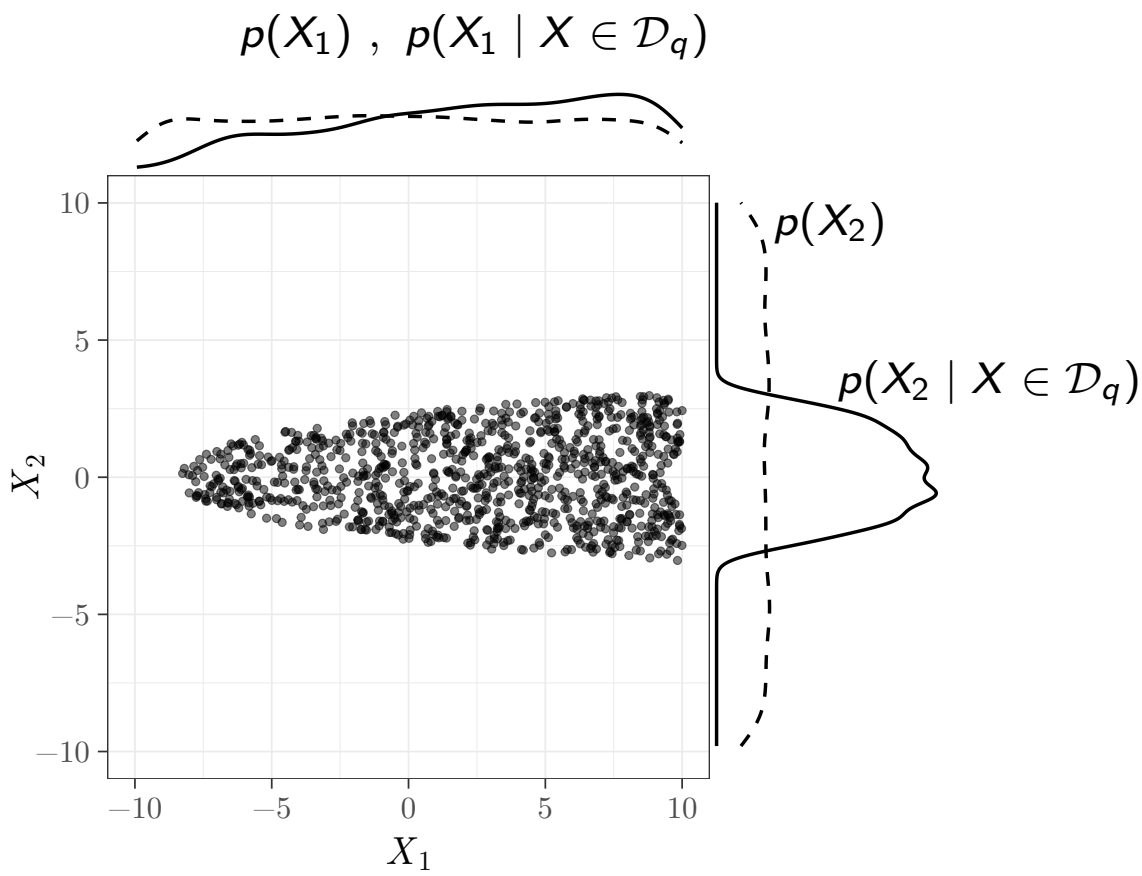






Dixon-Price function, $f(X) = (X_1 - 1)^2 + 2(X_2^2 - X_1)^2$
X_1 unimportant to reach $q = 10000$, both X_1 and X_2 important and coupled for $q = 50$

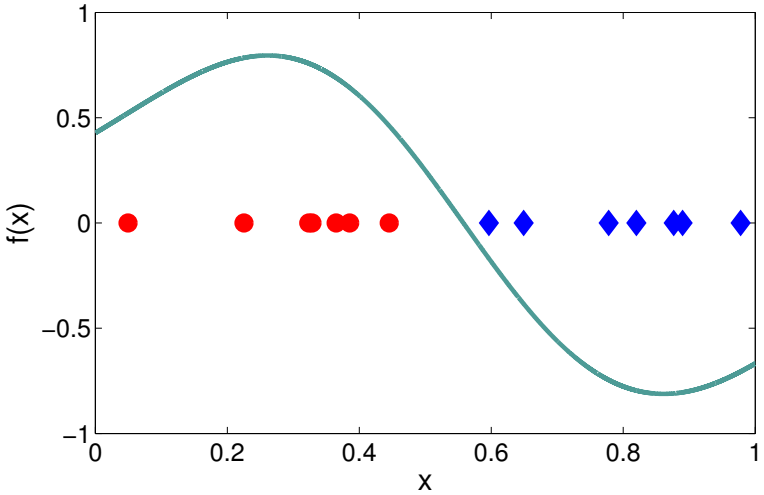
An optimization oriented sensitivity for X_i : distance between the non-informative $p(X_i)$ and the marginal distribution of the good points $p(X_i \mid X \in \mathcal{D}_q)$.



A robust statistics: the MMD

How to measure the distance between $\mathbf{P} \equiv p(X_i)$ and $\mathbf{Q} \equiv p(X_i) \mid X \in \mathcal{D}_q$?
Use the *Maximum Mean Discrepancy* (MMD), a kernel-based measure that is less sensitive to the number of points and dimension (adaptation to the data):

$$\mathrm{MMD}^2(\mathbf{P}, \mathbf{Q}) = \left(\sup_{f \in \mathcal{H}, \|f\| \leq 1} [\mathbb{E}_{\mathbf{P}}(f(X)) - \mathbb{E}_{\mathbf{Q}}(f(X))] \right)^2$$



$\mathcal{H}$ RKHS induced by kernel $k(.,.)$.

