Bayesian optimization with derivatives acceleration

Guillaume Perrin[1], Rodolphe Le Riche[2]

[1]COSYS, Univ. Gustave Eiffel
[2](speaker) CNRS LIMOS, Mines Saint-Étienne, Fr

Bayesian Optimization workshop, IHP, June 20th 2024

CIROQUO consortium

Abstract

For details and citation, please refer to [Perrin and Le Riche, 2023]

Bayesian optimization algorithms form an important class of methods to minimize functions that are costly to evaluate, which is a very common situation. These algorithms iteratively infer Gaussian processes from past observations of the function and decide where new observations should be made through the maximization of an acquisition criterion. Often, in particular in engineering practice, the objective function is defined on a compact set such as in a hyper-rectangle of a d-dimensional real space, and the bounds are chosen wide enough so that the optimum is inside the search domain. In this situation, this work provides a way to integrate in the acquisition criterion the a priori information that these functions, once modeled as GP trajectories, should be evaluated at their minima, and not at any point as usual acquisition criteria do. We propose an adaptation of the widely used Expected Improvement acquisition criterion that accounts only for GP trajectories where the first order partial derivatives are zero and the Hessian matrix is positive definite. The new acquisition criterion keeps an analytical, computationally efficient, expression. This new acquisition criterion is found to improve Bayesian optimization on a test bed of functions made of Gaussian process trajectories in dimensions 2, 3 and 5. The addition of first and second order derivative information is particularly useful for multimodal functions.

Bayesian Optimization (BO)

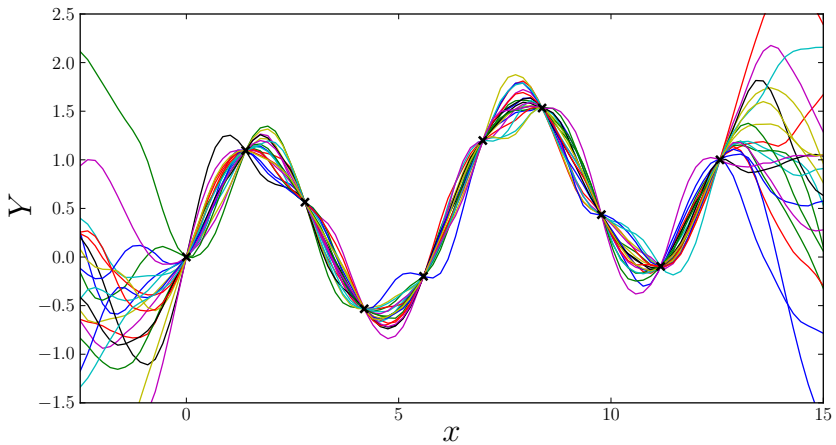
Our goal : find $x^\star \in \arg\min_{x \in \mathbb{X} \subset \mathbb{R}^d} y(x)$

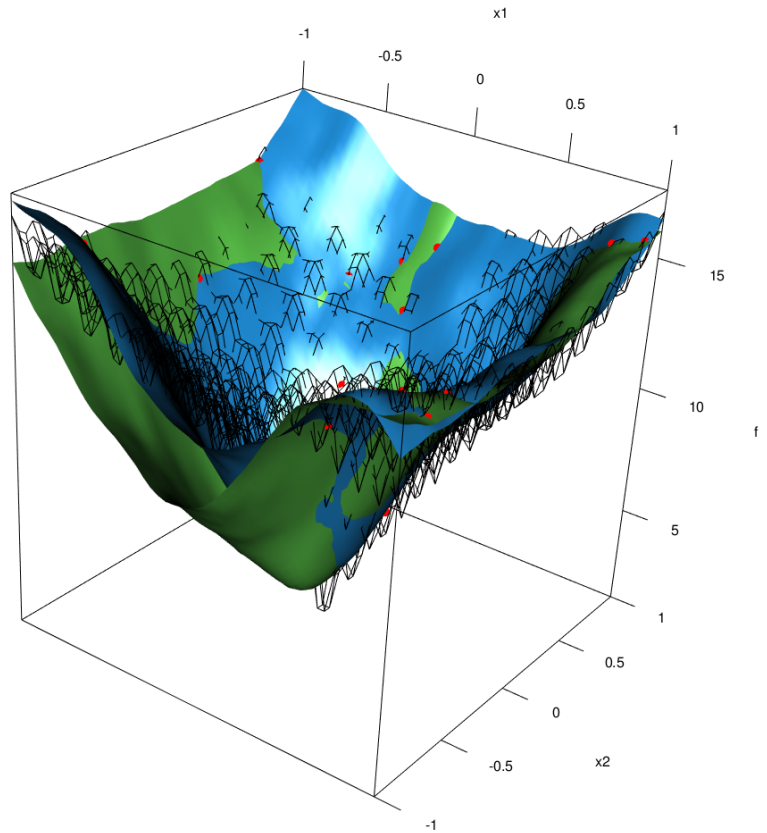
1. Learn a model (most often, a Gaussian Process $Y_N()$) of $y()$ from N point observations
2. Deduce from $Y_N()$ a candidate point to evaluate by maximizing an acquisition criterion
   $x^{N+1} = \arg\max_{x \in \mathbb{X}} \mathfrak{a}(x; Y_N)$
3. Calculate $y(x^{N+1})$, add it to the observations, $N \leftarrow N + 1$
4. stop or go to 1

BO general references: [Garnett, 2023, Gramacy, 2020, Frazier, 2018, Shahriari et al., 2015, Sobester et al., 2008, Jones, 2001]

Basic assumption of BO

Trajectories of the Gaussian process (GP) are possible functions underlying the observations:
$y^{(\omega)}(x) \sim \mathcal{N}\left(\mathbb{E}(Y_N(x)) \ , \ \mathrm{Cov}(Y_N(x), Y_N(x'))\right)$ are possible $y(x)$



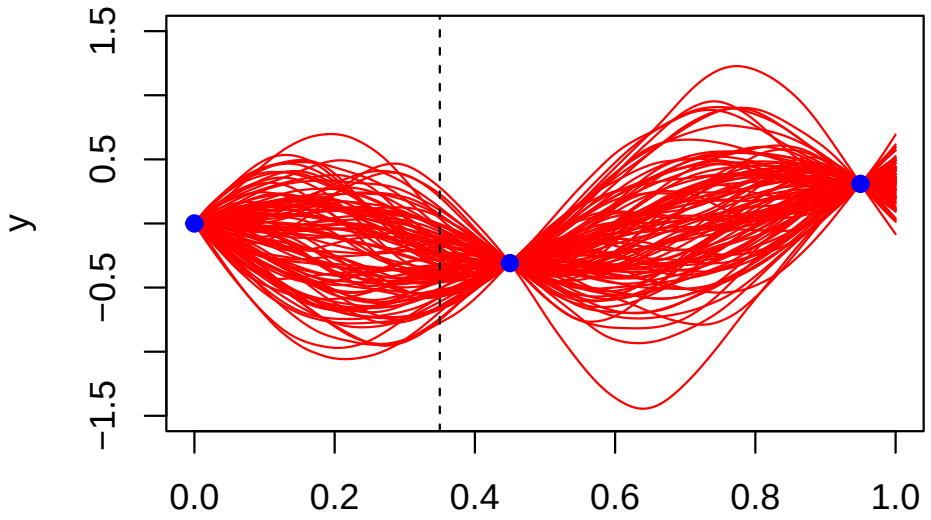


Acquisition criteria: pointwise

Look at the distribution of $Y_N(x)$ at x to give it a worth sampling value. Often Analytical.

(Notation: $\mathbb{E}(Y_N(x)) = \mu(x)$, $\mathrm{Cov}(Y_N(x), Y_N(x')) = c(x, x')$)

- Upper Confidence Bound : $\mathfrak{a}(x) = \mu(x) + \alpha\sqrt{c(x,x)}$
- Probability of improvement : $\mathfrak{a}(x) = \mathbb{P}(Y_N(x) < y_{\min})$
- Expected improvement : $\mathfrak{a}(x) = \mathrm{EI}(x) = \mathbb{E}\left[\max(0, y_{\min} - Y_N(x))\right]$
- and many others ...



