Sequential numerical designs for computer experiments

Pierre Barbillon

AgroParisTech / INRA MIA UMR 518

SPADRO 23/01/15

Outline

1. Computer experiments
   - Context
   - Meta-modeling
   - Design of numerical experiment

2. Sequential designs
   - Expected Improvement
   - Stepwise Uncertainty Reduction
   - Bandit optimization

Outline

Outline

Context

- A computer experiment is an evaluation of a determinist expensive black-box function describing a physical system:

$$f : \mathbf{x} \in E \subset \mathbb{R}^d \mapsto \mathbb{R}\,.$$

  Ref. : Fang et al. (2006), Koehler et Owen (1996), Santner et al. (2003).

- Expensive $\Leftrightarrow$ N possible calls to f.
- Uncertainties on inputs $\Rightarrow$ modelled by a random variable $\mathbf{X}$ (known distribution).

Goal

- Vizualisation,
- Optimization,
- Output distribution: mean, median, quantile, probability of rare events...
- Inverse problem.

Outline

Meta-modeling: prior distribution on f

[Sacks et al., 1989a].
f realization of a Gaussian process F / prior distribution F on f:
$\forall \mathbf{x} \in E$,

$$F(\mathbf{x}) = \sum_{k=1}^{Q} \beta_k h_k(\mathbf{x}) + \zeta(\mathbf{x}) = H(\mathbf{x})^T \boldsymbol{\beta} + \zeta(\mathbf{x}) \,,$$

where

- $h_1, \ldots, h_Q$ regression functions and $\boldsymbol{\beta}$ parameters vector,
- ζ centred Gaussians process with covariance function:

$$\mathrm{Cov}(\zeta(\mathbf{x}), \zeta(\mathbf{x}')) = \sigma^2 K(\mathbf{x}, \mathbf{x}') \,,$$

where K is a correlation kernel depending on parameters ψ.

Meta-modeling: posterior

- $y_1 = f(\mathbf{x}_1), \ldots, y_n = f(\mathbf{x}_n)$ evaluations of f on a design D_n ($n \leq N$).
- Conditioning F to $\{F(\mathbf{x}_1) = y_1, \ldots, F(\mathbf{x}_n) = y_n\} = F(D_n)$ and assuming parameters $\boldsymbol{\beta}, \sigma^2$ and ψ fixed;
  Gaussian Process with mean $m_{D_n}(\mathbf{x})$ and covariance $C_{D_n}(\mathbf{x}, \mathbf{x}')$ $\forall \mathbf{x}, \mathbf{x}'$.

  For all $\mathbf{x} \in E$,
  - $m_{D_n}(\mathbf{x})$ approximates $f(\mathbf{x})$,
  - $C_{D_n}(\mathbf{x}, \mathbf{x})$ uncertainty on this approximation.

For all $\mathbf{x}_i \in D_n$,
- $m_{D_n}(\mathbf{x}_i) = f(\mathbf{x}_i)$,
- $C_{D_n}(\mathbf{x}_i, \mathbf{x}_i) = 0$.

