Bayesian and evolutionary optimization: searching from both sides

Rodolphe Le Riche

Bayesian and evolutionary optimization: searching from both sides

Rodolphe Le Riche*

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

11 July 2023
talk at the Hasso-Plattner Institut, Potsdam

Context: global optimization

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

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

Both evolutionary and Bayesian optimization algorithms (EO & BO) apply.
EO & BO have different scientific communities (computer science vs. applied math, but it is changing).

Goal of the talk : introduce BO through a comparison with EO and draw common perspectives.

Context: BO for costly optimization

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

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 $\longrightarrow$ metamodel / surrogate based optimization. Gaussian process as metamodel : Bayesian Optimization (BO).

Beginnings and references

EO & BO both date from the 60s or 70s with the work of

- [Fogel et al., 1965], [Holland, 1973], [Rechenberg, 1973] and [Schwefel, 1977] for EO,
- and [Kushner, 1962], [Močkus, 1972] and [Saltenis, 1971] for BO.

Reference texts for BO : [Frazier, 2018],[Garnett, 2023].

Algorithms skeletons I

BO

1. make an initial design of experiments (DoE) $\mathbb{X}^t$ and calculate $\mathbb{F}^t = f(\mathbb{X}^t)$, $t = \text{length}(\mathbb{F})$
2. build a Gaussian Proc. from $(\mathbb{X}^t, \mathbb{F}^t) \rightarrow GP^t$
3. $x^{t+1} =$ $\arg\max_{x \in \mathcal{S}} \text{AcquisCrit}(x; GP^t)$
4. $(\mathbb{X}^{t+1}, \mathbb{F}^{t+1}) =$ $(\mathbb{X}^{t+1}, \mathbb{F}^{t+1}) \cup (x^{t+1}, f(x^{t+1}))$, increment t
5. stop $(t > t^{\max}, \dots)$ or go to 2.

EO

1. make an initial DoE $\mathbb{X}$ and calculate $\mathbb{F} = f(\mathbb{X})$, $t = \text{length}(\mathbb{F})$
2. Apply variation operators (selection, mutation, crossover) to get a new DoE: $\mathbb{X}' = \text{Variation}(\mathbb{X}, \mathbb{F})$
3. calculate $\mathbb{F}' = f(\mathbb{X}')$, $t = t + \text{length}(\mathbb{F}')$
4. Create a new $(\mathbb{X}, \mathbb{F})$ from old $(\mathbb{X}, \mathbb{F})$ and $(\mathbb{X}', \mathbb{F}')$
5. stop $(t > t^{\max}, \dots)$ or go to 2.

Algorithms skeletons II

BO

1. make an initial design of experiments (DoE) $\mathbb{X}^t$ and calculate $\mathbb{F}^t = f(\mathbb{X}^t)$, $t = \text{length}(\mathbb{F})$
2. build a Gaussian Proc. from $(\mathbb{X}^t, \mathbb{F}^t) \rightarrow GP^t$
3. $x^{t+1} = \arg\max_{x \in \mathcal{S}} \text{AcquisCrit}(x; GP^t)$
4. $(\mathbb{X}^{t+1}, \mathbb{F}^{t+1}) = (\mathbb{X}^{t+1}, \mathbb{F}^{t+1}) \cup (x^{t+1}, f(x^{t+1}))$, increment t
5. stop ($t > t^{\max}, \ldots$) or go to 2.

EO

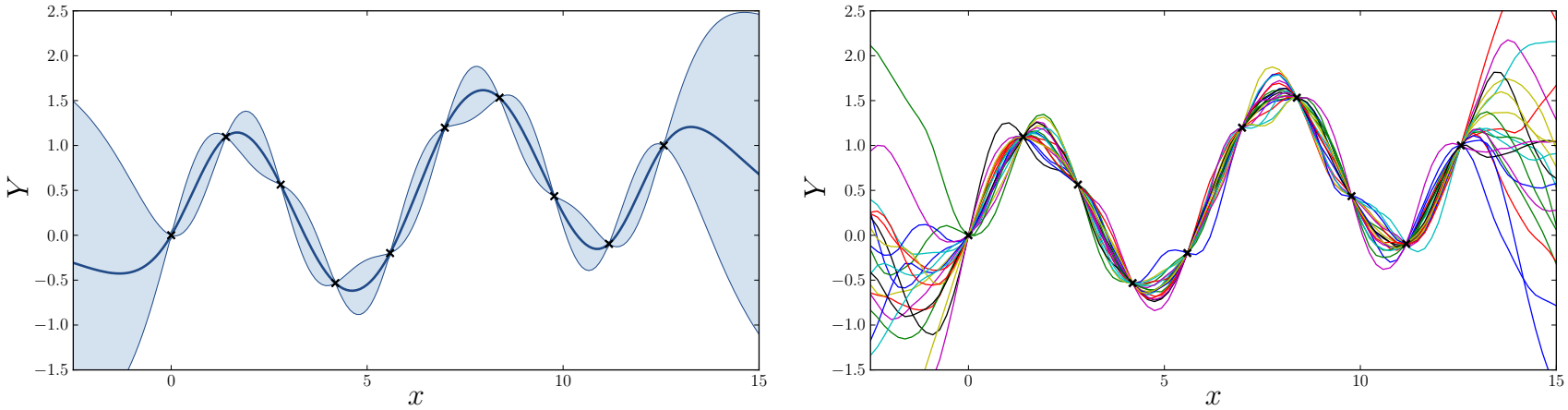
1. make an initial DoE $\mathbb{X}$ and calculate $\mathbb{F} = f(\mathbb{X})$, $t = \text{length}(\mathbb{F})$
2. Apply variation operators (selection, mutation, crossover) to get a new DoE: $\mathbb{X}' = \text{Variation}(\mathbb{X}, \mathbb{F})$
3. calculate $\mathbb{F}' = f(\mathbb{X}')$, $t = t + \text{length}(\mathbb{F}')$
4. Create a new $(\mathbb{X}, \mathbb{F})$ from old $(\mathbb{X}, \mathbb{F})$ and $(\mathbb{X}', \mathbb{F}')$
5. stop ($t > t^{\max}, \ldots$) or go to 2.