Acquisition criteria: global in scope

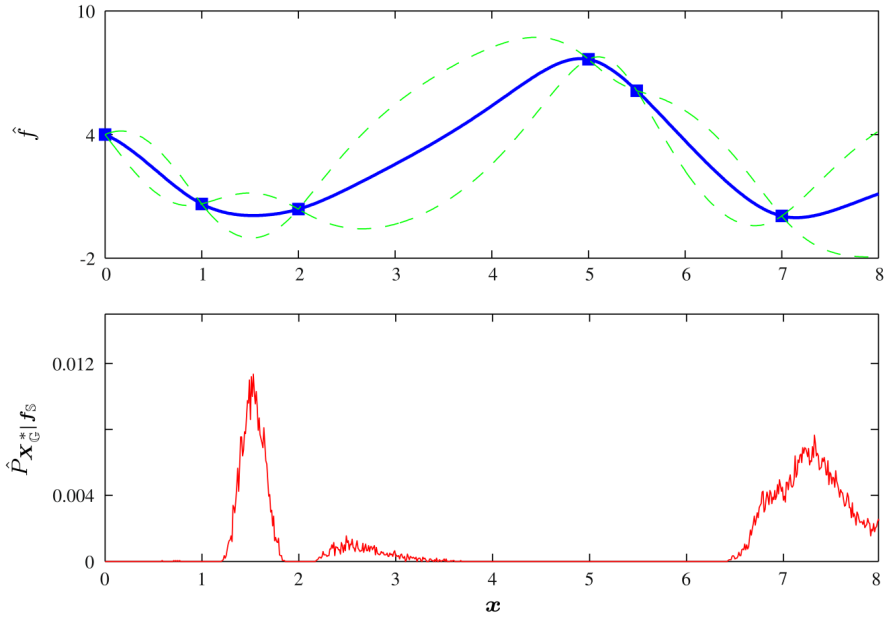
Look at the effect of adding an observation at x elsewhere in $\mathbb{X}$.
The criteria are no longer analytical.

- Pointwise criteria averaged over $\mathbb{X}$ (not analytical). Example: IECI (Integrated Expected Conditional Improvement) [Gramacy and Lee, 2011].
- Knowledge gradient [Frazier, 2018] : not analytical, not so local
  $\mathfrak{a}(x) = \mathbb{E}\left[\min_{x'} \mu(x') - \min_{x'} \mu^{(+x, Y_N(x))}(x')\right]$

Acquisition criteria: information-wise

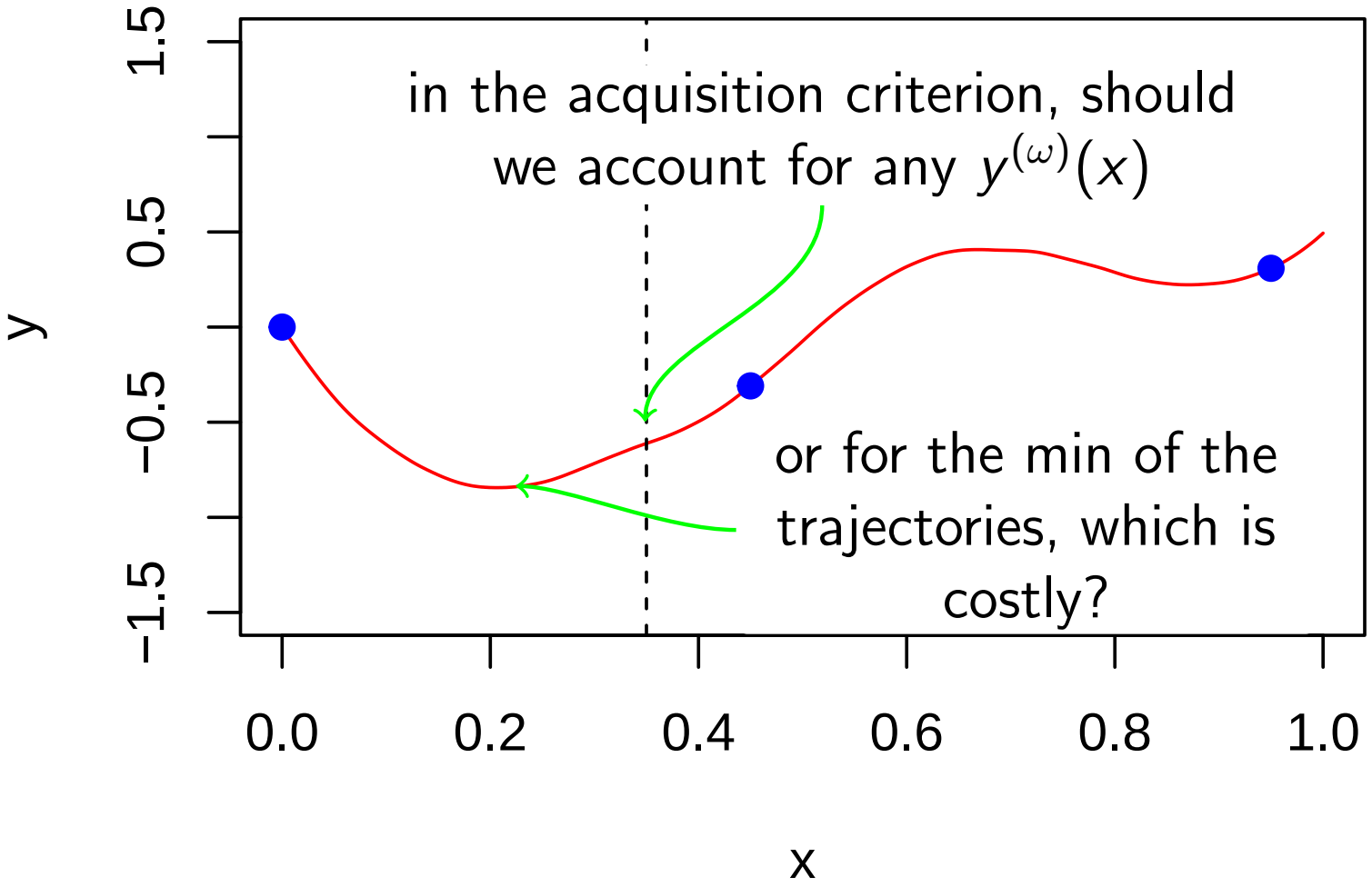
Measure how sampling at x globally provides information about $X^\star = \arg\min_x Y_N(x)$. Not analytical.

- Sample where the entropy of $X^\star$ is reduced the most [Villemonteix et al., 2009, Hennig and Schuler, 2012]
- as above, obtained through the entropy of $Y_N(x) \mid X^\star$ [Hernández-Lobato et al., 2014]



(from [Villemonteix et al., 2009])