See Gretton et al., [Smola et al., 2007, Fukumizu et al., 2009]

MMD estimation

$$\text{Mean embedding : } \mu_{\mathbf{P}}(.) = \int k(x,.)p(x)dx$$

$$\text{MMD}^2(\mathbf{P},\mathbf{Q}) = \left(\sup_{f\in\mathcal{H},\|f\|\leq 1} [\mathbb{E}_{\mathbf{P}}(f(X)) - \mathbb{E}_{\mathbf{Q}}(f(X))] \right)^2$$

$$\mathbb{E}_{\mathbf{P}}(f(X)) = \int f(x)p(x)dx = \int \langle k(x,.), f\rangle_{\mathcal{H}} p(x)dx = \langle \mu_{\mathbf{P}}(.), f\rangle_{\mathcal{H}}$$

$$\sup_{f\in\mathcal{H},\|f\|\leq 1} [\mathbb{E}_{\mathbf{P}}(f(X)) - \mathbb{E}_{\mathbf{Q}}(f(X))] = \sup_{f\in\mathcal{H},\|f\|\leq 1} \langle \mu_{\mathbf{P}}(.) - \mu_{\mathbf{Q}}(.), f\rangle_{\mathcal{H}} = \|\mu_{\mathbf{P}}(.) - \mu_{\mathbf{Q}}(.)\|_{\mathcal{H}}$$

$$\text{MMD}^2(\mathbf{P},\mathbf{Q}) = \langle \mu_{\mathbf{P}}(.) - \mu_{\mathbf{Q}}(.), \mu_{\mathbf{P}}(.) - \mu_{\mathbf{Q}}(.)\rangle_{\mathcal{H}}$$

develop, get terms like $\langle \mu_{\mathbf{P}}(.), \mu_{\mathbf{Q}}(.)\rangle_{\mathcal{H}} = \int\int k(x,x')p(x)q(x')dxdx'$ and take the empirical means from input sample $\mathbb{X}_i = \{x_i^1, \ldots, x_i^n\}$ and subsample $\widetilde{\mathbb{X}}_i = \{x_i^1, \ldots, x_i^m \mid x \in \mathcal{D}_q\}$

$$\text{MMD}^2(\mathbf{P},\mathbf{Q}) \approx \frac{1}{n(n-1)} \sum_{p=1}^{n}\sum_{q\neq p}^{n} k(x_i^p, x_i^q) + \frac{1}{m(m-1)} \sum_{p=1}^{m}\sum_{q\neq p}^{m} k(\widetilde{x}_i^p, \widetilde{x}_i^q) - \frac{2}{nm} \sum_{p=1}^{n}\sum_{q=1}^{m} k(x_i^p, \widetilde{x}_i^q)$$

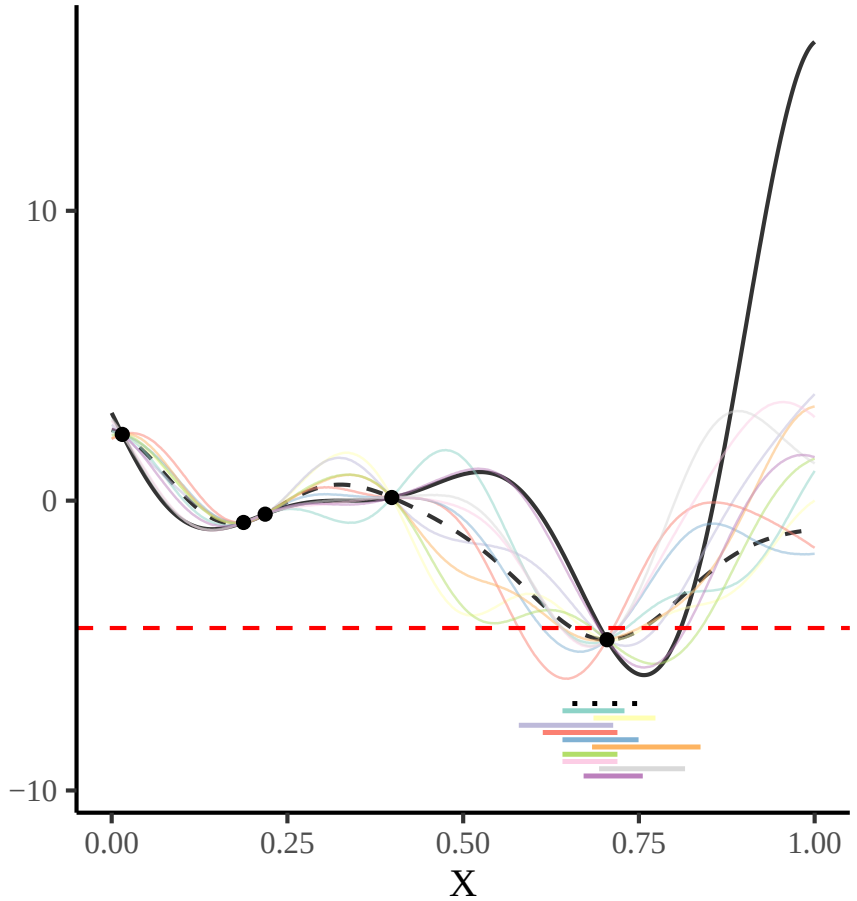
Easy to calculate. Equivalent to an independence measure between X_i and $\mathbb{1}(f(X) \leq q)$ [Spagnol et al., 2019].

Kernel-based sensitivity index

Sensitivity of variable i to reach the sublevel set $\mathcal{D}_q$:

$$S_i = \frac{\mathrm{MMD}^2(p(X_i), p(X_i \mid X \in \mathcal{D}_q))}{\sum_{j=1}^{d} \mathrm{MMD}^2(p(X_j), p(X_j \mid X \in \mathcal{D}_q))}$$

For costly functions, estimate S_i with the Gaussian process trajectories (account for model error) $\Rightarrow$ one $S_i^{(l)}$ per trajectory l.