Differences:

- Maximization of an acquisition criterion that depends on a Gaussian Process instead of the variation operators;
- EO is population oriented while all points are kept in BO. BO would be memory consumming for large t but not its typical range of use. The state of EO is implicitely stored in its population, efficient but loss prone.

A SHORT INTRODUCTION TO BAYESIAN OPTIMIZATION

Gaussian Process Regression (kriging) – I

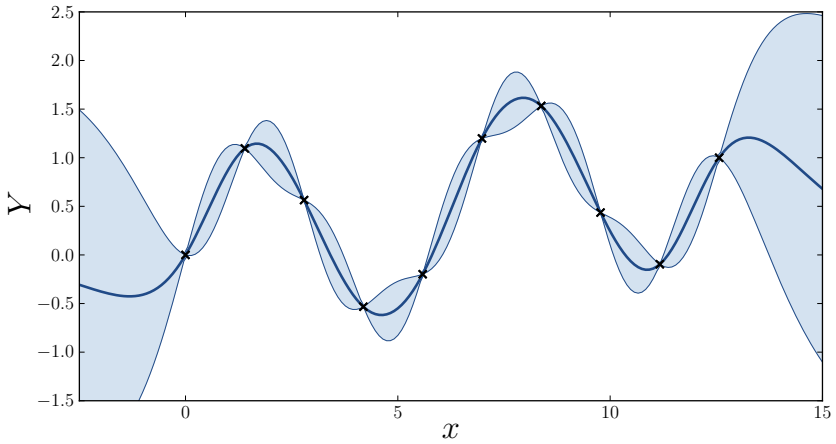


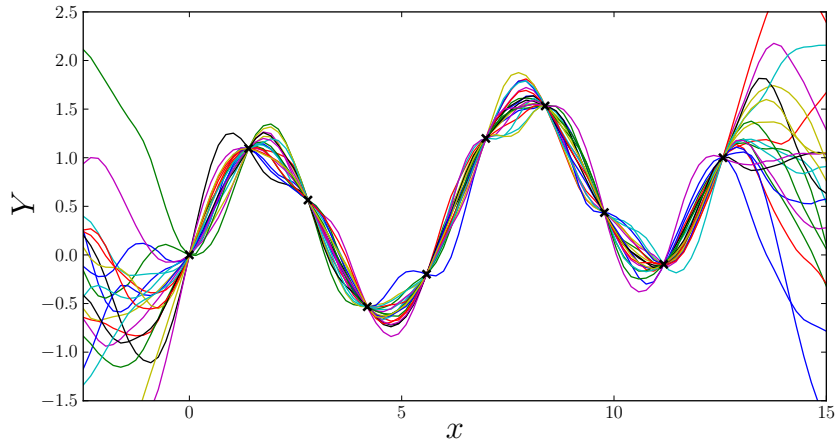
$Y(x)$ is $\mathcal{N}(\mu(x), k(x,x'))$
$Y(x)|Y(\mathbb{X})=\mathbb{F} \sim \mathcal{N}(m(x), c(x,x'))$ is also Gaussian, interpolating and depends on the kernel $k(.,.)$ and $\mu(.)$ through parameters θ.

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

Note: kernels are defined over all kinds of spaces. Example of mixed continuous-discrete BO in [Cuesta-Ramirez et al., 2022].

Gaussian Process Regression (kriging) – II





$Y(x)|Y(\mathbb{X})=\mathbb{F} \sim \mathcal{N}(m(x), c(x,x'))$, interpolating and depends on the kernel $k(.,.)$ and $\mu(.)$ through parameters θ.

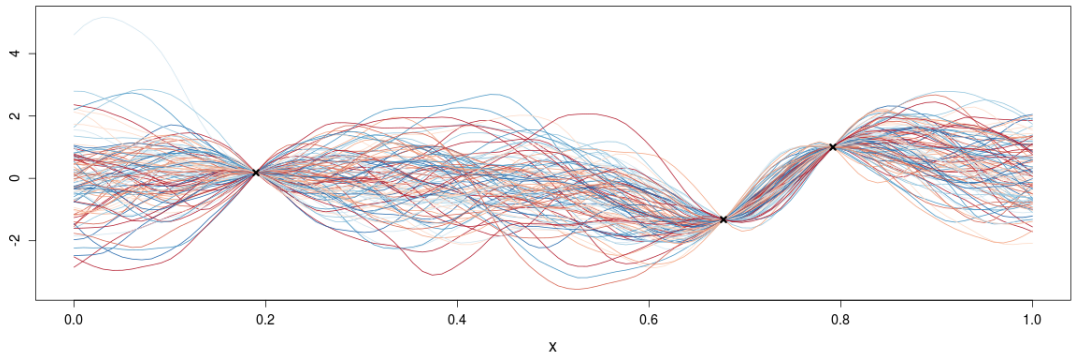
$$m(x) = \mathbb{E}[Y(x)|Y(\mathbb{X})=\mathbb{F}] = \mu(x) + k(x,\mathbb{X})k(\mathbb{X},\mathbb{X})^{-1}(\mathbb{F} - \mu(x)\mathbb{1})$$

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

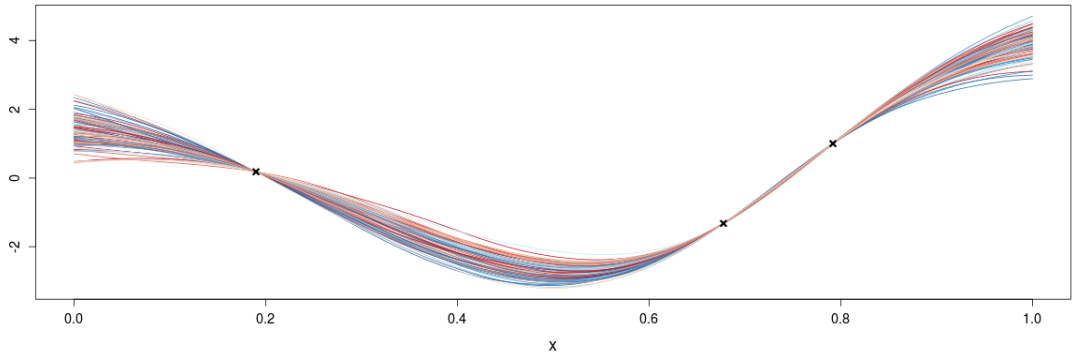
Learn the GP typically by max. (log) likelihood or leave-one-out crossvalidation, $\theta^\star = \arg\max_\theta LL(\theta;\mathbb{F})$.