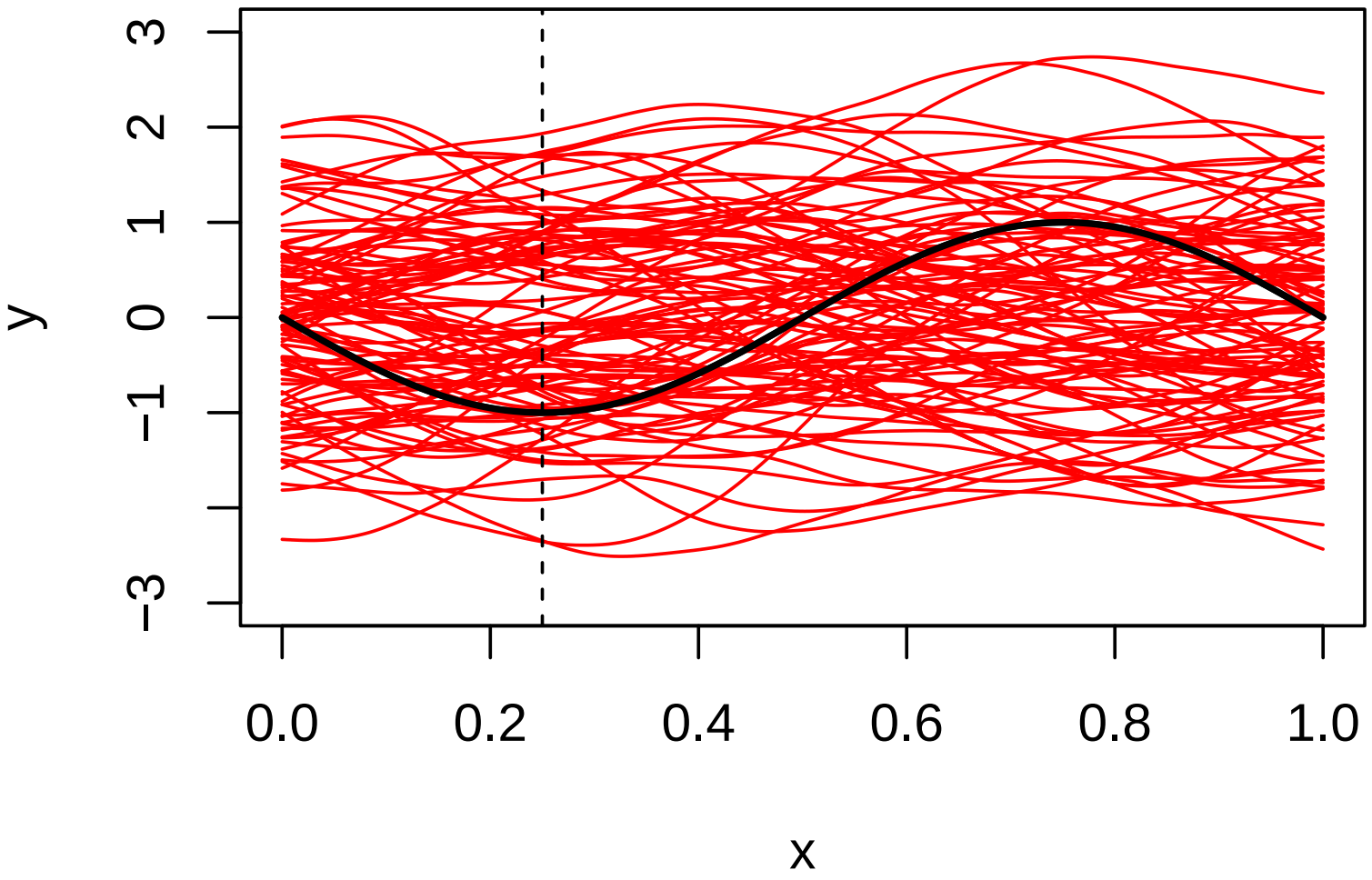
Trajectories and relevant information for optimization



Derivatives acceleration: outline of the talk

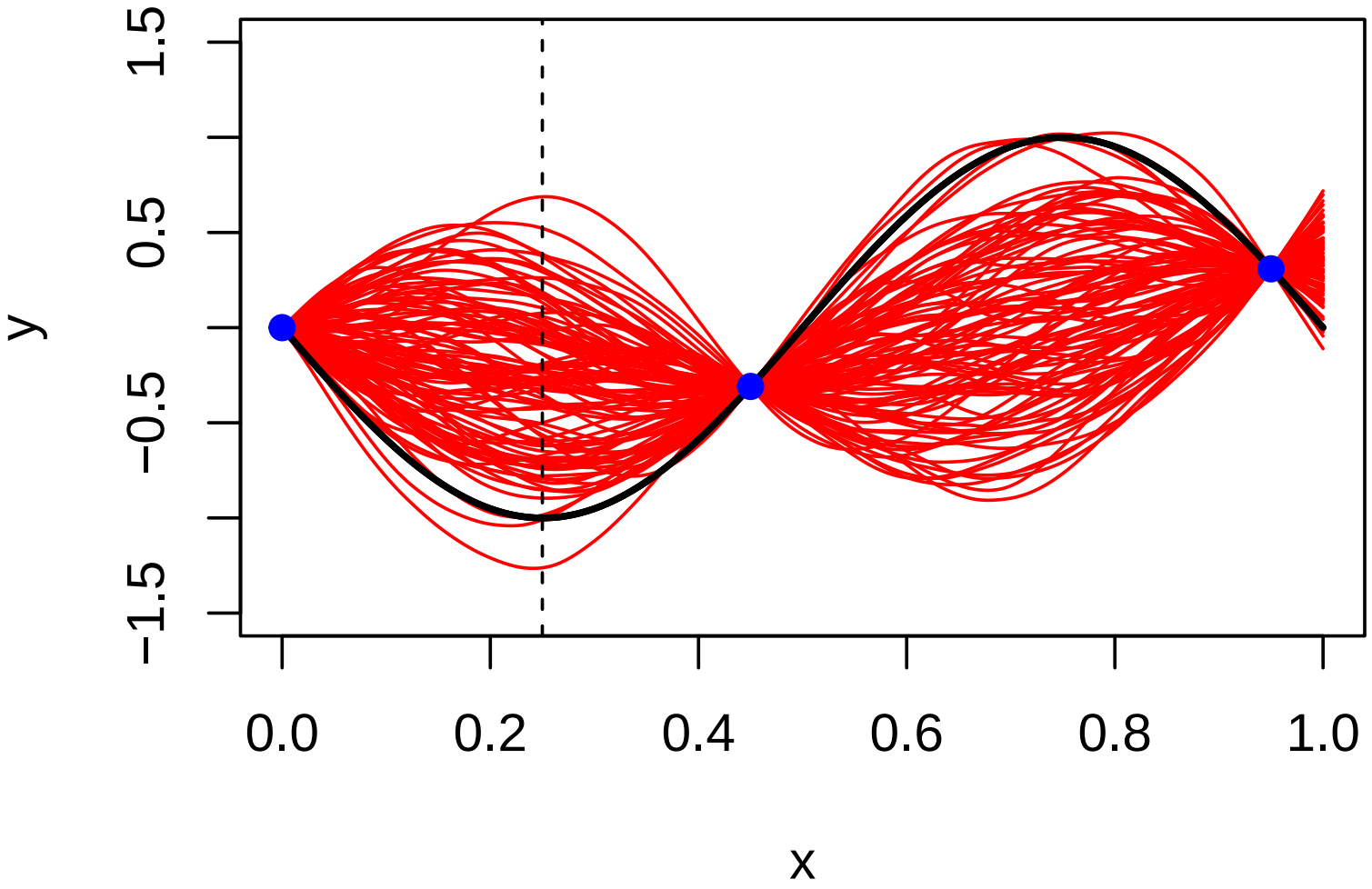
- $y^{(\omega)}(x) \sim \mathcal{N}\left(\mathbb{E}(Y_N(x)) \ , \ \mathrm{Cov}(Y_N(x), Y_N(x'))\right)$ are possible $y(x)$
- When evaluating the worth of x in the acquisition criterion (potential x^{N+1}), only account for trajectories that have a local minimum at $x \Rightarrow$ no need to optimize them!
- This is possible with GPs.
- Better, propose an analytical approximation to the acquisition criterion, deriv-EI.