KSA-BO

Kernel-based Sensitivity Analysis Bayesian Optimization

1. make an initial design of experiments X and calculate the associated F, $t = \texttt{length}(F)$
2. build a GP from (X, F) (max. likelihood)
3. Select active variables $a \in \{1, \ldots, d\}$: variable i selected if

p-value: $\mathbb{P}\left[\bar{S}_i = 1/N_{\text{traj}} \sum_l S_i^{(l)} < S \text{ random sample}\right] \leq 0.01$ or 0.05

determ.: or $\bar{S}_i > 1/d$

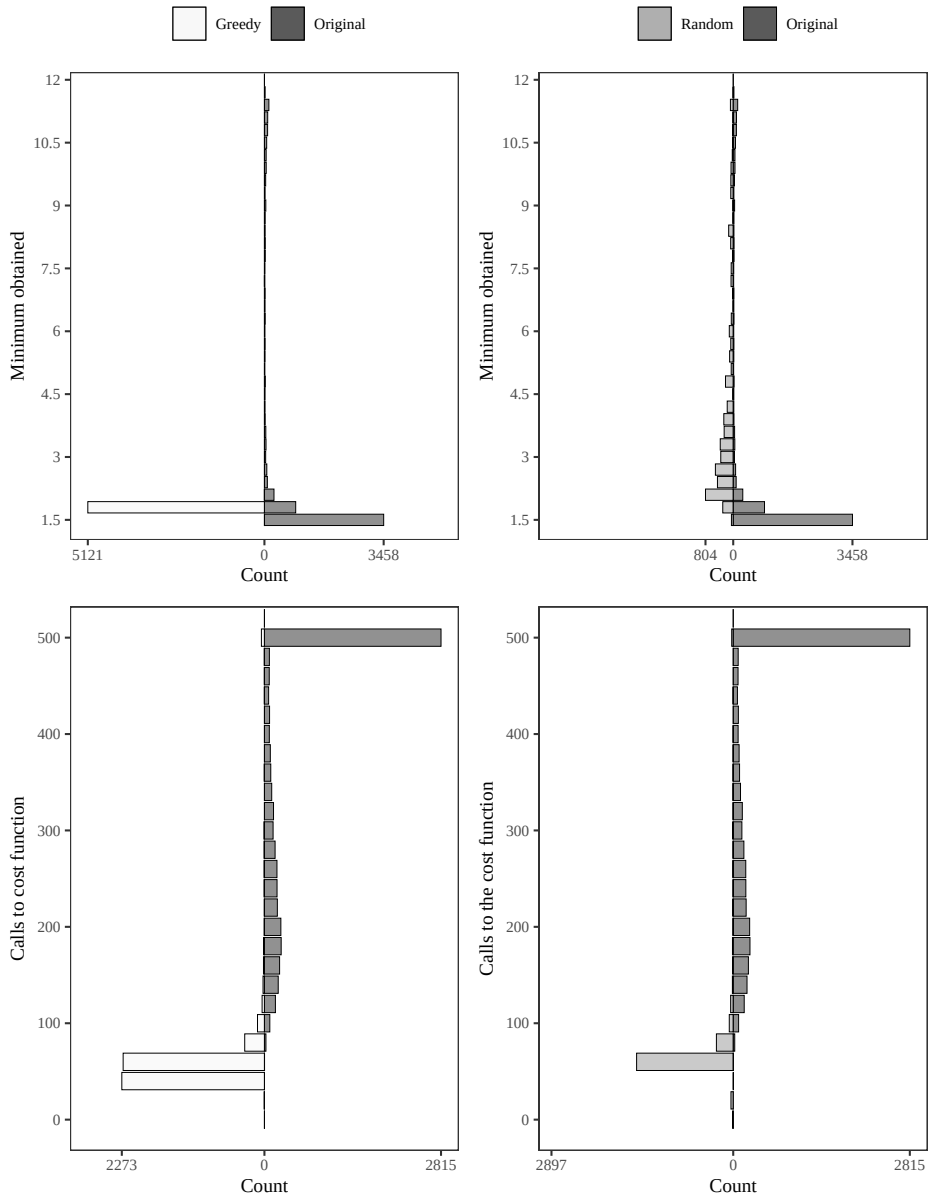
4. $x_a^{t+1} = \arg\max_{x_a} \text{EI}(x_a)$, $x_{\bar{a}}^{t+1} =$ best so far or random with proba 0.5, component-wise
5. calculate $F_{t+1} = f(X_{t+1})$, increment t
6. stop $(t > t^{\max})$ or go to 2.

Robustified version of the KSA-BO from [Spagnol et al., 2019]. Some tuning omitted here : how to choose $x_{\bar{a}}$, initial $p(X_i)$ and q? Details in [Spagnol, 2020].

Preliminary results

Welded beam problem, a priori selection of the active variables (no GP), $d = 4$ but $a = \{1, 4\}$ (deterministic strategy), 10000 repetitions of optimization.

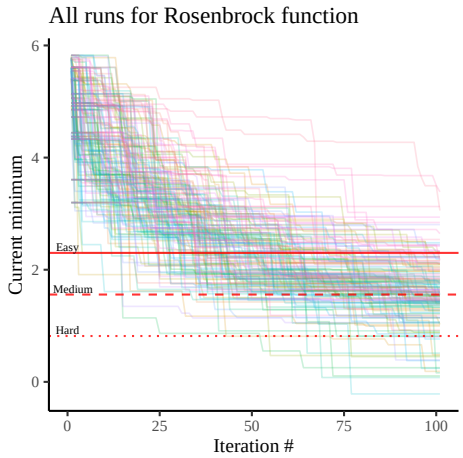
Note the compromise accuracy of the optimum vs. cost.



