Noisy EI and random metamodels

David Ginsbourger
IMSV, University of Bern (Switzerland)

Joint work with V. Picheny (Centrale Paris) and Y. Richet (IRSN)
Tuesday May 4th 2010
Atelier Mascot Num, IHP

Outline

1. Generalities on Kriging and Kriging-based Optimization
   - From ”deterministic” to ”noisy” Kriging
   - Deterministic Expected Improvement and EGO
2. Noisy EI and random metamodels
   - Noisy EI: past and future noises
   - Where random metamodels emerge
1. Conclusion and perspectives

General context: notation and objectives

Aim: to minimize a costly-to-evaluate function $y : \mathbf{x} \in D \subset \mathbb{R}^d \longrightarrow y(\mathbf{x}) \in \mathbb{R}$

Design of Experiments : $\mathbf{X}^n = \{\overbrace{\mathbf{x}^1, \dots \mathbf{x}^{n_0}}^{\text{Initial DoE}}, \overbrace{\mathbf{x}^{n_0+1}, \dots, \mathbf{x}^n}^{\text{Sequence}}\}$ s.t. $\forall i \leq n,\ \mathbf{x}^i \in D$

Assumption: y is one path of a Gaussian Process $(Y_{\mathbf{x}})_{\mathbf{x} \in D}$ with known kernel k and trend $\mu(\mathbf{x}) = \sum_{i=0}^{p} \beta_i f_i(\mathbf{x})$ known up to the linear coefficients $\beta_i \in \mathbb{R}$

General context: notation and objectives

Aim: to minimize a costly-to-evaluate function $y : \mathbf{x} \in D \subset \mathbb{R}^d \longrightarrow y(\mathbf{x}) \in \mathbb{R}$

Design of Experiments : $\mathbf{X}^n = \{\overbrace{\mathbf{x}^1, \ldots \mathbf{x}^{n_0}}^{\text{Initial DoE}}, \overbrace{\mathbf{x}^{n_0+1}, \ldots, \mathbf{x}^n}^{\text{Sequence}}\}$ s.t. $\forall i \leq n,\ \mathbf{x}^i \in D$

Assumption: y is one path of a Gaussian Process $(Y_\mathbf{x})_{\mathbf{x}\in D}$ with known kernel k and trend $\mu(\mathbf{x}) = \sum_{i=0}^{p} \beta_i f_i(\mathbf{x})$ known up to the linear coefficients $\beta_i \in \mathbb{R}$

Constraint: each $y(\mathbf{x}^i)$ is corrupted by an additive noise ε_i, where $var[\varepsilon_i] = \tau_i^2$

General context: notation and objectives

Aim: to minimize a costly-to-evaluate function $y : \mathbf{x} \in D \subset \mathbb{R}^d \longrightarrow y(\mathbf{x}) \in \mathbb{R}$

Design of Experiments : $\mathbf{X}^n = \{\overbrace{\mathbf{x}^1, \dots \mathbf{x}^{n_0}}^{\text{Initial DoE}}, \overbrace{\mathbf{x}^{n_0+1}, \dots, \mathbf{x}^n}^{\text{Sequence}}\}$ s.t. $\forall i \leq n,\ \mathbf{x}^i \in D$

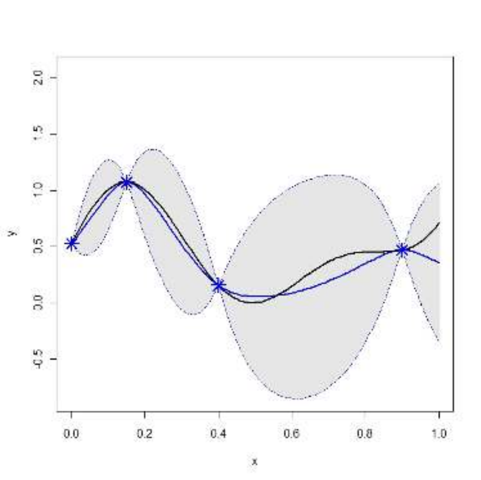
Assumption: y is one path of a Gaussian Process $(Y_{\mathbf{x}})_{\mathbf{x} \in D}$ with known kernel k and trend $\mu(\mathbf{x}) = \sum_{i=0}^{p} \beta_i f_i(\mathbf{x})$ known up to the linear coefficients $\beta_i \in \mathbb{R}$

Constraint: each $y(\mathbf{x}^i)$ is corrupted by an additive noise ε_i, where $var[\varepsilon_i] = \tau_i^2$