GPs plasticity: no conditioning



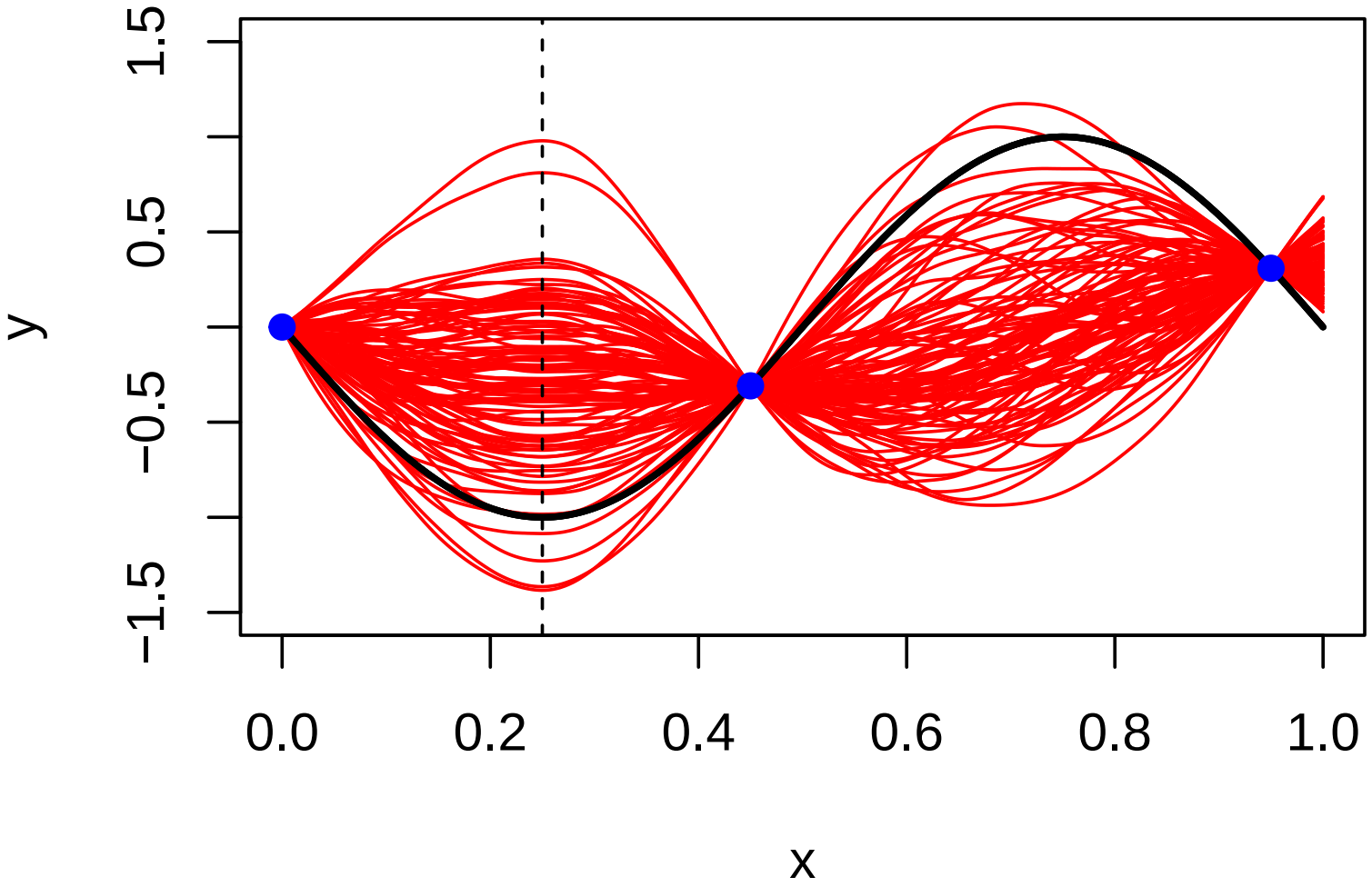
Note: with a stationary GP, $Y_N(x)$, $\partial Y_N(x)$ and $\partial^2 Y_N(x)$ are independent.

GPs plasticity: with y observations



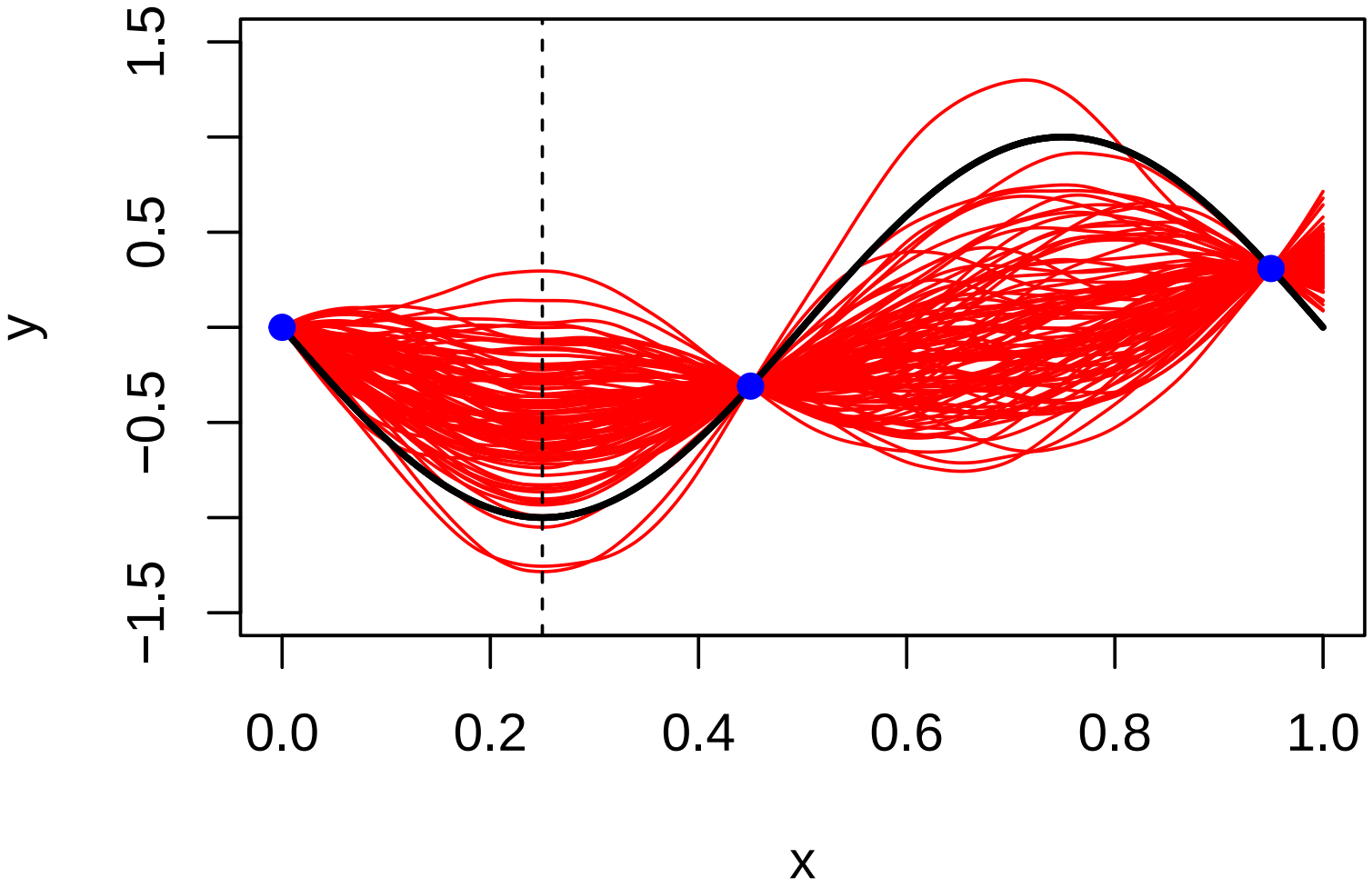
The usual Gaussian Process Regression (kriging) picture

GPs plasticity: y observations + null derivatives



All trajectories interpolate the observations and have $\partial y^{(\omega)}(x) = 0$ at dotted x. Some are maxima, others minima.

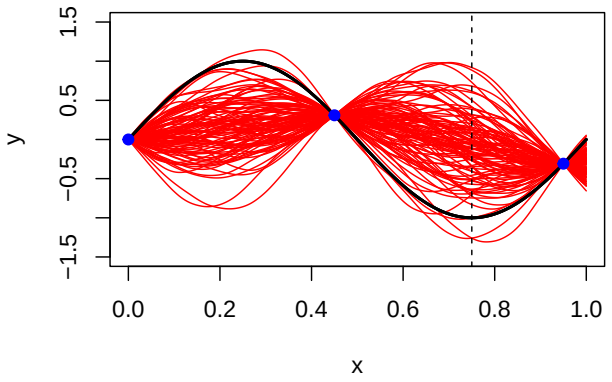
GPs plasticity: y observations + null derivatives + positive curvatures



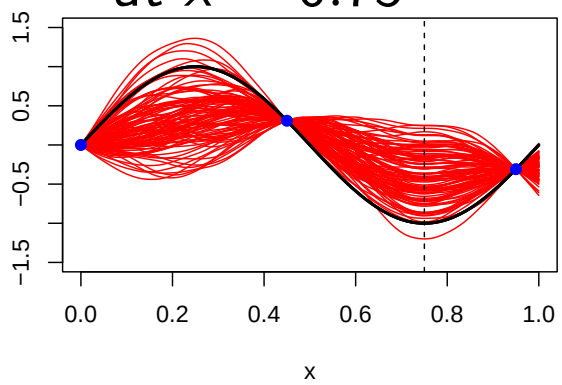
All trajectories interpolate the observations, have $\partial y^{(\omega)}(x) = 0$ and $\partial^2 y^{(\omega)}(x) > 0$ at dotted x. They are local minima.