Results : test set

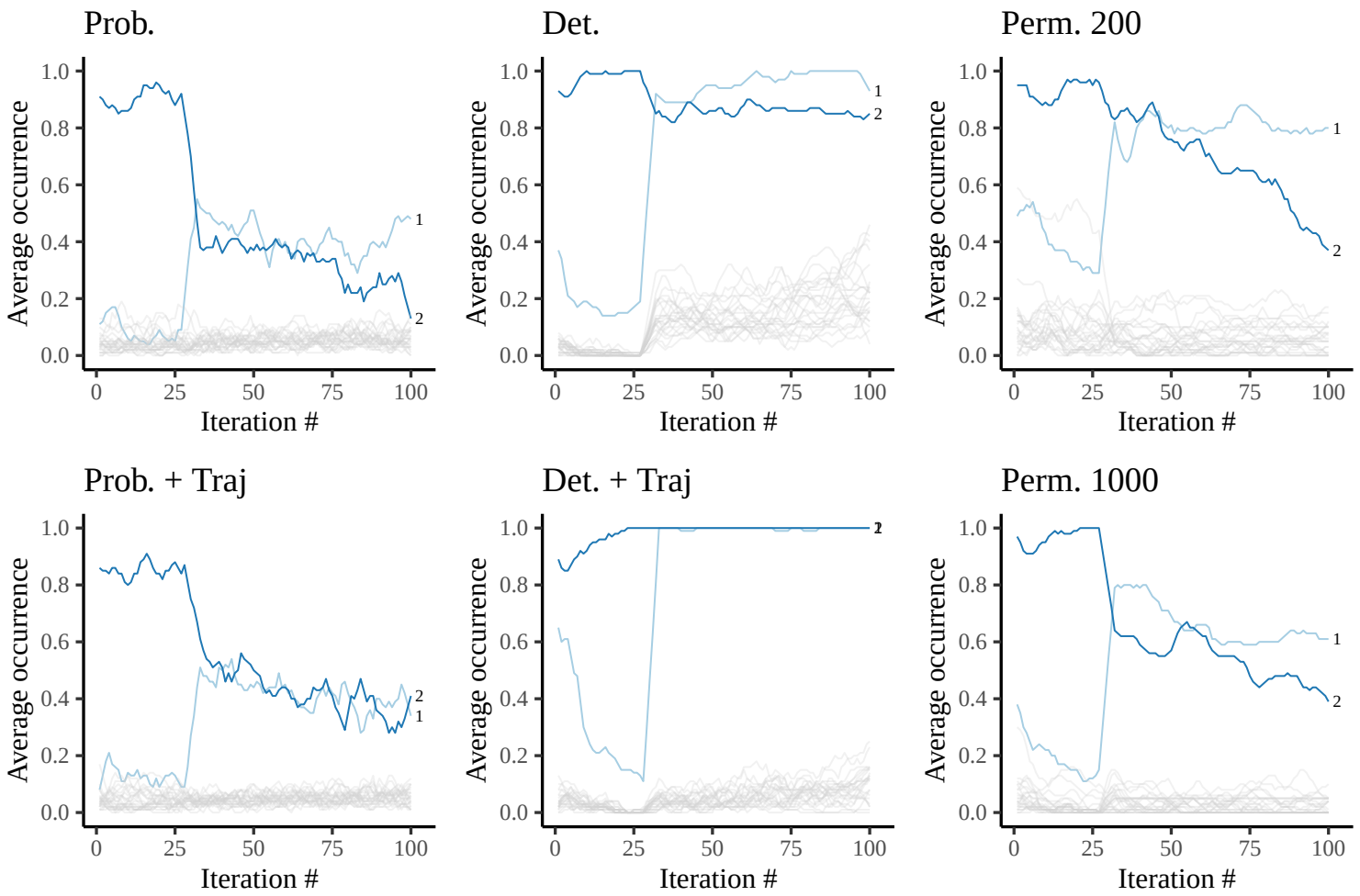
20 repetitions on

Name	d_{eff}	d	Expression
Branin	2	25	$f(\mathbf{X}) = \left(X_2 - \frac{5.1}{4\pi^2}X_1^2 + \frac{5}{\pi}X_1 - 6\right)^2 + 10\left(1 - \frac{1}{8\pi}\right)\cos(X_1) + 10$
Rosenbrock	5	20	$f(\mathbf{X}) = \sum_{i=1}^{d-1} 100\left(X_{i+1} - X_i^2\right)^2 + (X_i - 1)^2$
Borehole	8	25	$f(\mathbf{X}) = \frac{2\pi X_3(X_4 - X_6)}{\ln(X_2/X_1)(1 + \frac{2X_7X_3}{\ln(X_2/X_1)X_1^2X_8} + \frac{X_3}{X_5})}$
Ackley	6	20	$f(\mathbf{X}) = -20\exp\left(-0.2\sqrt{\frac{1}{d}\sum_{i=1}^{d}X_i^2}\right) - \exp\left(\frac{1}{d}\sum_{i=1}^{d}\cos(2\pi X_i)\right) + 20 + \exp(1)$
Schwefel	20	20	$f(\mathbf{X}) = \sum_{i=1}^{d}\left(\sum_{j=1}^{i}X_j\right)^2$
Stybtang	20	20	$f(\mathbf{X}) = \frac{1}{2}\sum_{i=1}^{d}(X_i^4 - 16X_i^2 + 5X_i)$