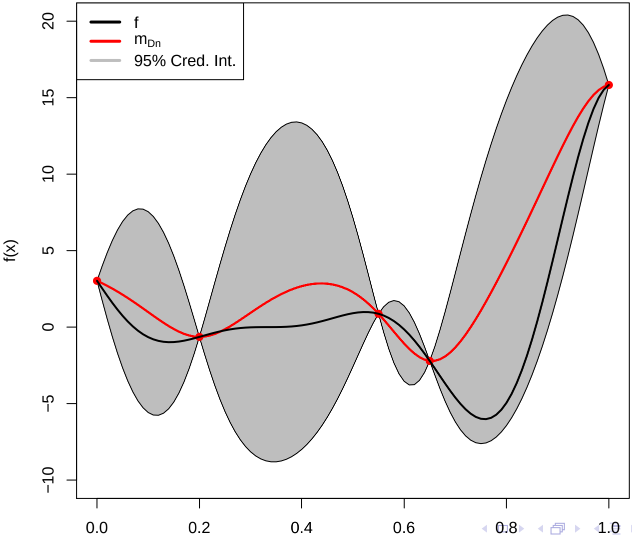
Conditioned GP distribution

For $\mathbf{x} \notin D_N$,

$$\begin{aligned} m_{D_N}(\mathbf{x}) &= H(\mathbf{x})^T \boldsymbol{\beta} + \Sigma_{\mathbf{x}D_N}^T \Sigma_{D_N D_N}^{-1} (z_{D_N} - H_{D_N} \boldsymbol{\beta}) \\ K_{D_N}(\mathbf{x}, \mathbf{x}) &= \sigma^2 (1 - \Sigma_{\mathbf{x}D_N}^T \Sigma_{D_N D_N}^{-1} \Sigma_{\mathbf{x}D_N}) \, , \end{aligned}$$

where

- $H_{D_N} = (H(\mathbf{x}_1), \ldots H(\mathbf{x}_N))^T$,
- $(\Sigma_{D_N D_N})_{1 \leq i,j \leq N} = \mathrm{corr}(F(\mathbf{x}_i), F(\mathbf{x}_j)) = K(\mathbf{x}_i, \mathbf{x}_j)$,
- $\Sigma_{\mathbf{x}D_N} = (\mathrm{corr}(F(\mathbf{x}_1), F(\mathbf{x})), \ldots \mathrm{corr}(F(\mathbf{x}_N), F(\mathbf{x})))^T$.


f
m_{Dn}
95% Cred. Int.
f(x)
20
15
10
5
0
−5
−10
0.0
0.2
0.4
0.6
0.8
1.0

Hyperparameters

[Santner et al., 2003]

1. Plug-in ML estimates for $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{\psi}$,
2. Prior distribution on $\boldsymbol{\beta}$ (Laplace distribution or Gaussian distribution), conditioned process still Gaussian with a closed-form for mean and covariance,
3. Prior distribution on σ, Student process and closed-form for mean and covariance,
4. Prior distribution on $\boldsymbol{\psi}$, no closed-form expression, posterior sampling....

Interpolation in RKHS

[Schaback, 1995]

- K positive definite kernel $\Rightarrow \mathcal{H}_K$ Reproducing Kernel Hilbert Space,
- If $f \in \mathcal{H}_K$, m_{D_n} interpolates f on D_n with minimal $||m_{D_n}||_{\mathcal{H}_K}$,
- $\forall \mathbf{x}_0$,

$$|f(\mathbf{x}_0) - m_{D_n}(\mathbf{x}_0)| \leq ||f||_{\mathcal{H}_K} \; C_{D_n}(\mathbf{x}_0, \mathbf{x}_0) \, .$$

For large class of kernels K,

$$\sup_{\mathbf{x} \in E} C_{D_n}(\mathbf{x}, \mathbf{x}) \leq G_K(u(D_n)) \, ,$$

with

- $G_K : h \mapsto G_K(h)$ non-decreasing, with limit 0 when $h \to 0$,
- $u(D_n) = \sup_{\mathbf{z} \in E} \min_{1 \leq j \leq n} \|\mathbf{z} - \mathbf{x}_j\|$ with $D_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$.

BUT

Driscoll's theorem: $\mathbb{P}(\omega : F(\omega) \in \mathcal{H}_K) = 0$

Outline

Space-Filling design

- Exploratory design,
- space-filling since the precision of the approximation depends on the distance between a new point and a design point,
  - low discrepancy designs, quasi-uniform sampling,
  - maximin and minimax designs,
  - optimal designs.