Expl: enforcing trajectories with local minima

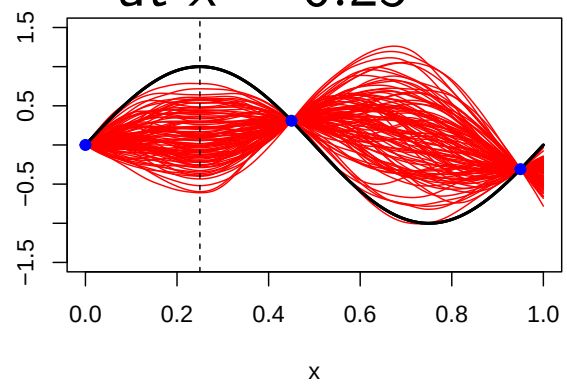
classical GPR



GPR with loc. min. at $x = 0.75$



GPR with loc. min. at $x = 0.25$



When enforcing local minima in the optimal region, the true function (black line) is better represented, and vice versa $\Rightarrow$ an acquisition criterion with derivatives acceleration should learn faster.

GP and optimality conditions

Recall our goal : find $x^\star \in \arg\min_{x \in \mathbb{X} \subset \mathbb{R}^d} y(x)$

2nd order optimality conditions on the GP

If x is an interior point and $y^{(\omega)}()$ C^2, x is a local minimum of $y^{(\omega)}()$ if gradient is null and Hessian positive definite at x:

$$\partial y^{(\omega)}(x) = 0 \qquad \text{and} \qquad \partial^2 y^{(\omega)}(x) > 0$$

No derivative of the true function needed

Do not mistake the true function, $y(x)$, with the GP trajectories, $y^{(\omega)}(x)$. The GP trajectories need to be C^2, not the true function.

GP with derivatives (1/2)

Derivation is a linear operator: $(\partial Y_N(x))_{i=1,\dots,d}$ and $(\partial^2 Y_N(x))_{i,j\in\{1,\dots,d\}^2}$ are GPs with known means and covariances.

Examples: $Y_N(x) \sim \mathcal{N}(\mu(x), c(x,x'))$

- $\frac{\partial Y_N(x)}{\partial x} \sim \mathcal{N}(\frac{\partial \mu(x)}{\partial x}, \frac{\partial^2 c(x,x')}{\partial x \partial x'})$
- $\frac{\partial^2 Y_N(x)}{\partial x^2} \sim \mathcal{N}(\frac{\partial^2 \mu(x)}{\partial x^2}, \frac{\partial^4 c(x,x')}{\partial x^2 \partial x'^2})$
- $\text{Cov}\left(Y_N(x), \frac{\partial Y_N(x')}{\partial x'}\right) = \frac{\partial c(x,x')}{\partial x'}$
- $\text{Cov}\left(\frac{\partial Y_N(x)}{\partial x}, \frac{\partial^2 Y_N(x')}{\partial x'^2}\right) = \frac{\partial^3 c(x,x')}{\partial x \partial x'^2}$
- ...

GP with derivatives (2/2)

In general, for any linear operator $\mathcal{L}$, $\mathcal{L}Y$ is also a Gaussian process, with

$$\mathbb{E}[\mathcal{L}Y(x)] = \mathcal{L}\mu(x), \quad \text{Cov}(\mathcal{L}Y(x), \mathcal{L}Y(x')) = \mathcal{L}C(x, x')\mathcal{L}^T.$$

Here, of interest is the operator creating the $\mathbb{R}^{1+d(d+3)/2}$ vector of Y and its first and second derivatives,

$$\mathcal{L} : Y \mapsto \mathcal{L}Y := \left\{ Y, \frac{\partial Y}{\partial x_1}, \ldots, \frac{\partial Y}{\partial x_d}, \frac{\partial^2 Y}{\partial x_1^2}, \ldots, \frac{\partial^2 Y}{\partial x_1 \partial x_2}, \ldots, \frac{\partial^2 Y}{\partial x_d^2} \right\}$$

GP conditioned for optimality conditions (1/2)

Key Gaussian vector :

$$\left(Y(x), \partial^2 Y(x), \overbrace{\partial Y(x), Y(x^1), \ldots, Y(x^N)}^{\text{to be conditioned}} \right)$$

Conditioning by the observations :
$Y_N(x) = Y(x) \mid Y(x^1) = y(x^1) \ldots Y(x^N) = y(x^N)$

$$(Y_N(x), D^2 Y_N(x) | \partial Y_N(x) = 0) \sim \mathcal{N}\left(\begin{pmatrix} m \\ \ddot{m}_1 \\ \vdots \\ \ddot{m}_d \end{pmatrix}, \begin{bmatrix} s^2 & \rho_{1,1} & \cdots & \rho_{1,d} \\ \rho_{1,1} & \ddot{s}_1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho_{d-1,d} \\ \rho_{d,1} & \cdots & \rho_{d,d-1} & \ddot{s}_d \end{bmatrix} \right)$$

GP conditioned for optimality conditions (2/2)

For example,

$$m = \mu(x) + \partial c(x,x)^\top \left[\partial^2 c(x,x)\right]^{-1} (0 - \partial\mu(x))$$
$$s^2 = c(x,x) - \partial c(x,x)^\top \left[\partial^2 c(x,x)\right]^{-1} \partial c(x,x)$$

Enforcing the positive definiteness of $\partial^2 Y_N(x)$ is more complicated.

Simulation with gradient and Hessian constraints

Approximate $\partial^2 Y_N(x) > 0$ by $D^2 Y_N(x) > 0$
$(D^2 Y_N(x) = \text{diag}(\partial^2 Y_N(x)))$

Use algorithm of [Perrin and Da Veiga, 2021] to simulate the GP $Y_N(x), D^2 Y_N(x) \mid \partial Y_N(x) = 0$ under the constraints $D^2 Y_N(x) > 0$

EI with derivatives acceleration

Expected Improvement [Saltenis, 1971, Schonlau, 1997] :

$$\begin{aligned}\mathfrak{a}(x) = \mathsf{EI}(x) &:= \mathbb{E}\left[\max(0, y_{\min} - Y_N(x))\right] \\ &= \sqrt{c(x,x)}\left[U(x)\Phi(U(x)) + \phi(U(x))\right] \\ &\text{where } U(x) := (y_{\min} - \mu(x))/\sqrt{c(x,x)}\end{aligned}$$

EI with derivatives acceleration

$$\text{deriv-EI}(x) := \mathbb{E}\left[\mathbb{1}_{R(x)\partial Y_N(x) \in \mathcal{B}(\varepsilon), \partial^2 Y_N(x) > 0} \max(0, y_{\min} - Y_N(x))\right]$$

where $\mathcal{B}(\varepsilon) := \left\{\dot{y} \in \mathbb{R}^d,\ \|\dot{y}\| \leq \varepsilon\right\}$ is the d-dimensional hypersphere of radius ε, $R(x)$ is a matrix such that $R(x)\text{Cov}(\partial Y_N(x))R(x)^T = I_d$.
All the explanations in [Perrin and Le Riche, 2023].

Approximation to deriv-EI I

Acquisition criteria will be optimized a lot : they should be inexpensive to calculate if possible $\Rightarrow$ we propose an analytical approximation to deriv-EI.

Assumptions:

- A1 ε, the size of the sphere to which $\|R(x)\partial Y_N(x)\|$ belongs, is small.
- A2 The positive definiteness of $\partial^2 Y_N(x) \mid \partial Y_N(x) = 0, Y_N(x) = y$ can be approximately checked as all main curvatures are independent and positive, $((\partial^2 Y_N(x))_{i,i} \mid \partial Y_N(x) = 0, Y_N(x) = y) > 0$.