Gaussian Process Regression (kriging)

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



$\theta = 0.1$



$\theta = 0.5$

(Matérn kernel, $\sigma = 1$)

An example of acquisition criterion: the $\mathbb{E}I$

Measure of progress: the improvement,

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

Acquisition criterion: $\mathbb{E}I$ [Saltenis, 1971, Schonlau, 1997], to maximize at each iteration.

The $\mathbb{E}I$ is analytically known and does not involve calls to f,

$\mathbb{E}I(x) = \int_{-\infty}^{+\infty} I(x)\, dy(x) = \cdots =$

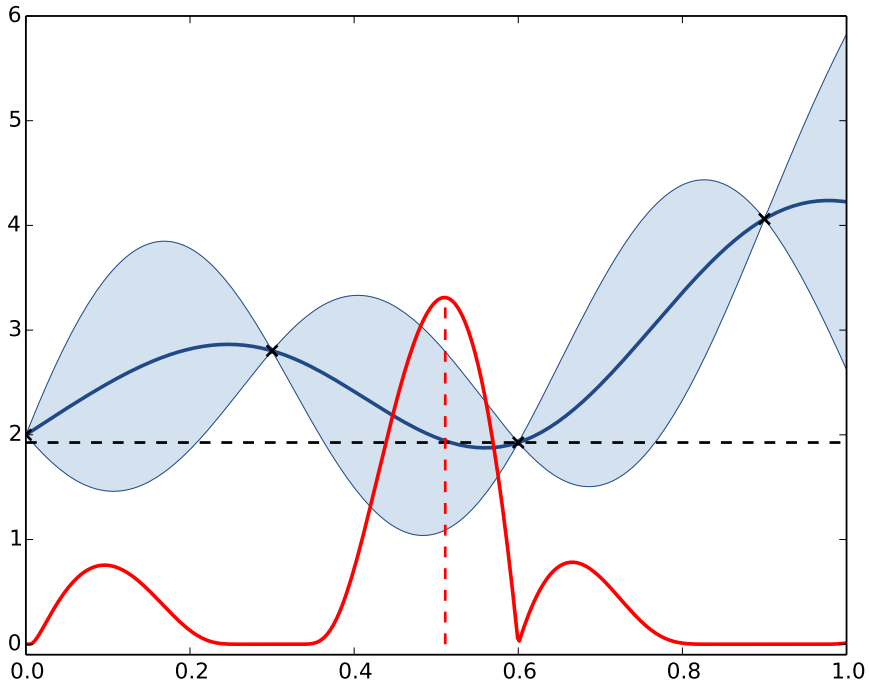
$$\sqrt{c(x,x)}\,[w(x)\mathrm{cdf}_{\mathcal{N}}(w(x)) + \mathrm{pdf}_{\mathcal{N}}(w(x))]$$

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

A parameter-less intensification (1st term) – exploration (2nd term) compromise.