Minimax and Maximin criterion

[Johnson et al., 1990]

Definition

$D_n = \{\mathbf{x}_1, \ldots \mathbf{x}_n\}$ is a minimax design if it minimizes over E^n :

$$u(D_n) = \sup_{\mathbf{z} \in E} \min_{1 \le j \le n} \|\mathbf{z} - \mathbf{x}_j\| .$$

$$\sup_{\mathbf{x} \in E} C_{D_n}(\mathbf{x}, \mathbf{x}) \le G_K(u(D_n)) .$$

Definition

$D_n = \{\mathbf{x}_1, \ldots \mathbf{x}_n\}$ is a maximin design if it maximizes over E^n :

$$v(D_n) = \min_{1 \le i,j \le n} \|\mathbf{x}_i - \mathbf{x}_j\| .$$

If D_n is maximin, we have:

$$\sup_{\mathbf{x} \in E} C_{D_n}(\mathbf{x}, \mathbf{x}) \le G_K(u(D_n)) \le G_K(v(D_n)) .$$

maximin Latin Hypercube Sampling

[Morris and Mitchell, 1995]

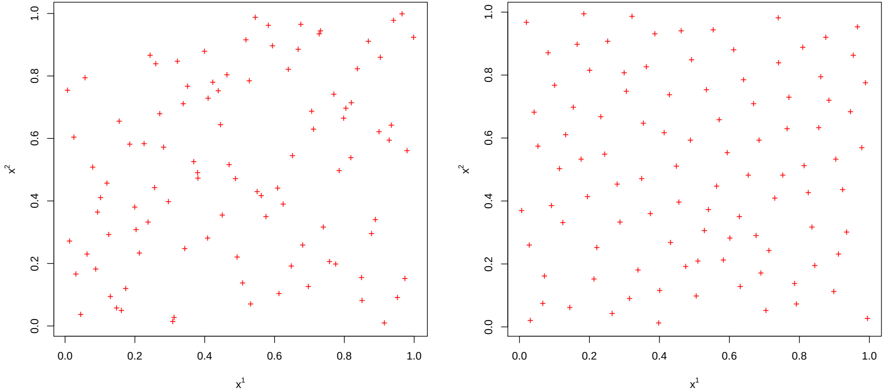


Figure : left: LHS, right: maximin-LHS

Outline

Overview

Based on a first exploratory design, procedure to augment the design sequentially:

- one step ahead,
- batch of experiments (parallelization),

to achieve:

- optimization of the code,
- better estimates of the probability of a rare event,
- better estimates of the quantiles,
- calibration of the computer code,
- ...

2 frameworks:

- **deterministic code**,
- stochastic code.

Principle

Algorithm

1. Construct a first exploratory design: D_n s. t. $n \leq N$ and estimate hyperparameters,
2. For $i = n + 1 ... N$ do
   1. find $\mathbf{x}_i$ which maximizes $Crit(\cdot|f(D_{i-1}))$
   2. $D_i = D_{i-1} \cup \{\mathbf{x}_i\}$ and update conditional process $F|f(D_i)$.

- $Crit(\cdot|f(D_{i-1}))$ is adapted to a given goal (optimization, estimation of a given probability, quantile, ...).
- It can be computed as an expectation with respect to the prior on f and the $i - 1$ first evaluations of f.
- Step updating the conditional process may be with fixed or re-estimated hyperparameters.

Different approaches

- Expected Improvement
  for efficient global optimization EGO [Jones et al., 1998]
- Stepwise Uncertainty Reduction for estimating probability of a rare event. [Chevalier et al., 2014, Picheny, 2014]
- Gaussian Process Upper Confidence Bound for optimization, quantile. [Srinivas et al., 2009, Contal et al., 2014, Contal et al., 2013, Grunewalder et al., 2010, Jala et al., 2014]

Outline

Expected Improvement criterion

- Goal: Find the global extremum (here minimum e.g.) of f,
- Expected improvement criterion proposed by [Jones et al., 1998]:

$$EI_{i-1}(\mathbf{x}) = \mathbb{E}((min_{i-1} - F(\mathbf{x}))^+ | F(D_{i-1})) \, ,$$

where min_n is the current minimum value:

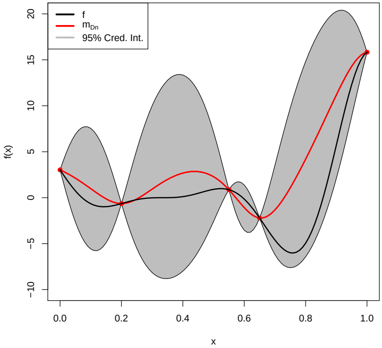
$$min_{i-1} = \min_{1,\ldots,i-1} f(\mathbf{x}_i)$$