Approximation to deriv-EI II

Under A1 & A2,

$$\textbf{deriv-EI}(x) \approx \text{LikelyMin}(x) \times \text{cond-EI}(x)$$

$$\text{LikelyMin}(x) := \text{cte}(\varepsilon^d) \times \exp\left(-\frac{\dot{m}^T \dot{S}^{-1} \dot{m}}{2}\right) \times \prod_{i=1}^{d} \Phi\left(\frac{\ddot{\tau}_i}{\sqrt{1 - r_i{}^2}}\right)$$

$$\text{cond-EI}(x) := s\left[(z_{\min} - a)\Phi(z_{\min}) + \phi(z_{\min})\right]$$

where $\partial Y_N(x) \sim \mathcal{N}(\dot{m}, \dot{S})$, $\ddot{\tau}_i = \frac{\ddot{m}_i}{\ddot{s}_i}$, $r_i = \frac{\rho_{1i}}{s\ddot{s}_i}$, $z_{\min} = \frac{y_{\min} - m}{s}$,
$a =$ cf. [Perrin and Le Riche, 2023], $\to 0$ when $r_i \to 0$ or $\ddot{\tau}_i \to 0$.

Approximation to deriv-EI III

- The (analytical) deriv-EI is marginally more expensive than EI: just model $Y(x), D^2Y(x) \mid Y(\mathbf{X}) = y(\mathbf{X}), \partial Y(x) = 0$ instead of $Y(x) \mid Y(\mathbf{X}) = y(\mathbf{X})$, i.e., only d extra-observations in the covariance matrix to invert.

Start of the proof:

$$\text{deriv-EI}(x) = \int_{y=-\infty}^{y_{\min}} \int_{\dot{y}\in\mathcal{E}(x,\varepsilon)} \int_{\ddot{Y}\,\text{def.pos.}} (y_{\min} - y)^p d\mathbb{P}(y, \dot{y}, \ddot{Y})$$

$$\approx \text{Vol}(\mathcal{E}(x,\varepsilon)) f_{\partial Y_N(x)}(0) \int_{y=-\infty}^{y_{\min}} \int_{\ddot{y}\in[0,+\infty[^d} (y_{\min} - y)^p f_{Y_N(x), D^2Y_N(x) \mid \partial Y_N(x)=0}(y, \ddot{y}) dy d\ddot{y}$$

$$= \text{Vol}(\mathcal{E}(x,\varepsilon)) f_{\partial Y_N(x)}(0) \int_{y=-\infty}^{y_{\min}} \int_{\ddot{y}\in[0,+\infty[^d} (y_{\min} - y)^p f_{Y_N(x) \mid \partial Y_N(x)=0}(y)$$

$$f_{D^2Y_N(x) \mid \partial Y_N(x)=0, Y_N(x)=y}(\ddot{y}) dy d\ddot{y}$$

$= \ldots$ independence of the $(D^2Y_N(x))_i \mid \partial Y_N(x) = 0, Y_N(x) = y$ makes the product appear $\ldots$

Numerical experiments

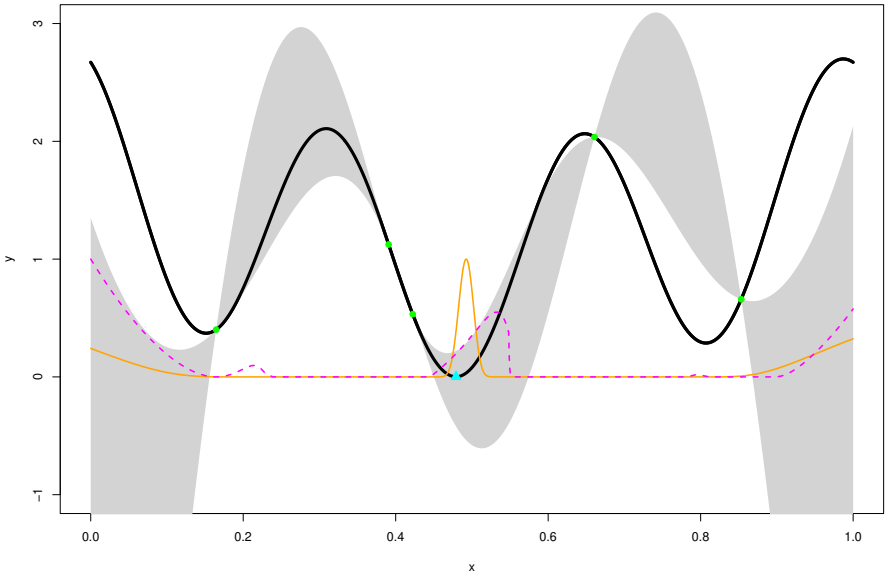
- GPs: constant mean, Matérn 5/2 kernels : trajectories are twice continuously differentiable, as needed,

$$x \in \mathbb{X} = [0,1]^d \ , \ \mathrm{Cov}(Y(x), Y(x')) = \sigma^2 \prod_{i=1}^{d} \kappa\left(\frac{|x_i - x_i'|}{\theta\sqrt{d/2}}\right) \ ,$$

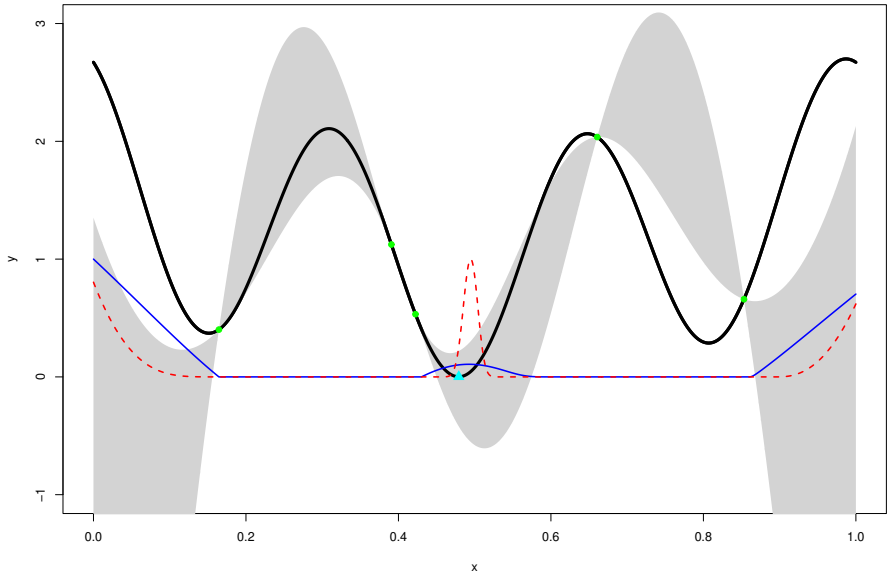
$$\kappa(u) := \left(1 + \sqrt{5}u + \frac{5}{3}u^2\right)\exp\left(-\sqrt{5}u\right)$$

- Test functions: y^{1D} and y^{2D} + GPs with this (known) Matérn covariance and interior optima: results only depend on the acquisition criterion, not artifacts.
- deriv-EI and EI are either optimized by exhaustive search (in 1 and 2D) or 10^5 random evaluations followed by 10 Nelder-Mead search starting from the best points.