Expected Improvement

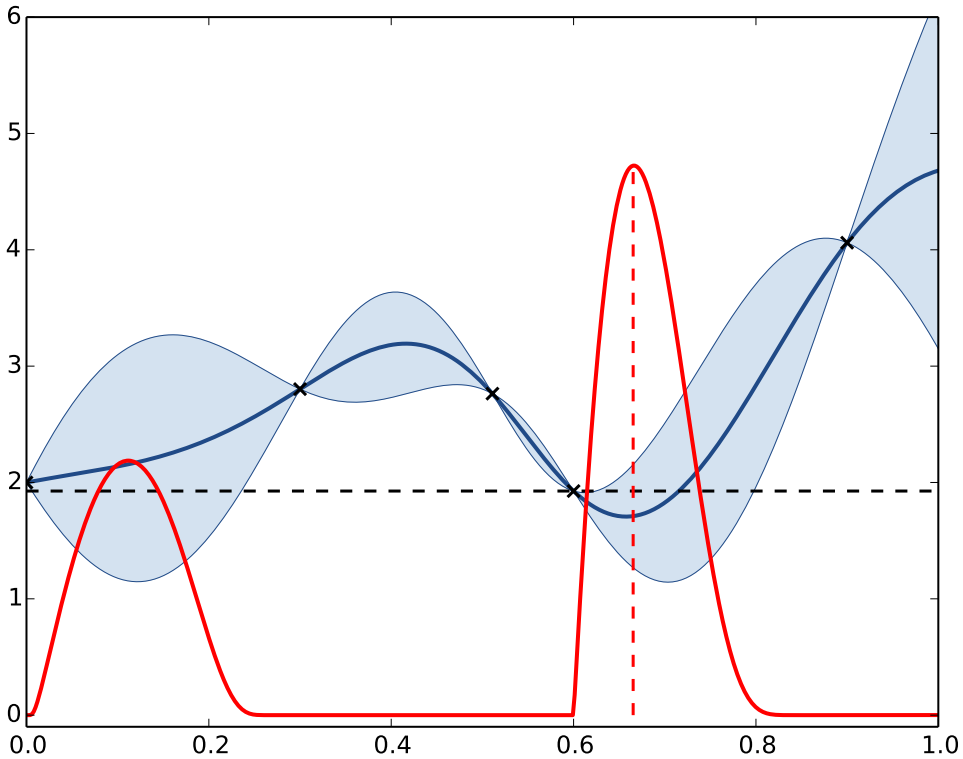
Let's see how it works... iteration 4



Expected Improvement

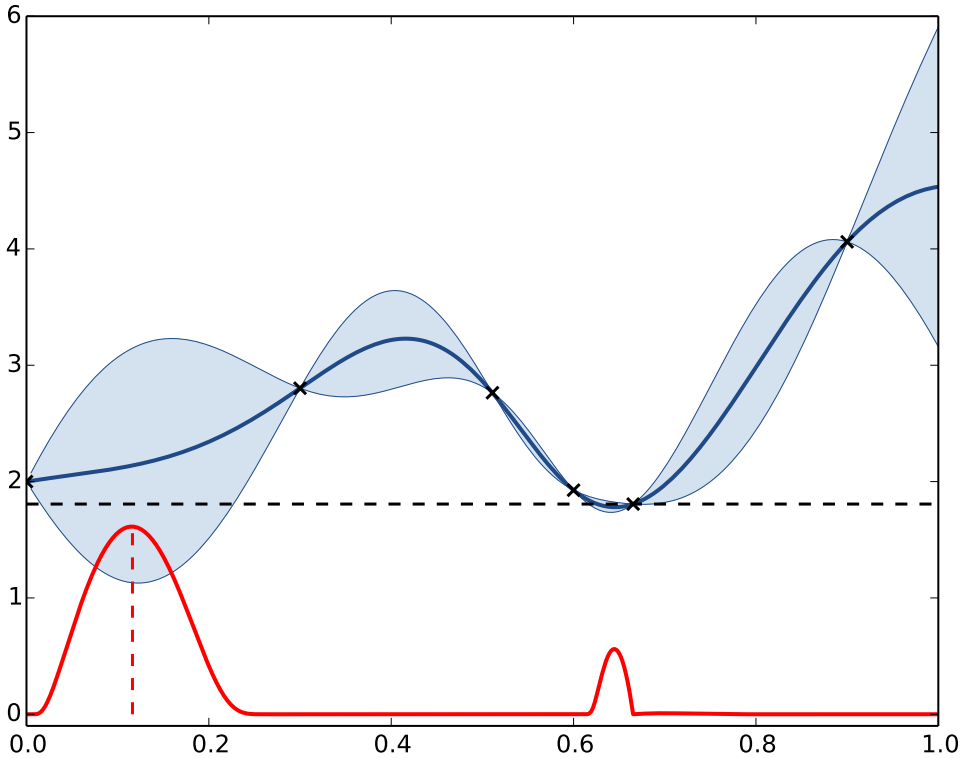
$$x^{t+1} = \arg\max_{x \in \mathcal{S}} \mathbb{E}I(x)$$

Let's see how it works... iteration 4+1



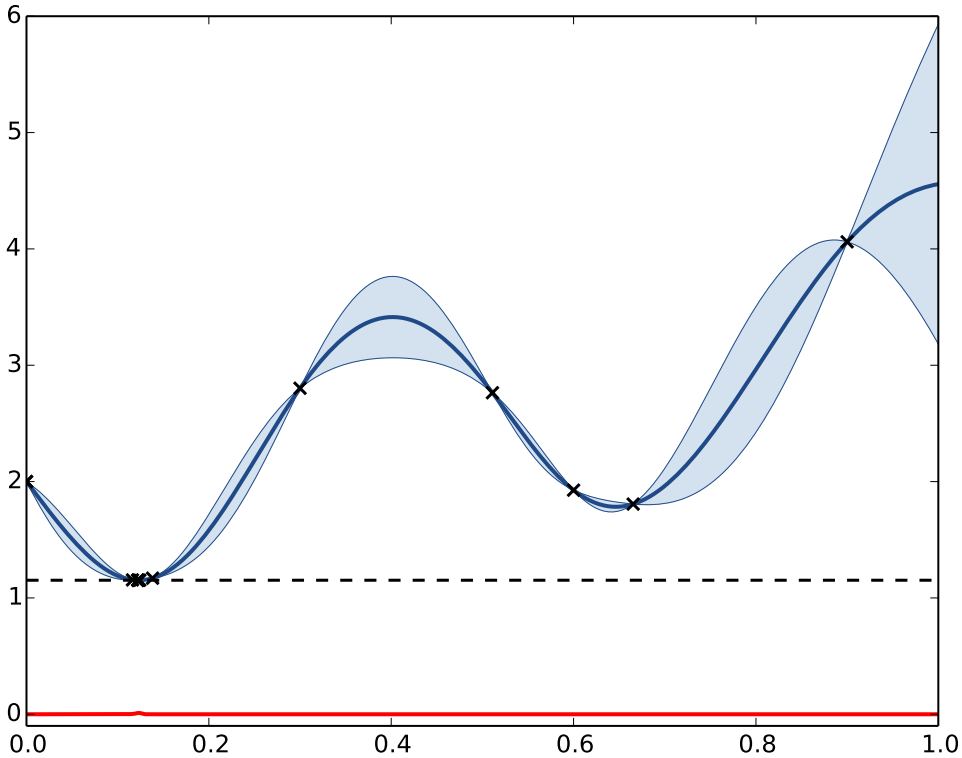
Expected Improvement

$x^{t+1} = \arg\max_{x \in \mathcal{S}} \mathbb{E}I(x)$... iteration 4+2



Expected Improvement

$x^{t+1} = \arg\max_{x \in \mathcal{S}} \mathbb{E}I(x)$... iteration 4+5



Expected Improvement: comments

There are other acquisition criteria:

- Upper Confidence Bound [Kushner, 1962, Srinivas et al., 2010], $= m(x) - \alpha c(x, x)$. Need to choose α.
- Knowledge gradient [Močkus, 1972, Frazier and Powell, 2007], $\min_{x \in \mathcal{S}} m^t(x) - \mathbb{E}[\min_{x \in \mathcal{S}} m^{t+1}(x)]$. No analytical expression.
- Criteria based on the reduction in entropy of the optimum [Villemonteix et al., 2009, Hernández-Lobato et al., 2014]. No analytical expression, complex.
- ...
- $\mathbb{E}I$ is a myopic criterion compared to the above criteria. It is still a good entry criterion thanks to its simplicity and ease of interpretation.