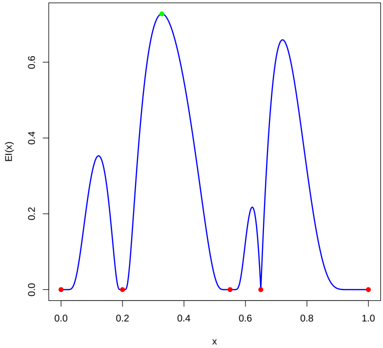
- Closed-form computation:

$$\begin{aligned} EI_{i-1}(\mathbf{x}) &= (min_{i-1} - m_{D_{i-1}}(\mathbf{x}))\Phi\left(\frac{min_{i-1} - m_{D_{i-1}}(\mathbf{x})}{\sqrt{C_{D_{i-1}}(\mathbf{x},\mathbf{x})}}\right) \\ &+ \sqrt{C_{D_{i-1}}(\mathbf{x},\mathbf{x})}\,\phi\left(\frac{min_{i-1} - m_{D_{i-1}}(\mathbf{x})}{\sqrt{C_{D_{i-1}}(\mathbf{x},\mathbf{x})}}\right) . \end{aligned}$$

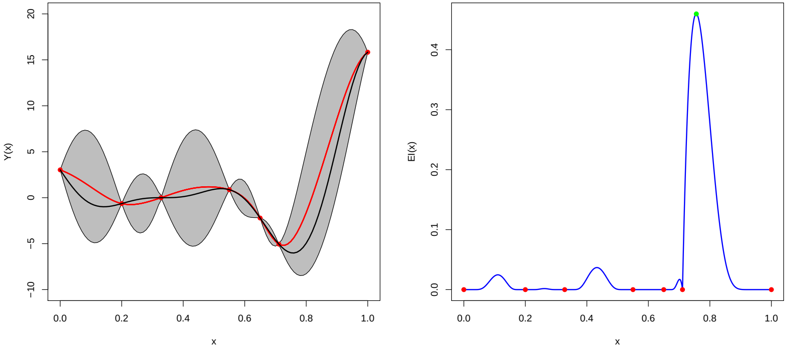
where Φ and ϕ are respectively the cdf and the pdf of $\mathcal{N}(0,1)$.