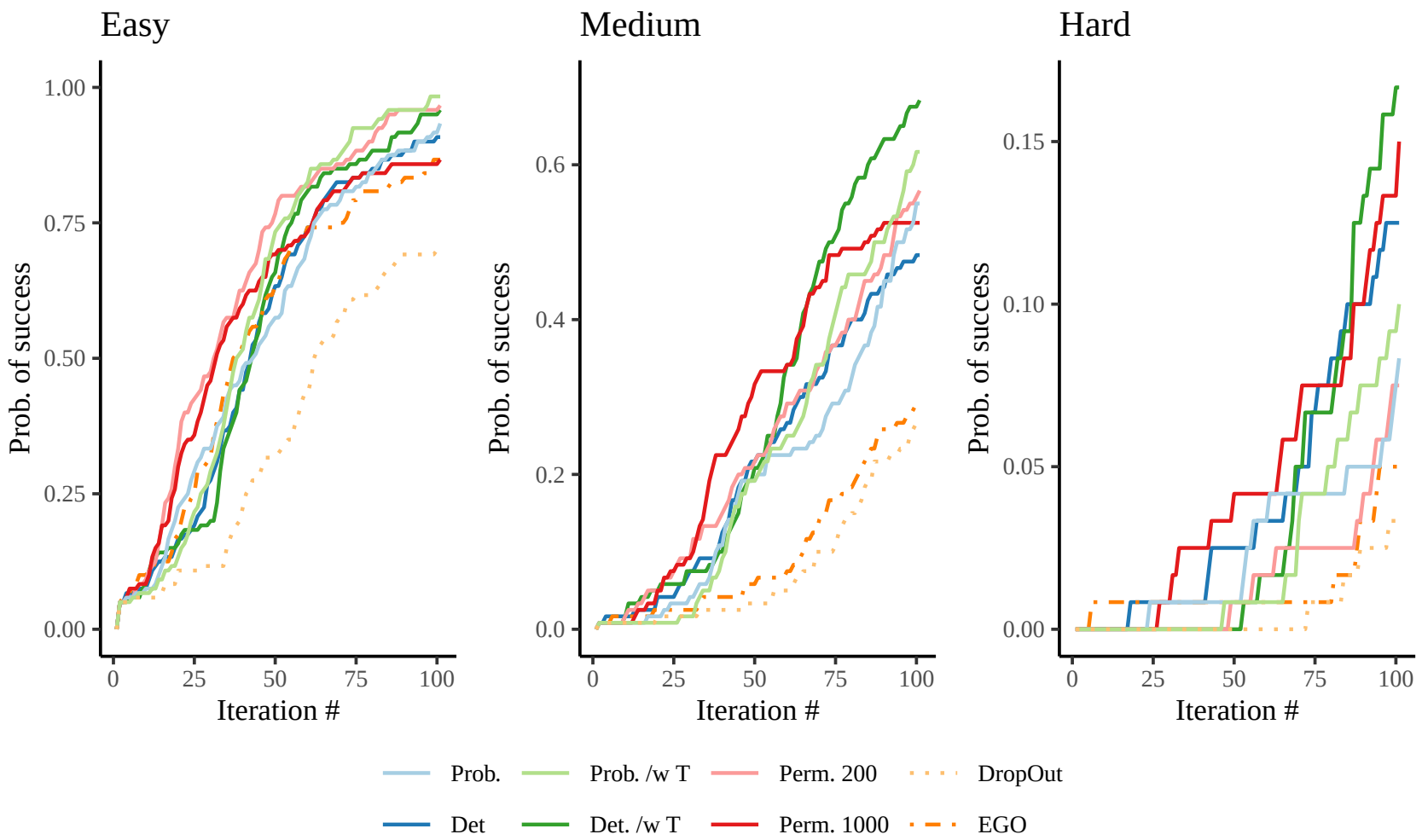
easy, medium, hard = 90 , 50 , 10% solved

Results : variables selection rates



Branin 25d, 2 first variables are active, 23 dummy.
Idem on other functions: **variables are correctly selected**.
@30 iterations set more ambitious goals: (p, q) go from $(100\%, 30\%)$ to $(30\%, 5\%)$.

Results : task solving rate



KSA-BO outperforms EGO and Dropout. Versions with trajectories perform better. Deterministic approach better overall.

Outline

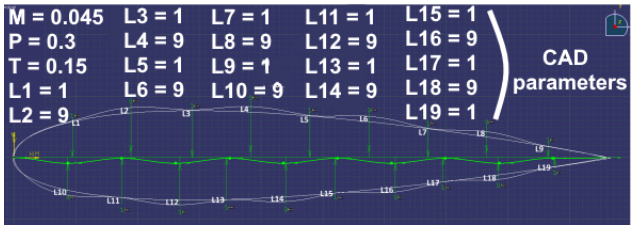
Bayesian optimizers are very competitive at low number of function evaluations but they loose this advantage with dimension.

2 research efforts to reduce dimension:

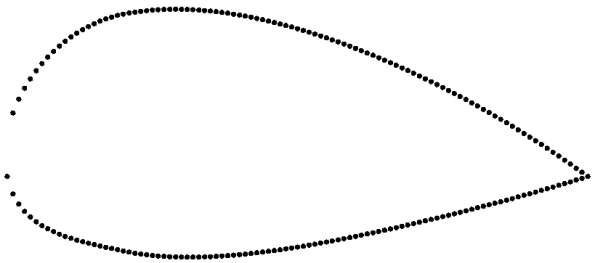
1. Reduce dimension by selecting variables.
2. Gaussian process and optimization in reduced dimension for shapes. Joined work with David Gaudrie and Victor Picheny [Gaudrie et al., 2020].