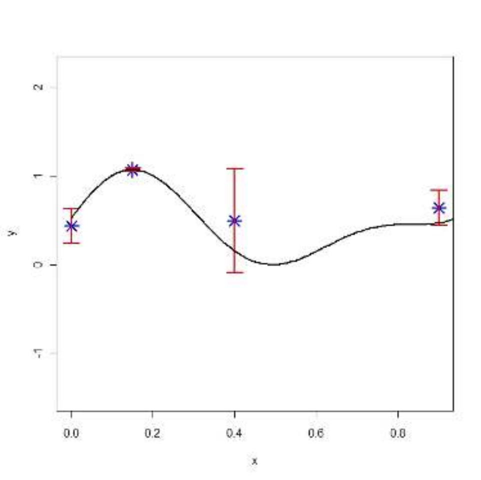
$$+ \text{ notations: } \begin{cases} \forall i \leq n,\ A_n \equiv \text{event } \{Y(\mathbf{x}^1) = y(\mathbf{x}^1), \dots, Y(\mathbf{x}^n) = y(\mathbf{x}^n)\} \\ \forall i \leq n,\ m_i(.),\ s_i^2(.) \equiv \text{Kriging mean and variance } |A_n \\ \mathbf{x}^{n+1}, Y_{n+1} := Y(\mathbf{x}^{n+1}) \equiv \text{next point and random response} \end{cases}$$

Outline

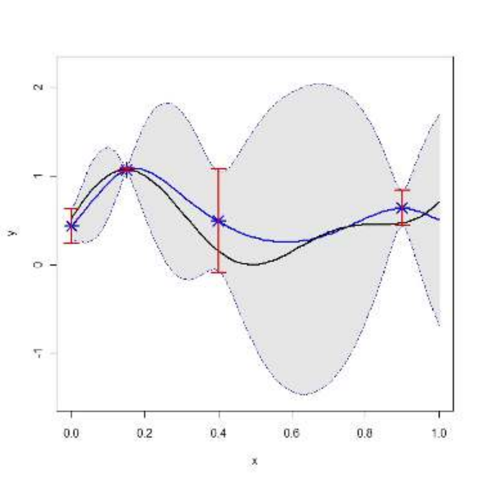
Example of Ordinary Kriging Interpolation



Kriging with heterogeneous noise?



Kriging with heterogeneous noise



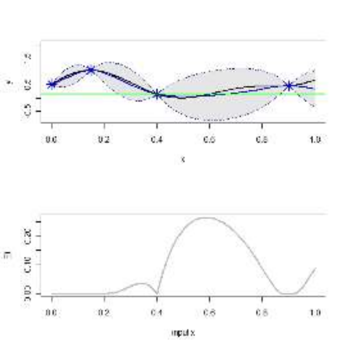
Outline

Introducing the noiseless EI criterion

Expected Improvement

$$EI_n(\mathbf{x}) = \mathbb{E}\left[\left(\min(Y(\mathbf{X}^n)) - Y(\mathbf{x})\right)^+ \,|\, Y(\mathbf{X}^n) = \mathbf{Y}^n \right]$$

M. Schonlau, W.J. Welch and D.R. Jones.
Efficient Global Optimization of Expensive Black-box Functions
Journal of Global Optimization, 1998.

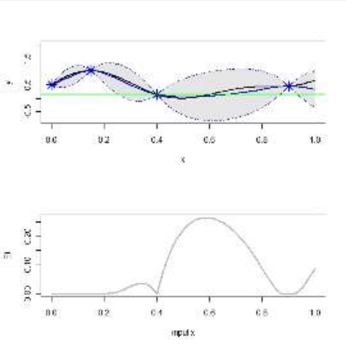


Introducing the noiseless EI criterion

Expected Improvement

$$EI_n(\mathbf{x}) = \mathbb{E}\left[\left(\min(Y(\mathbf{X}^n)) - Y(\mathbf{x}) \right)^+ | Y(\mathbf{X}^n) = \mathbf{Y}^n \right]$$

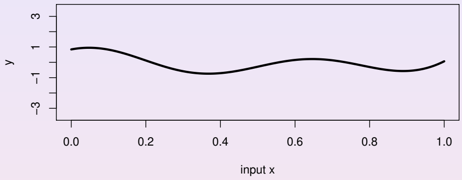
M. Schonlau, W.J. Welch and D.R. Jones.
Efficient Global Optimization of Expensive Black-box Functions
Journal of Global Optimization, 1998.