Illustration of deriv-EI on y^{1D}



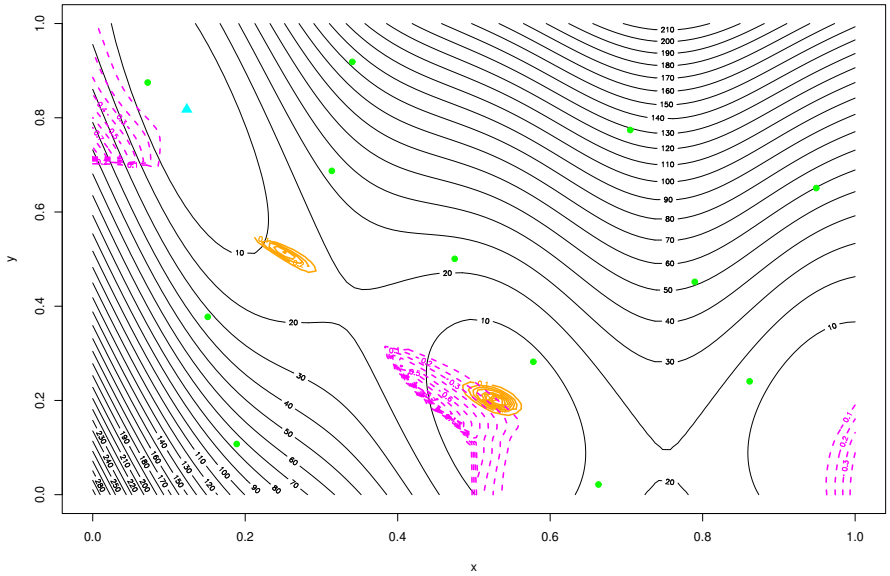
LikelyMin (——) and cond-EI (- - -)



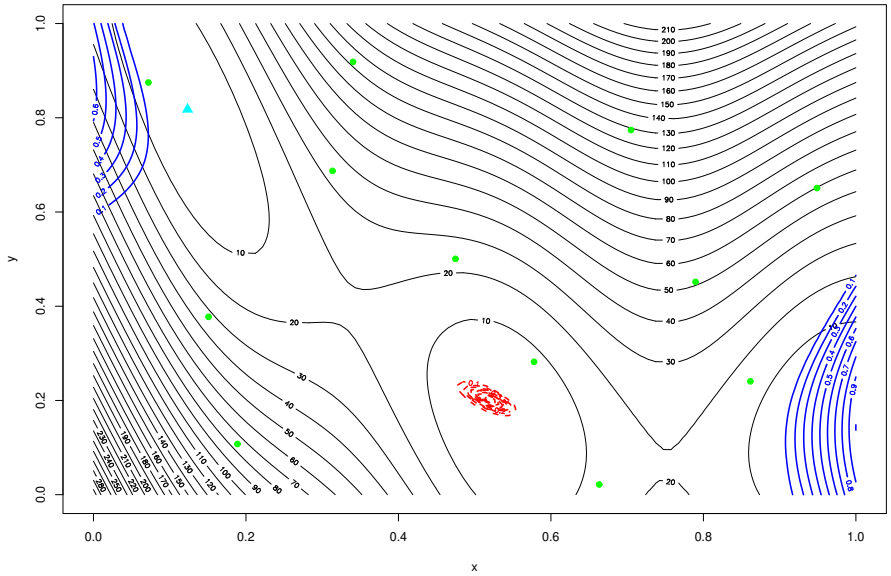
EI (——) and deriv-EI (- - -)

- deriv-EI has more concentrated high-values regions than EI
- deriv-EI favors the center of $\mathbb{X}$ while EI favors the edges

Illustration of deriv-EI on y^{2D}



LikelyMin (——) and cond-EI (- - -)

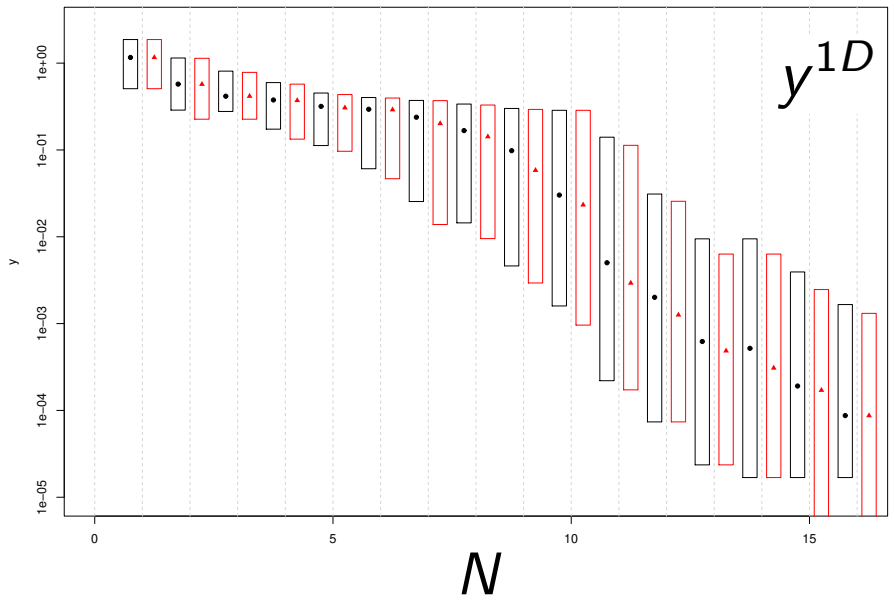


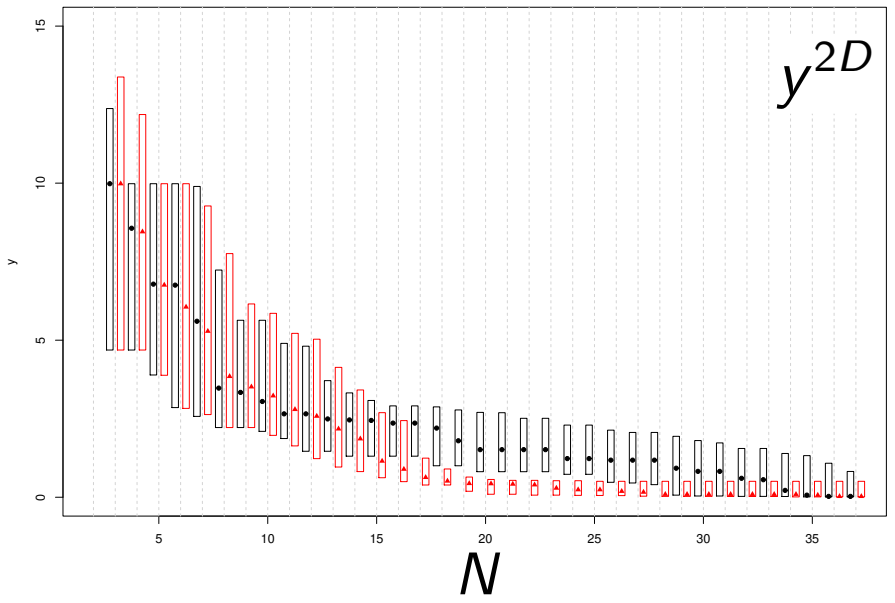
EI (——) and deriv-EI (- - -)

- deriv-EI has more concentrated high-values regions than EI
- deriv-EI favors the center of $\mathbb{X}$ while EI favors the edges