Dimension reduction for shapes : summary

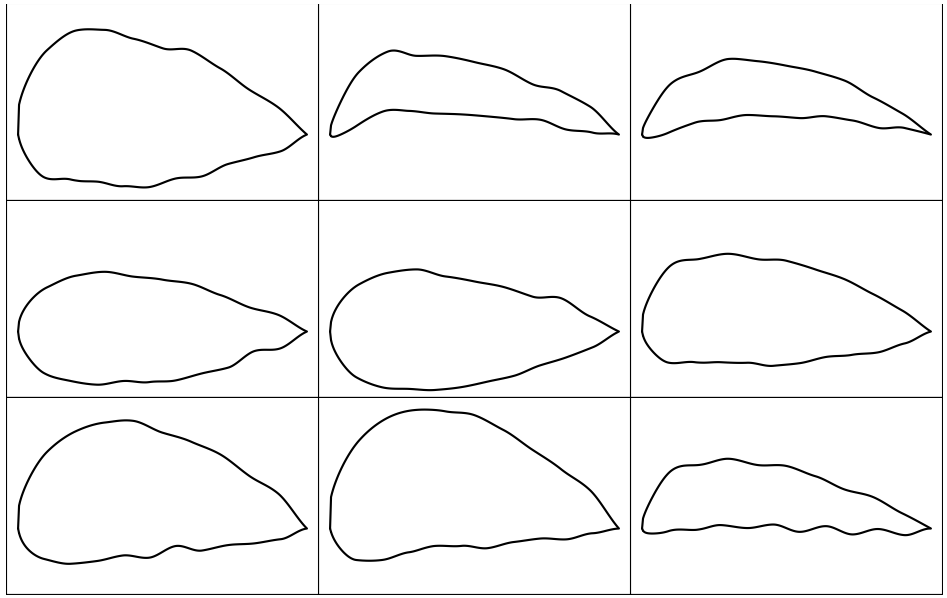
Shapes are described by CAD parameters $x \in \mathbb{R}^d$



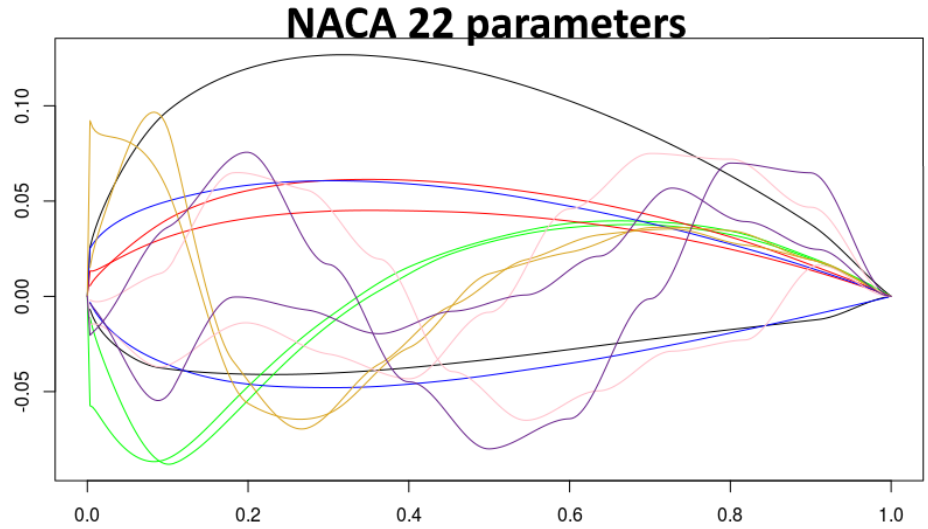
Nonlinear map to a high dimensional space $\phi(x) \in \mathbb{R}^D$, $D \gg d$ (free from biases created by CAD choices): here by contour discretization [Stegmann and Gomez, 2002]



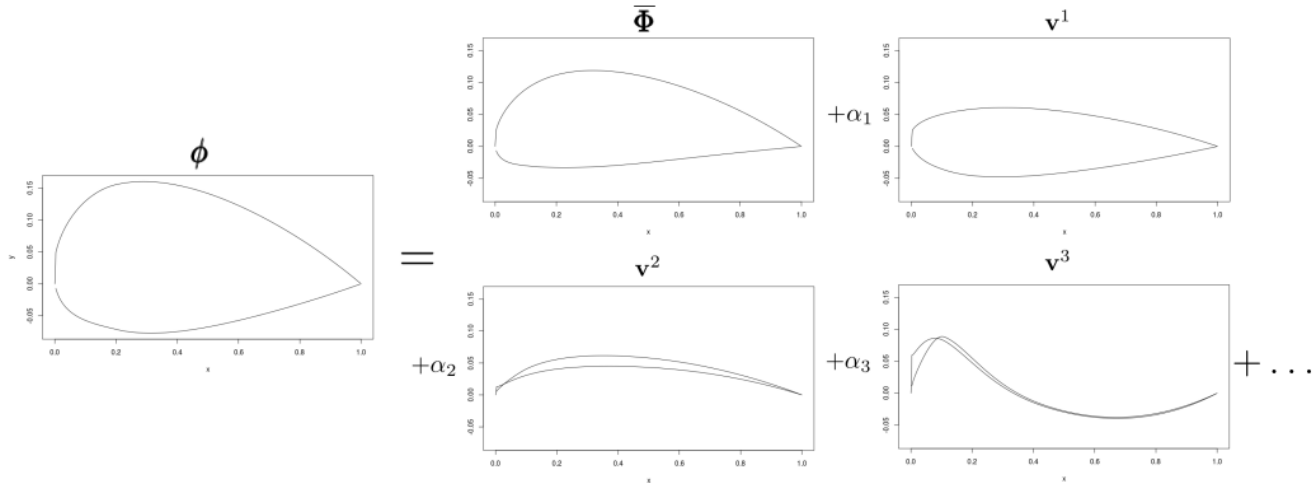
From a database of possible shapes $[\phi(x^{(1)}), \ldots, \phi(x^{(n)})]$,



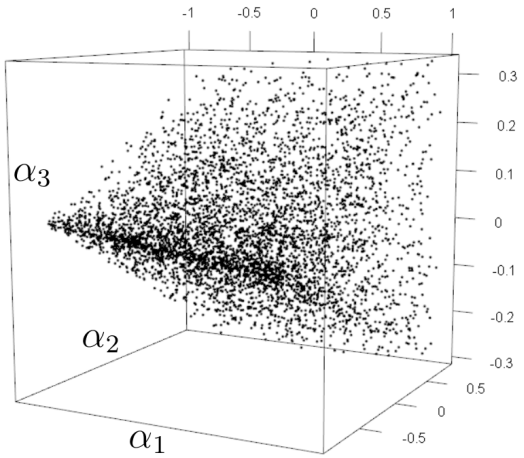
extract a basis of most important shapes by principal component analysis, $\{V^1, \ldots, V^\delta\}$.