Example step 1

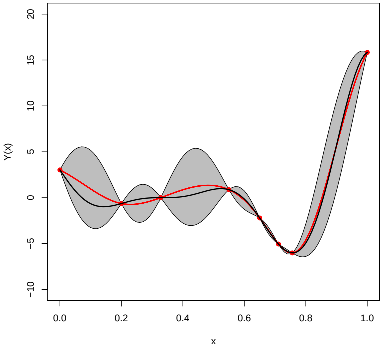


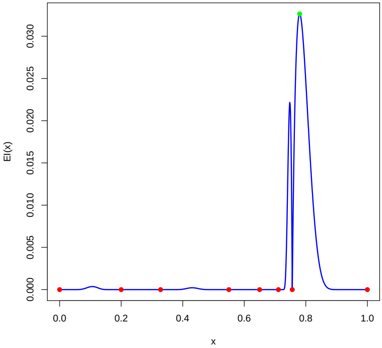


Example step 2

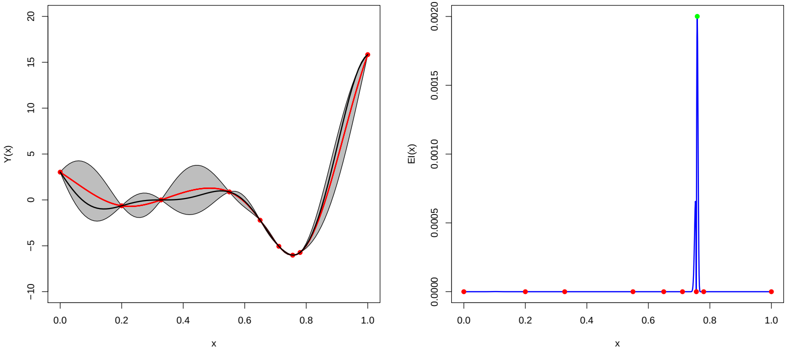


Example step 3

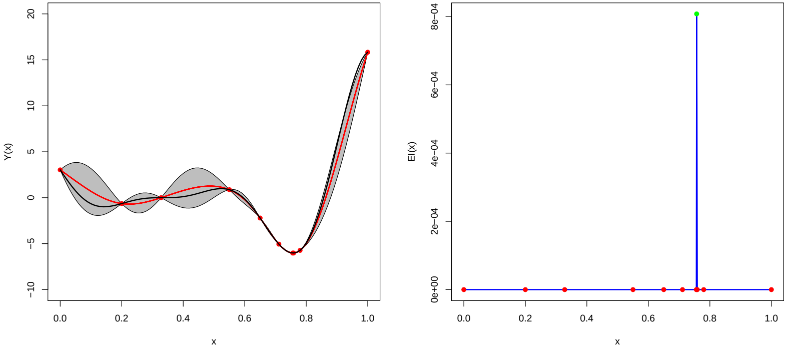




Example step 4



Example step 5



Theoretical results: dense in E

[Vazquez and Bect, 2010]

Theorem

If covariance function K satisfies to No-Empty-Ball property ($K(\mathbf{z}, \mathbf{x}_n) \to 0 \Rightarrow \mathbf{x}_n \to z$),

1. for all $x_{init} \in E$ and for all $h \in \mathcal{H}_K$, $(x_n)_n$ generated from a sequential algorithm based on maximization of the EI criterion is dense in E,
2. for all $x_{init} \in E$, $(X_n)_n$ generated from a sequential algorithm based on maximization of the EI criterion is $\mathbb{P}_F$ almost-surely dense in E.

Any fixed sequence (independent from the first evaluations to f) which fills E has the same theoretical guarantees.

Theoretical results: convergence rate

[Bull, 2011]

Theorem: Minimax convergence

For $\psi \in \mathbb{R}_+$ (length-scale), $R > 0$,

$$\inf_{u} \sup_{\|f\|_{\mathcal{H}_{K_\psi}} \leq R} \mathbb{E}^u \left(f(x_n^*) - \min_{\mathbf{x} \in E} f(\mathbf{x}) \right) = \Theta(n^{-\nu/d}),$$

where x_n^* is the estimated minimum location given n evaluations and $\nu < \infty$ (assumption $\hat{K} = \Theta_\infty(\|\mathbf{x}\|^{-2\nu-d}$, $\hat{K}$ is the Fourier transform of K) and u stands for a sequential strategy (may be stochastic) for evaluating f.

The upper bound is provided by a naive strategy: quasi-uniform sequence $(\mathbf{x}_n)_n$ fixed in advance and x_n^* taken as the minimizer of an interpolant of the data $f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)$.

Theoretical results: convergence rate (2)

Theorem

For F GP prior on f with ψ as length-scales. For any $R > 0$,

$$\sup_{\|f\|_{\mathcal{H}_{K_\psi}} \leq R} \mathbb{E}^{EI(F)} \left(M_n - \min_{\mathbf{x} \in E} f(\mathbf{x}) \right) = \Theta(n^{-\nu/d} \log(n)^\alpha) \text{ if } \nu \leq 1 ,$$

$$= \Theta(n^{-1/d}) \text{ if } \nu > 1 ,$$

where $M_n = \min(f(\mathbf{x}_1, \ldots, \mathbf{x}_n))$.

Adaptation of the algorithm with a mixed strategy greedy and EI:

- with proba $1 - \epsilon$ choose the next point with EI strategy,
- with proba ϵ choose uniformly at random from E.

Theoretical results: estimated parameters (3)

If the parameters are re-estimated at each step, counter-example to convergence of the EI algorithm.
[Bull, 2011] proposed other estimated than MLE to ensure convergence of EI with re-estimation.