deriv-EI vs. EI, one step

500 LHS of size N $\in [2d+1, 20d]$, quartiles of $y(x^{N+1})$,
EI in black vs. deriv-EI in red.

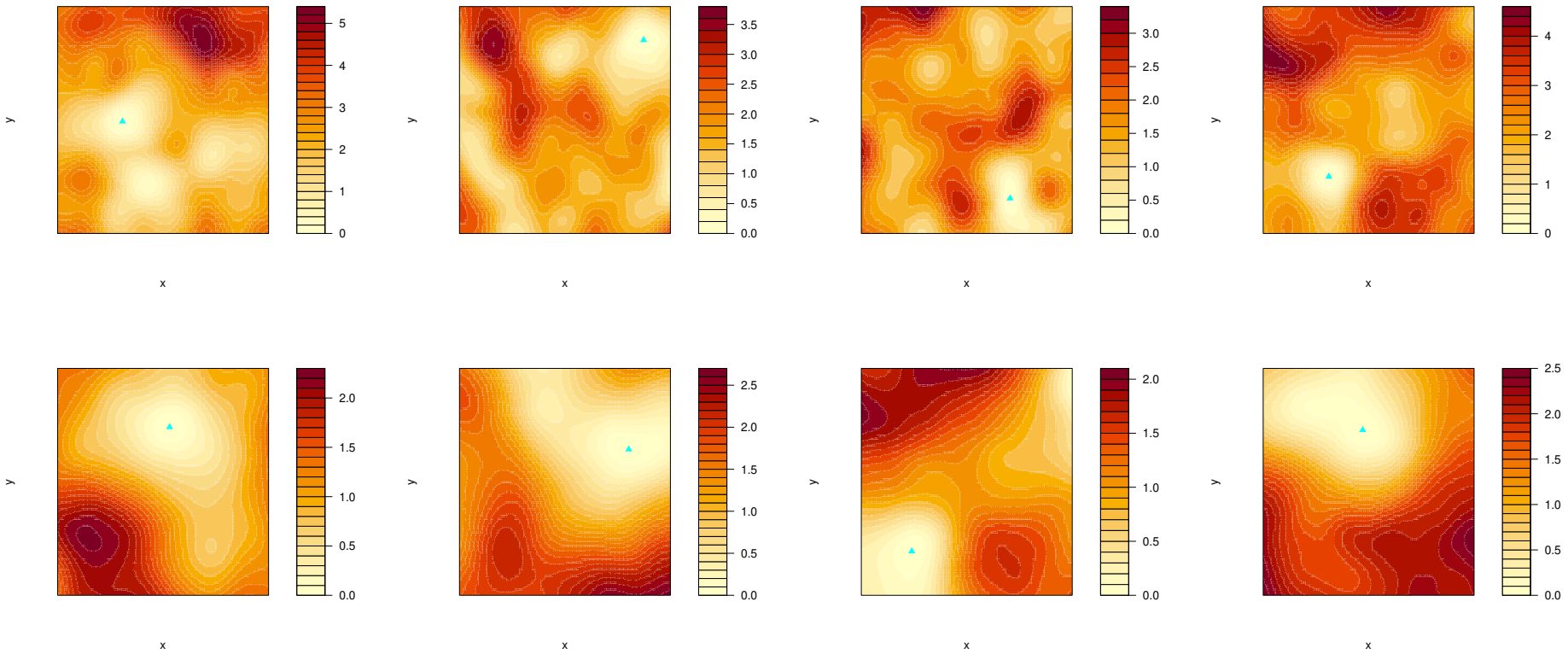




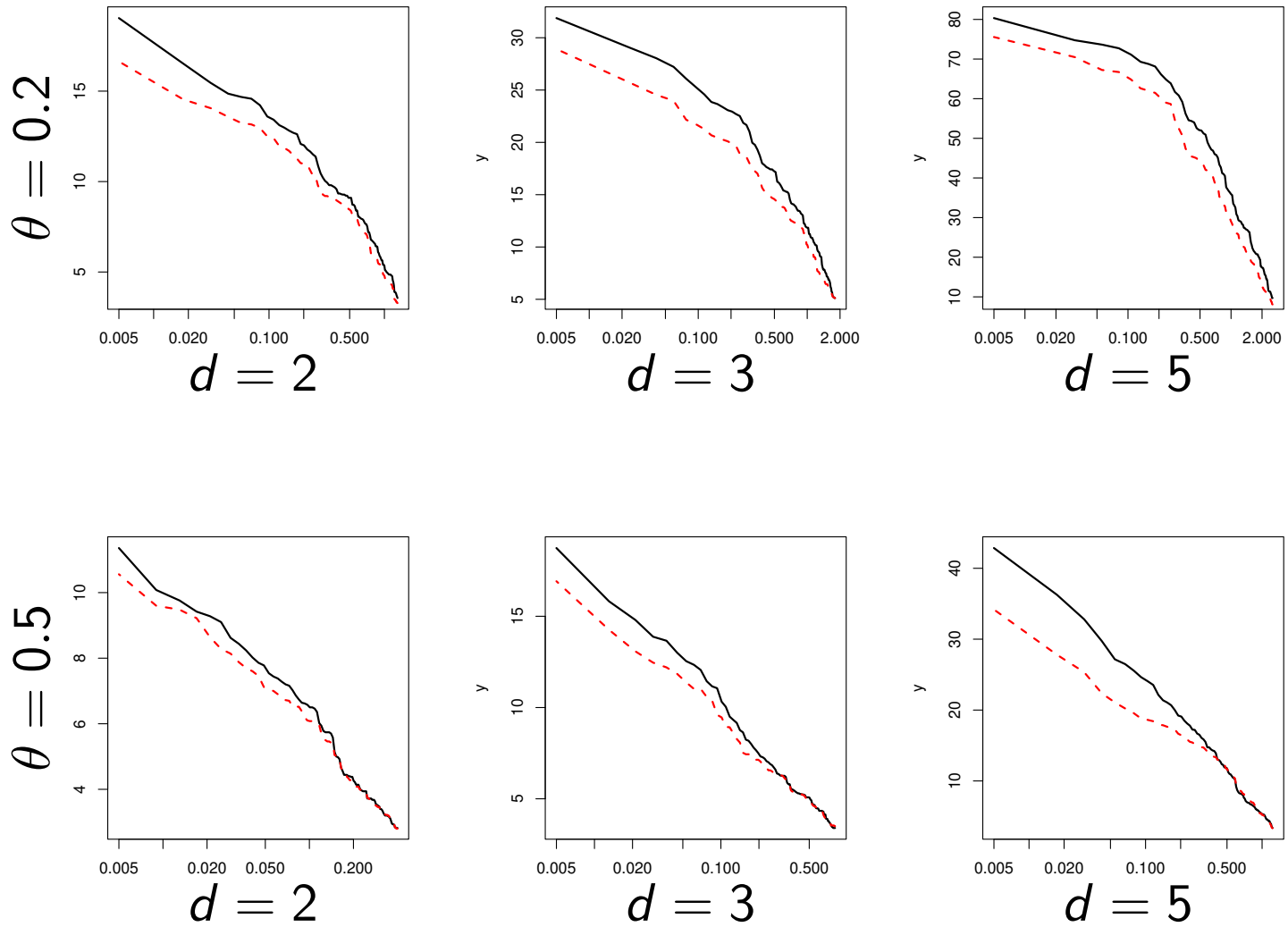
deriv-EI yields better results, in particular in the middle of the search. Indeed, EI and deriv-EI equivalent at the beginning (Y, ∂Y and $\partial^2 Y$ independent) and at the end (EI samples in regions of local optima).

Tests on GPs

100 functions are generated as GPs in $d = 2, 3$ and 5.
Examples in 2D: $\theta = 0.2$ top row, $\theta = 0.5$ bottom row.



Tests on GPs: mean time to target vs. target



Conclusions

- deriv-EI is an EI calculated only on optima at no extra cost.
- Testing on problems with the proper structure (GPs and interior optima), $d \leq 5$, show that deriv-EI improves over EI.
- deriv-EI samples less than EI on the bounds ? What if the optimum is on the bounds ?
  - $\Leftarrow$ deriv-EI does not prevent from sampling on the bounds, it strikes a compromise between null gradient, positive curvatures and possible Y values. It may correct in high-dimension a known flaw of EI.
  - $\Leftarrow$ Extension: use $Y_N(x)$ trajectories to estimate the probability that $x^\star$ is on the bound. If large, do something else (use EI, set x_i to the bound value, ...).
- Other acquisition criteria could benefit from the derivatives acceleration: EI^2 (done in the paper), probability of improvement, ...

References I

Frazier, P. I. (2018).

A tutorial on Bayesian optimization.
arXiv preprint arXiv:1807.02811.

Garnett, R. (2023).

Bayesian optimization.
Cambridge University Press.

Gramacy, R. B. (2020).

Surrogates: Gaussian process modeling, design, and optimization for the applied sciences.
Chapman and Hall/CRC.

Gramacy, R. B. and Lee, H. K. H. (2011).

Optimization under unknown constraints.
In Bayesian Statistics, volume 9, pages 229–256.

Hennig, P. and Schuler, C. J. (2012).

Entropy search for information-efficient global optimization.
Journal of Machine Learning Research, 13(6).

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.

Jones, D. R. (2001).

A taxonomy of global optimization methods based on response surfaces.
Journal of Global Optimization, 21:345–383.

References II

Perrin, G. and Da Veiga, S. (2021).

Constrained gaussian process regression: an adaptive approach for the estimation of hyperparameters and the verification of constraints with high probability.
Journal of Machine Learning for Modeling and Computing, 2(2).

Perrin, G. and Le Riche, R. (2023).

Bayesian optimization with derivatives acceleration.
preprint.

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.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and De Freitas, N. (2015).

Taking the human out of the loop: A review of Bayesian optimization.
Proceedings of the IEEE, 104(1):148–175.

Sobester, A., Forrester, A., and Keane, A. (2008).

Engineering design via surrogate modelling: a practical guide.
John Wiley & Sons.

Villemonteix, J., Vazquez, E., and Walter, E. (2009).

An informational approach to the global optimization of expensive-to-evaluate functions.
Journal of Global Optimization, 44:509–534.