Then work (build a GP, optimize) in this basis,



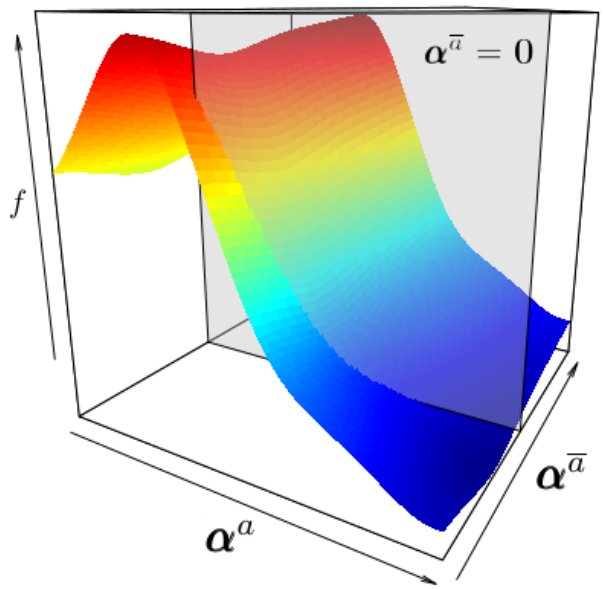
i.e. in the $(\alpha_1, \ldots, \alpha_\delta)$ manifold.



The choice of $\phi(x)$ is important. Other choices: characteristic function [Raghavan et al., 2013], signed distance to contour [Raghavan et al., 2014]

Further reduce dimension of the GP within the α-space of eigencomponents:

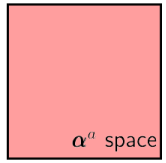
- Likelihood that favors sparsity [Yi et al., 2011]:
  max_θ Likelihood$(\theta; f(\mathbb{X})) - \lambda\|\theta^{-1}\|_1$
- GP with zonal anisotropy [Allard et al., 2016]:
  $Y(\alpha) = Y^a(\alpha_a) + Y^{\bar{a}}(\alpha_{\bar{a}})$, $Y^a(\alpha_a)$ detailed (anisotropic), $Y^{\bar{a}}(\alpha_{\bar{a}})$ isotropic

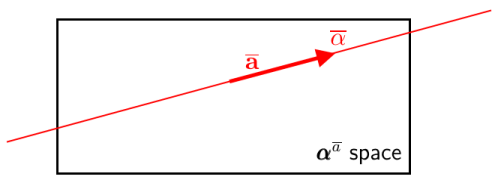


Expl NACA22 : Card$(a) = 3$, $\delta = 10$, $d = 22$, $D = 600$

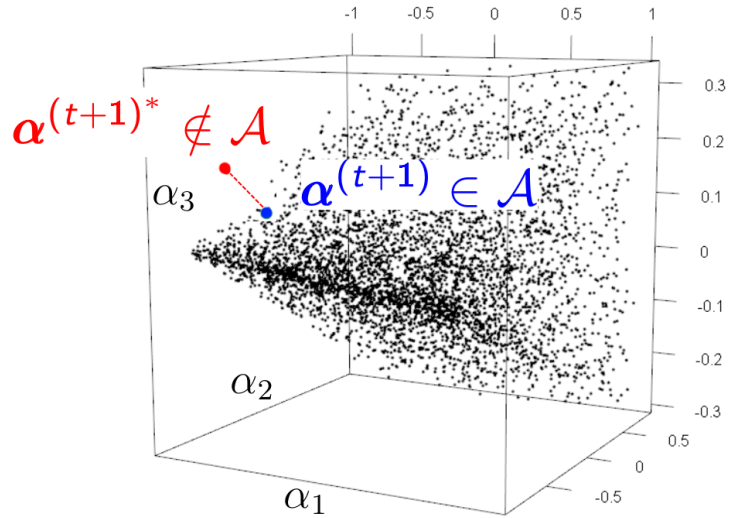
and optimize in the reduced dimensional space:

- $\alpha^{(t+1)*}$ comes from $\max([\underbrace{\alpha_a, \overline{\alpha}}_{\in \mathbb{R}^{\delta+1}}])$,