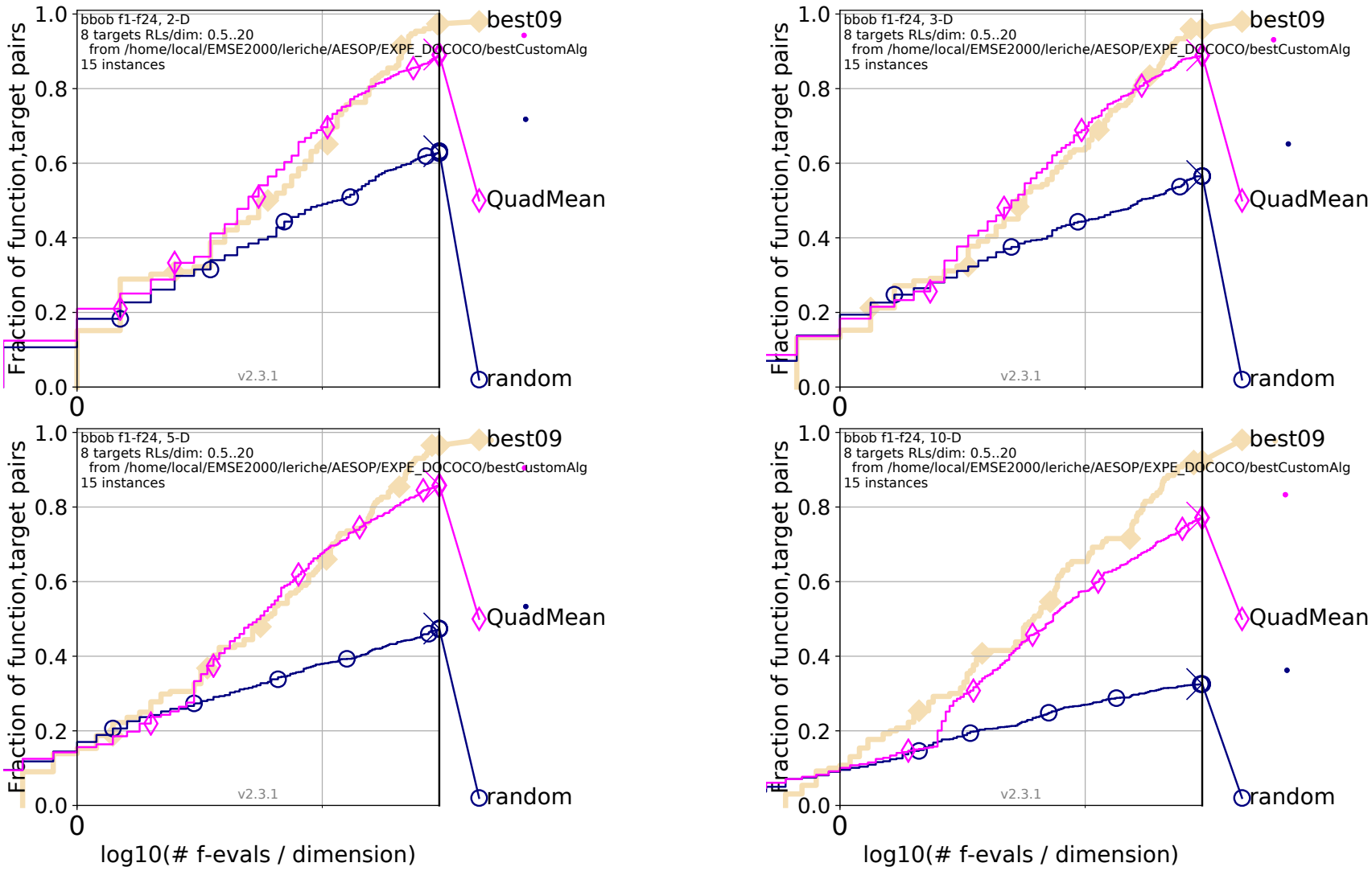
Bayesian optimization and dimension

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

- $\Leftarrow$ Loss of GP accuracy?
- $\Leftarrow$ EI sample too often at boundary?

Next slide: COCO tests [Hansen et al., 2016] from [Le Riche and Picheny, 2021]

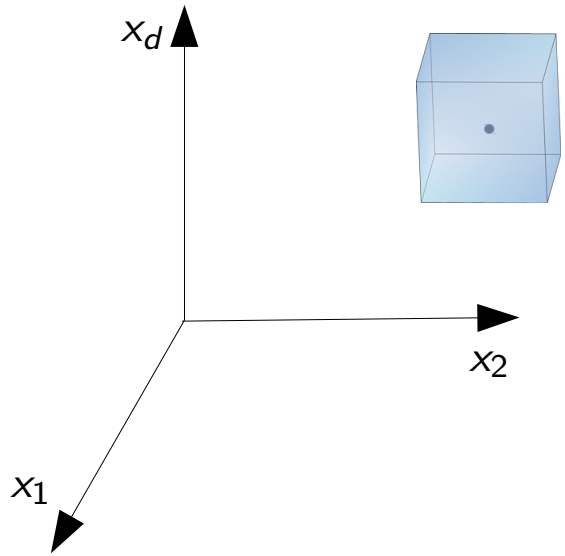
Effect of dimension on Bayesian Optimization



QuadMean = BO with quadratic mean, runs length = 30 d.