Outline

Principle for estimating the probability of a rare event

- Uncertainties on $\mathbf{x} \Rightarrow$ random variable $\mathbf{X}$,
- Goal: estimate the probability $\alpha = \mathbb{P}(f(\mathbf{X}) > c)$ under the constraint of a limited number N of calls to f.
- For a design D_n, estimation of α:

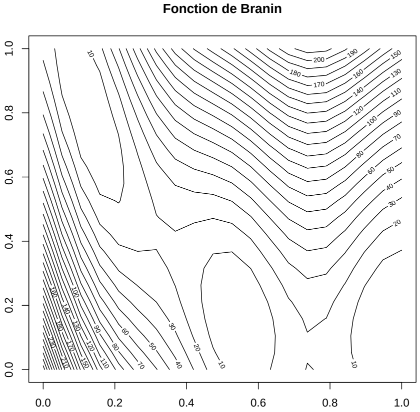
$$\hat{\alpha}_n = \mathbb{E}\left[\mathbb{P}(F(\mathbf{X}) > c)|F(D_n)\right] = \int \Phi\left(\frac{m_{D_n}(\mathbf{x}) - c}{\sqrt{C_{D_n}(\mathbf{x}, \mathbf{x})}}\right) d\mathbb{P}_{\mathbf{X}}(x)$$

- Sequential design criterion:

$$\mathbf{x}_i \in \arg\max_{\mathbf{x}} \mathbb{E}\left[\mathbb{E}((\hat{\alpha}_i - \alpha)^2 | F(D_{i-1}) \cap \mathbf{x}_i = \mathbf{x}) | F(D_{i-1})\right] ,$$

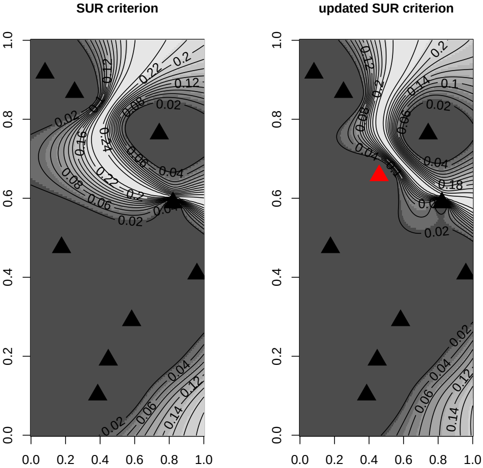
proposed in [Chevalier et al., 2014].

Example

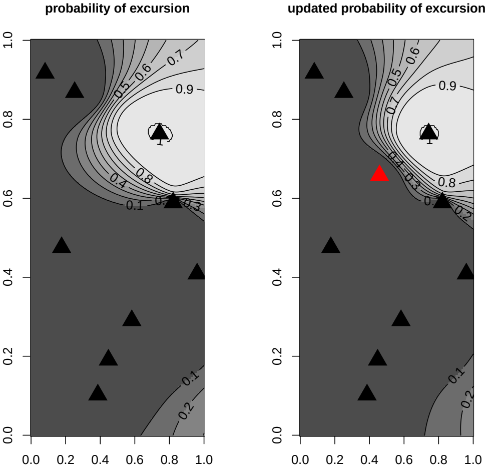


- $\mathbf{X} \sim \mathcal{U}([0,1]^2)$,
- estimate $\mathbb{P}(f(\mathbf{X}) > 80)$.

SUR criterion



point-wise probability of exceeding the threshold



Theoretical results and implementation

- sub-optimal procedure of dynamic programming (one-look ahead),
- interesting when closed-form computation of the SUR criterion, otherwise can require GP trajectories simulations...

Outline

GP-UCB

Gaussian Process - Upper Confidence Bound [Srinivas et al., 2009]:

- Add point $\mathbf{x}_i$ s. t.

$$\mathbf{x}_i \in \arg\max_{\mathbf{x}} m_{D_{i-1}}(\mathbf{x}) + \beta_i^{1/2} \cdot \sqrt{C_{D_{i-1}}(\mathbf{x}, \mathbf{x})}$$

- Bounds on cumulative regret

$$R_T = \sum_{i=1}^{T} f(\mathbf{x}^*) - f(\mathbf{x}_i)$$

  for a well chosen sequence $(\beta_i)_i$ and depending on the covariance kernel of the GP.