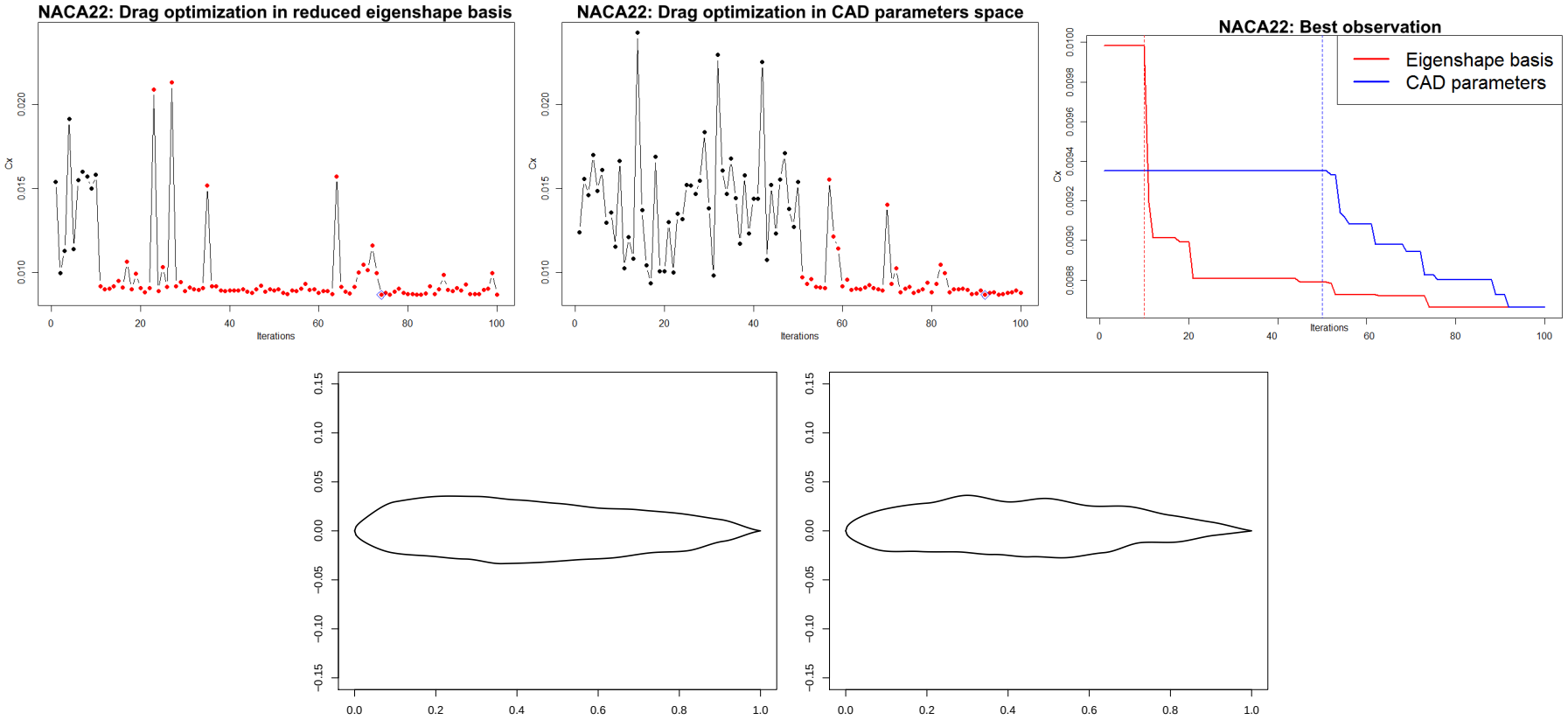




  $\overline{\alpha}$ coordinate along a random line in non-active space
- Solve pre-image problem:
  $x^{(t+1)} = \underset{x \in \mathcal{S}}{\arg\min} \, \|\mathbf{V}^\top(\phi(x) - \overline{\phi}) - \alpha^{(t+1)*}\|^2$
  and evaluate $f(x^{(t+1)})$. Eigencomp. $\alpha^{(t+1)} = \mathbf{V}^\top(\phi(x^{(t+1)}) - \overline{\phi})$
- 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

BO's performance degrades with dimensionality. 2 techniques for reducing dimensions in BO:

- variable selection specific to optimization because based on sublevel sets; select variables from a robust statistics, the maximum mean discrepancy in the RKHS; $x = (x_{\mathbf{a}}, x_{\bar{\mathbf{a}}})$, optimize on the active $x_{\mathbf{a}}$.
- build an embedding, $\phi(x)$, and identify its most active directions (eigenshapes), $V^{\mathbf{a}}$, from the regularized likelihood; build a GP and optimize with more details in $V^{\mathbf{a}}$ while not completely overlooking $V^{\bar{\mathbf{a}}}$.
- Perpectives: generalize and cumulate: create embeddings for general optimization problem and select variables from sublevel sets in this better parameterized space.

References I


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


Durrande, N. and Le Riche, R. (2017).
Introduction to Gaussian Process Surrogate Models.
Lecture at 4th MDIS form@ter workshop, Clermont-Fd, France.
HAL report cel-01618068.


Fukumizu, K., Gretton, A., Lanckriet, G. R., Schölkopf, B., and Sriperumbudur, B. K. (2009).
Kernel choice and classifiability for rkhs embeddings of probability distributions.
In *Advances in neural information processing systems*, pages 1750–1758.


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.

References II


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.


Hansen, N. and Ostermeier, A. (2001).
Completely derandomized self-adaptation in evolution strategies.
Evol. Comput., 9(2):159–195.


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.


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.


Raghavan, B., Le Quilliec, G., Breitkopf, P., Rassineux, A., Roelandt, J.-M., and Villon, P. (2014).
Numerical assessment of springback for the deep drawing process by level set interpolation using shape manifolds.
International journal of material forming, 7(4):487–501.

References III


Smola, A., Gretton, A., Song, L., and Schölkopf, B. (2007).
A hilbert space embedding for distributions.
In *International Conference on Algorithmic Learning Theory*, pages 13–31. Springer.


Sobol, I. M. (1993).
Sensitivity estimates for nonlinear mathematical models.
Math. Model. Comput. Exp, 1(4):407–414.


Spagnol, A. (2020).
Kernel-based sensitivity indices for high-dimensional optimization problems.
PhD thesis, École Nationale Supérieure des Mines de Saint-Etienne.


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.


Stegmann, M. B. and Gomez, D. D. (2002).
A brief introduction to statistical shape analysis.
Informatics and mathematical modelling, Technical University of Denmark, DTU, 15(11).

References IV

Yi, G., Shi, J., and Choi, T. (2011).
Penalized Gaussian process regression and classification for high-dimensional nonlinear data.
Biometrics, 67(4):1285–1294.