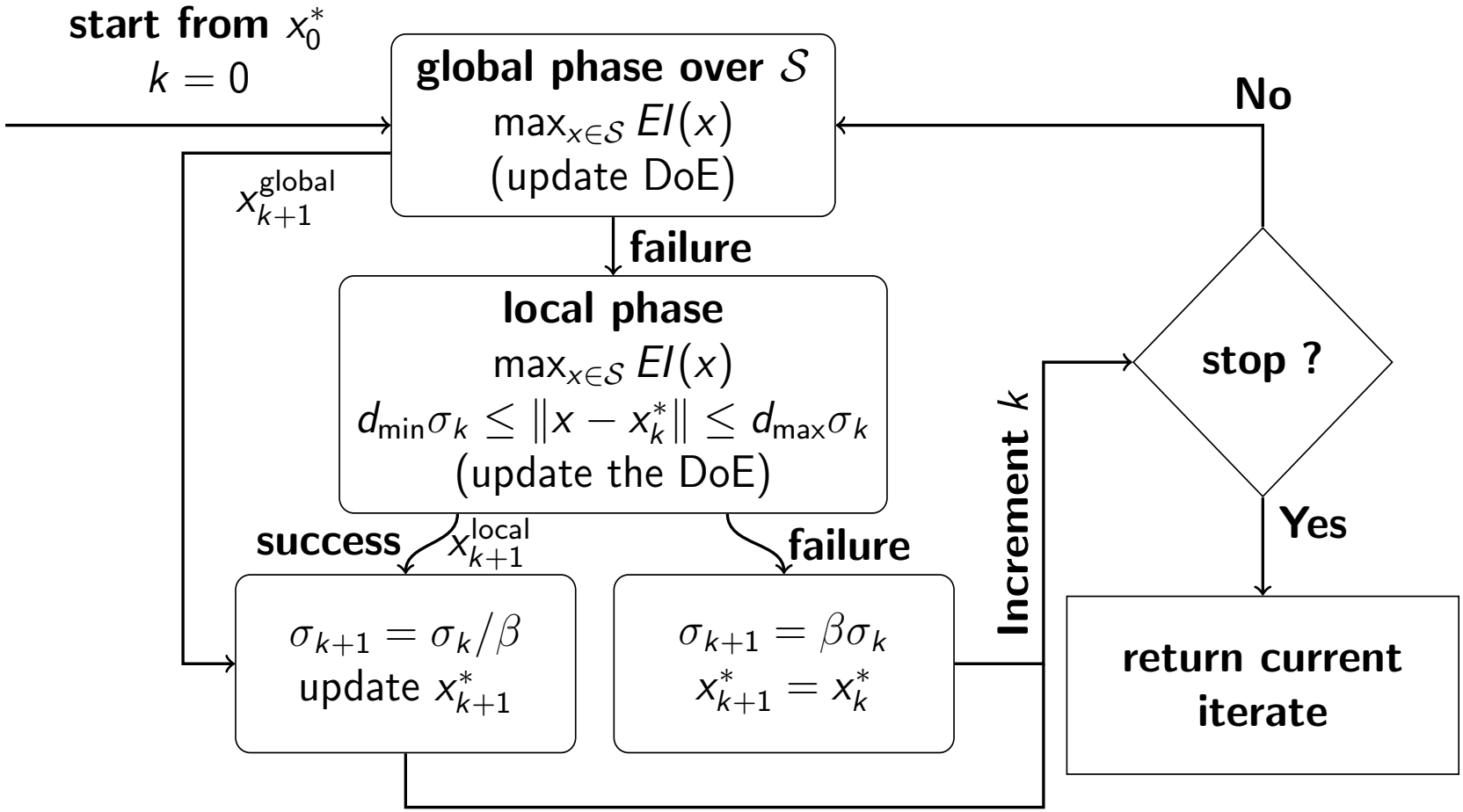
BO and trust regions

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., 2023] : mix searches inside (local) and outside (global) the trust region.

TREGO algorithm



Parameters : σ_0, $\beta < 1$

Sufficient decrease condition for success of the local phase,

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

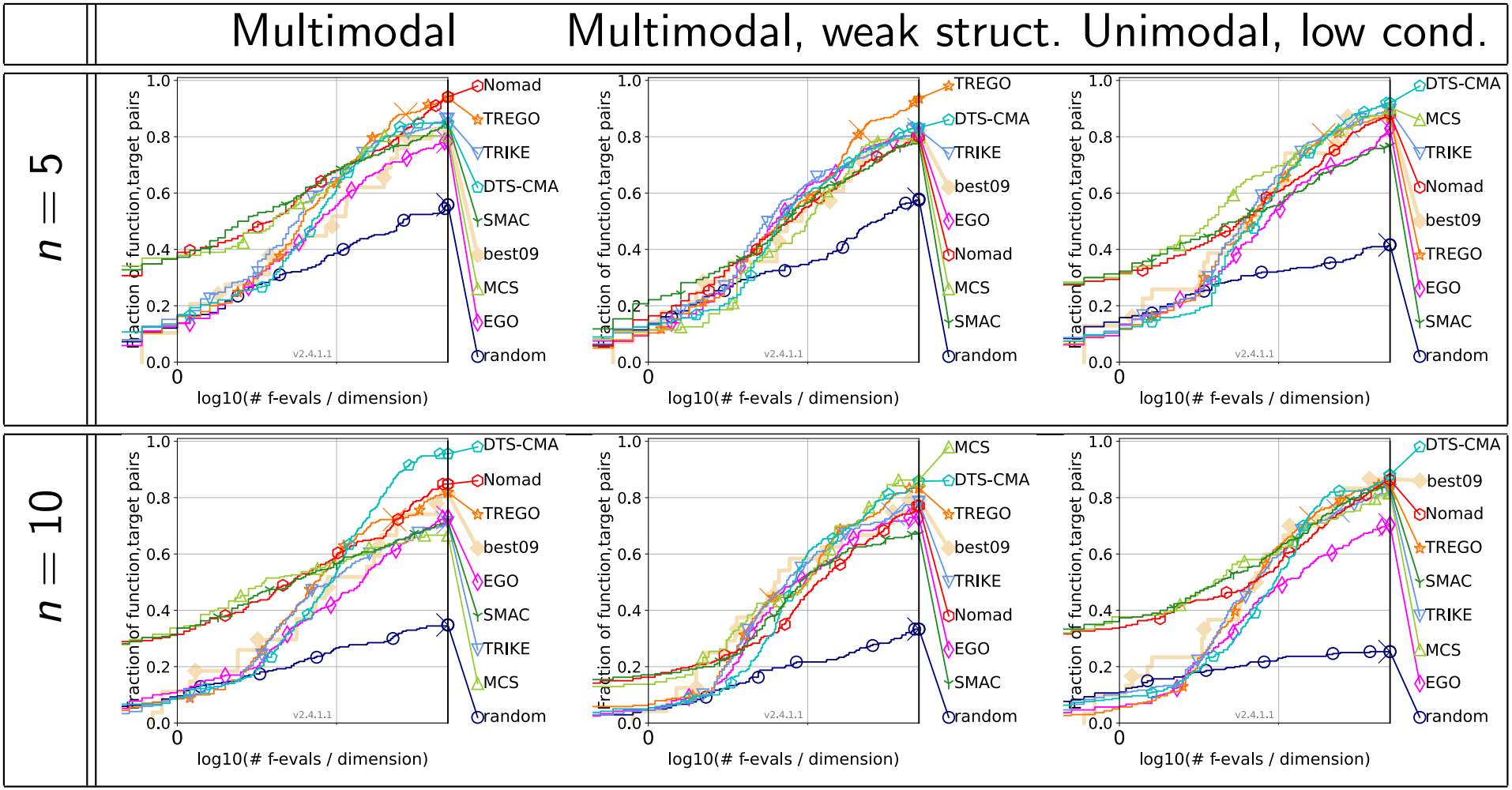
TREGO properties

From [Diouane et al., 2023],

- 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).