  With high probability:

$$R_T = O(\sqrt{dT\gamma_T}),$$

  where d is the dimension of E and γ_T is the maximal information gain in T rounds.

- in the practical cases studied in the paper, similar performance of EI and $GP - UCB$.

GP-MI

Gaussian Process - Mutual Information [Contal et al., 2014]:

- Add point $\mathbf{x}_i$ s. t.

$$\mathbf{x}_i \in \arg\max_{\mathbf{x}} m_{D_{i-1}}(\mathbf{x}) + \phi_i(\mathbf{x}) \,,$$

  where

$$\phi_i(\mathbf{x}) = \sqrt{\alpha}\left(\sqrt{\sigma_i^2(\mathbf{x}) + \hat{\gamma}_{i-1}} - \sqrt{\hat{\gamma}_{i-1}}\right) \,.$$

- $\hat{\gamma}_i$ forms a lower bound on the information acquired on f on the query points D_i. Updating formula:

$$\hat{\gamma}_i = \hat{\gamma}_{i-1} + C_{D_i}(\mathbf{x}_i, \mathbf{x}_i) \,,$$

  with $\hat{\gamma}_0 = 0$.

- Bound on the cumulative regret

$$R_T = O((\log T)^{(d+1)/2}) \,,$$

  for a RBF kernel.

References

Bull, A. D. (2011).
Convergence rates of efficient global optimization algorithms.
12:2879–2904.

Chevalier, C., Bect, J., Ginsbourger, D., Vazquez, E., Picheny, V., and Richet, Y. (2014).
Fast parallel kriging-based stepwise uncertainty reduction with application to the identification of an excursion set.
Technometrics, 56(4):455–465.

Contal, E., Buffoni, D., Robicquet, A., and Vayatis, N. (2013).
Parallel gaussian process optimization with upper confidence bound and pure exploration.
In *Machine Learning and Knowledge Discovery in Databases*, pages 225–240. Springer.

Contal, E., Perchet, V., and Vayatis, N. (2014).
Gaussian process optimization with mutual information.
eprint arXiv:1311.4825.

Grûnewâlder, S., Audibert, J.-Y., Opper, M., and Shawe-Taylor, J. (2010).
Regret bounds for gaussian process bandit problems.
In *International Conference on Artificial Intelligence and Statistics*, pages 273–280.

Jala, M., Lévy-Leduc, C., Moulines, E., Garivier, A., Conil, E., and Wiart, J. (2014).
Sequential design of computer experiments for the estimation of a quantile with application to numerical dosimetry.
Technometrics.

Johnson, M. E., Moore, L. M., and Ylvisaker, D. (1990).
Minimax and maximin distance designs.
Journal of Statistical Planning and Inference, 26(2):131 – 148.

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.

Morris, M. D. and Mitchell, T. J. (1995).
Exploratory designs for computer experiments.
Journal of Statistical Planning and Inference, 43:381–402.

Picheny, V. (2014).
Multiobjective optimization using gaussian process emulators via stepwise uncertainty reduction.
Statistics and Computing, pages 1–16.

Sacks, J., Schiller, S. B., Mitchell, T. J. et Wynn, H. P. (1989).
Design and analysis of computer experiments (with discussion).
Statistical Science, 4 409–435.

Santner, T. J., Williams, B., and Notz, W. (2003).
The Design and Analysis of Computer Experiments.
Springer-Verlag.

Schaback, R. (1995).
Error estimates and condition numbers for radial basis function interpolation.
Advances in Computational Mathematics, (3):251–264.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. (2009).
Gaussian process optimization in the bandit setting: No regret and experimental design.
arXiv preprint arXiv:0912.3995.

Vazquez, E. and Bect, J. (2010).
Convergence properties of the expected improvement algorithm with fixed mean and covariance functions.
Journal of Statistical Planning and Inference, 140(11):3088–3095.