RELATIONS BETWEEN BAYESIAN AND EVOLUTIONARY OPTIMIZATION

The probabilistic representations of f

BO and EO represent f differently

BO	**EO**
Global model of the function (in $\mathbb{R}$)	Model the distribution of good points (in $\mathcal{S}$)
$p(f \mid x) = \mathcal{N}(m(x), c(x,x))$	$p(x \mid f(x) \leq \text{threshold}) \approx \frac{1}{\text{card}(\mathbb{X})} \sum_{x_i \in \mathbb{X}} \delta_{x_i}(x)$

Both views are tied by Bayes rule (see later).

Fundamental degrees of freedom

BO

Choosing and learning the kernel = choosing the feature space of the $\phi(x)$'s: by Mercer's theorem, $k(x, x') = \sum_{i=1}^{D} \lambda_i \phi_i(x) \phi_i(x')$ (N possibly infinite).
Quality metrics: regression metrics (Q2, LOO normality test).

Acquisition criterion: controls the exploration-intensification tradeoff.
Quality metrics: optimization metrics (convergence, performance profiles,...).

EO

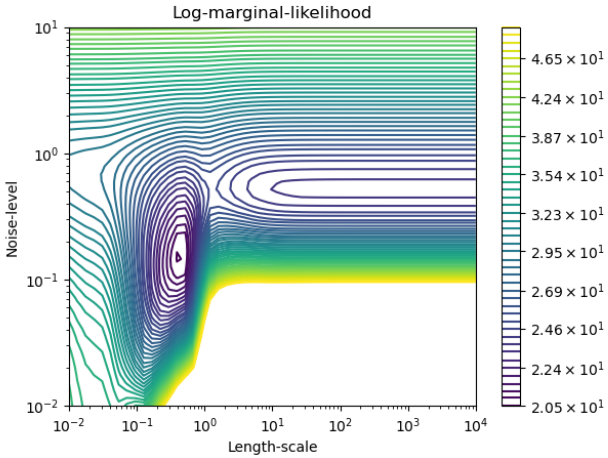
Choosing and learning the representation and variation operators.
Quality metrics: ergodicity (completness), unbiasedness, locality (genotype-phenotype local correlation).

Both views will be linked later (perspectives).

EO for BO

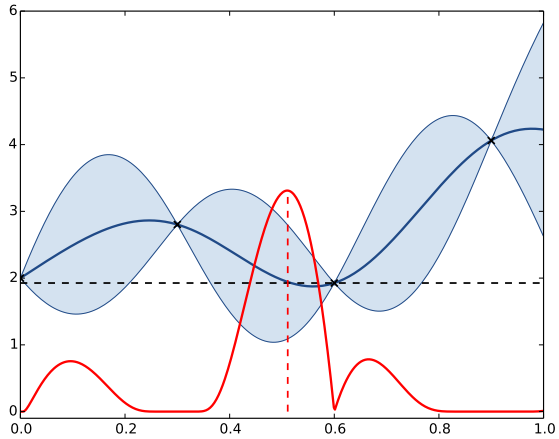
BO has 2 internal optimizations that are not costly and multimodal: they benefit from EO (CMA-ES is a typical choice for maximizing the $\mathbb{E}I$).

$\max_\theta$ likelihood$(\theta; \mathbb{X}, \mathbb{F})$



(from `scikit-learn.org`)

$\max_x \mathbb{E}I(x; GP)$



BO for EO: adding a GP – I

Principle : EO algorithms need to order new x's for selection.
Replace calls to f by calls to GPs.

Like surrogate-based EO, but with prediction uncertainties. Ranking from

- UCB [Büche et al., 2005], $m(x) - \alpha c(x,x)$,
- prob. of improvement [Ulmer et al., 2003], $\text{cdf}_{\mathcal{N}}\left((\min \mathbb{F} - m(x))/\sqrt{c(x,x)}\right)$,
- partial (Pareto) ranks of (function prediction,- function uncertainty) [Volz et al., 2017], $(m(x), -\sqrt{c(x,x)})$,
- . . . cf. [Bajer et al., 2019] for other criteria.

BO for EO: adding a GP – II

Algorithms are variants around the following sequence:

- $\mathbb{X}' = \mathit{Variation}(\mathbb{X})$
- Rank points in $\mathbb{X}'$ using the GP and one of the ranking criteria (UCB, PI, $\mathbb{E}I$, m, ...)
- Check the quality of the GP:
  - number of generations without GP update
  - OR compare ranking from true f and from GP on a small portion of $\mathbb{X}'$
- GP update from a reduced number of points chosen in $\mathbb{X}' \cup \mathbb{X}$ where true f is calculated

a sequence which is decomposed, repeated, varied ... inside EO implementations.

The aforementionned interfaces between BO and EO are external.

PERSPECTIVES:

A DEEPER LINK BETWEEN EO AND BO

through 2 examples.

Crossover as a barycenter in feature space I

Summary

- Feature learning (from data, through the kernels) can be seen as a search for an appropriate representation where linear reasoning works.
- Interprete crossover as a barycenter in feature space.

A kernel can always be seen as an inner product in feature space[†], $k(x, x') = \langle \phi(x), \phi(x') \rangle$, where the feature is a map $\phi : \mathcal{S} \to \mathbb{R}^D$ and D can be infinite.

[†] For example it can be the eigendecomposition of Mercer's theorem seen earlier, $k(x, x') = \sum_{i=1}^{\infty} \sqrt{\lambda_i}\phi_i(x)\sqrt{\lambda_i}\phi_i(x')$ with the usual L2 scalar product, or the canonical decomposition in the RKHS, $k(x, x') = \langle k(x, .), k(x', .) \rangle_{\mathcal{H}}$.

Crossover as a barycenter in feature space II

Definition (Crossover as barycenter of features)

Let $x_C \in \mathcal{S}$ be the result of the crossover of $x_A \in \mathcal{S}$ with $x_B \in \mathcal{S}$. The features of x_C are the α-barycenter of the features of x_A and x_B, $0 \leq \alpha \leq 1$:

$$x_C := \arg\min_{x \in \mathcal{S}} \alpha \|\phi(x) - \phi(x_A)\|^2 + (1-\alpha)\|\phi(x) - \phi(x_B)\|^2$$

Crossover as a barycenter in feature space III

Calculation:
$\|\phi(x) - \phi(x_A)\|^2 = \langle \phi(x) - \phi(x_A), \phi(x) - \phi(x_A) \rangle$
$= \langle \phi(x), \phi(x) \rangle + \langle \phi(x_A), \phi(x_A) \rangle - 2\langle \phi(x), \phi(x_A) \rangle$
$= k(x, x) + k(x_A, x_A) - 2k(x, x_A)$
and idem with $\|\phi(x) - \phi(x_B)\|^2$.
Assume stationarity, $k(x, x) = k(x_A, x_A) = k(x_B, x_B) = \sigma^2$, then

$$x_C = \arg\min_{x \in \mathcal{S}} 2\sigma^2 - 2\alpha k(x, x_A) - 2(1 - \alpha)k(x, x_B)$$

or equivalently

$$x_C = \arg\max_{x \in \mathcal{S}} \alpha k(x, x_A) + (1 - \alpha)k(x, x_B)$$

i.e., maximize weighted similarities with x_A and x_B as seen by the (learned) kernels.

For a crossover, choose $\alpha \sim \mathcal{U}[0, 1]$.

A Bayesian variation operator I

Link the probabilistic representations of BO and EO (cf. earlier slide) to define a variation (crossover and mutation) operator.

- T, a threshold (e.g., $T = \min \mathbb{F}$)
- $F(x) := Y(x) \mid f(\mathbb{X}) = \mathbb{F} \sim \mathcal{N}(m(x), c(x,x))$, the GP
- $F(x) < T$, a logical random variable. It can be written $F(X) < T \mid X = x$, X random variable in $\mathcal{S}$
- Bayes:
$$p(F(X) < T \mid X = x) \,.\, p(X = x) = p(X = x | F(X) < T) \,.\, p(F(X) < T)$$
- GP : $p(F(X) < T \mid X = x) = \text{cdf}_{\mathcal{N}}\left((T - m(x))/\sqrt{c(x,x)}\right)$
- $p(X = x) = \mathcal{U}[\mathcal{S}]$. Other more intensifying choices possible, at the risk of premature convergence.

A Bayesian variation operator II

$p(X = x|F(X) < T)$ is the model for the variation operator of EO. It is sampled through

$$p(X = x|F(X) < T) \propto \text{cdf}_{\mathcal{N}}\left((T - m(x))/\sqrt{c(x,x)}\right) p(X = x)$$

by a Metropolis-Hastings (or any appropriate MCMC) scheme.

References I


Bajer, L., Pitra, Z., Repickỳ, J., and Holeňa, M. (2019).
Gaussian process surrogate models for the CMA evolution strategy.
Evolutionary computation, 27(4):665–697.


Büche, D., Schraudolph, N. N., and Koumoutsakos, P. (2005).
Accelerating evolutionary algorithms with Gaussian process fitness function models.
IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 35(2):183–194.


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, 9(1).


Diouane, Y., Picheny, V., Le Riche, R., and Di Scotto, A. (2023).
TREGO: a trust-region framework for efficient global optimization.
Journal of Global Optimization, 86(1):1–23.


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.

References II


Fogel, L., Owens, A., and Walsh, M. (1965).
Artificial intelligence through a simulation of evolution, in biophysics and cybernetics systems.
In *Proceedings of the Second Cybernetics Sciences Symposium, ed. by M. Maxfield, A. Callahan, and LJ Fogel, Spartan Books, Washington*.


Frazier, P. and Powell, W. (2007).
The knowledge gradient policy for offline learning with independent normal rewards.
In *2007 IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning*, pages 143–150. IEEE.


Frazier, P. I. (2018).
A tutorial on Bayesian optimization.
arXiv preprint arXiv:1807.02811.


Garnett, R. (2023).
Bayesian optimization.
Cambridge University Press.


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 III


Hernández-Lobato, J. M., Hoffman, M. W., and Ghahramani, Z. (2014).
Predictive entropy search for efficient global optimization of black-box functions.
Advances in neural information processing systems, 27:918–926.


Holland, J. H. (1973).
Genetic algorithms and the optimal allocation of trials.
SIAM journal on computing, 2(2):88–105.


Kushner, H. J. (1962).
A versatile stochastic model of a function of unknown and time varying form.
Journal of Mathematical Analysis and Applications, 5(1):150–167.


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.


Močkus, J. (1972).
On Bayesian methods of search for extremum.
Automatics and Computers, 3:53–62.

References IV


Rechenberg, I. (1973).
Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution.
Frommann-Holzboog-Verlag, Stuttgart.


Regis, R. G. (2016).
Trust regions in kriging-based optimization with expected improvement.
48:1037–1059.


Saltenis, V. R. (1971).
One method of multiextremum optimization.
Avtomatika i Vychislitel'naya Tekhnika (Automatic Control and Computer Sciences), 5(3):33–38.


Schonlau, M. (1997).
Computer experiments and global optimization.
PhD thesis, University of Waterloo.


Schwefel, H.-P. (1977).
Numerische Optimierung von Computer Modellen mittels der Evolutionsstrategie.
Birkhaeuser, Basel/Stuttgart.
volume 26 of ISR.

References V


Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. (2010).
Gaussian process optimization in the bandit setting: No regret and experimental design.
In *ICML*.


Ulmer, H., Streichert, F., and Zell, A. (2003).
Evolution strategies assisted by Gaussian processes with improved preselection criterion.
In *The 2003 Congress on Evolutionary Computation, 2003. CEC'03.*, volume 1, pages 692–699. IEEE.


Villemonteix, J., Vazquez, E., and Walter, E. (2009).
An informational approach to the global optimization of expensive-to-evaluate functions.
Journal of Global Optimization, 44(4):509.


Volz, V., Rudolph, G., and Naujoks, B. (2017).
Investigating uncertainty propagation in surrogate-assisted evolutionary algorithms.
In *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 881–888.