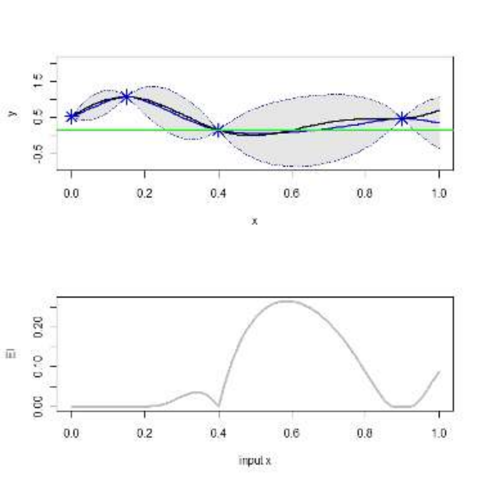
$$EI_n(\mathbf{x}) := (\min(\mathbf{Y}^n) - m_n(\mathbf{x}))\, \Phi\left(\frac{\min(\mathbf{Y}^n) - m_n(\mathbf{x})}{s_n(\mathbf{x})} \right) + s_n(\mathbf{x}) \phi\left(\frac{\min(\mathbf{Y}^n) - m_n(\mathbf{x})}{s_n(\mathbf{x})} \right),$$

where Φ and ϕ are respectively the cdf and the pdf of the standard gaussian law, and $\min(\mathbf{Y}^n) = \min\{y(\mathbf{x}^1), \ldots, y(\mathbf{x}^n)\}$ is the currently smallest observed value.

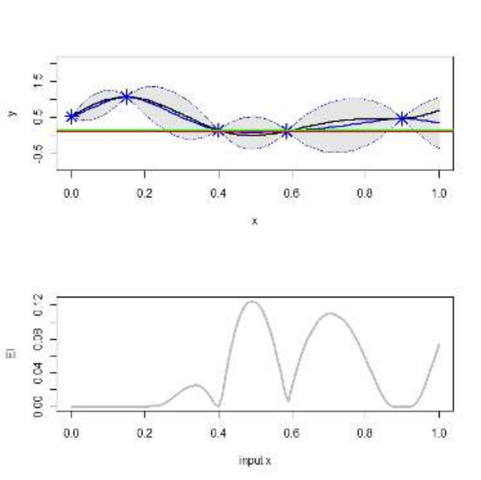
Illustrating EI using conditional simulation



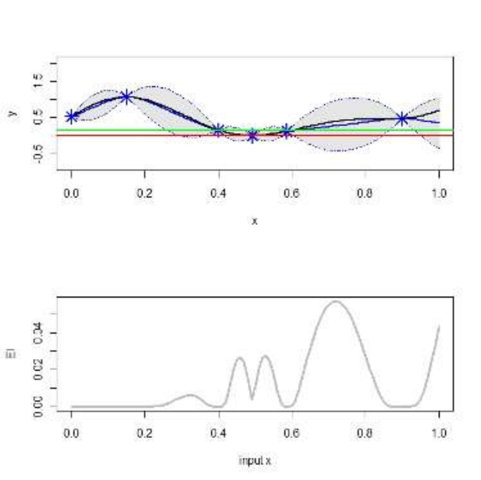
Kriging-based optimization with EGO



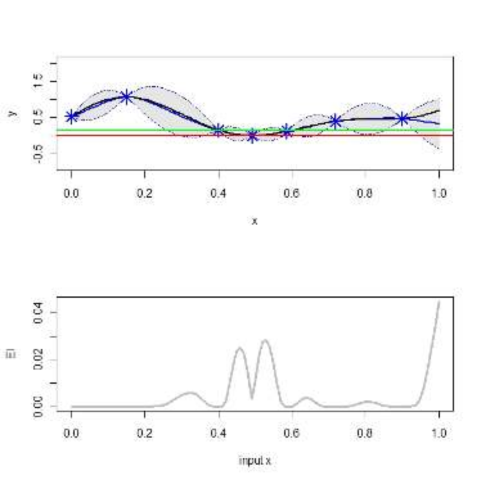
Kriging-based optimization with EGO



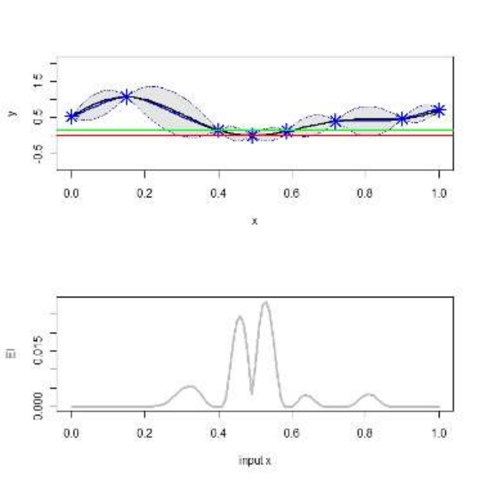
Kriging-based optimization with EGO



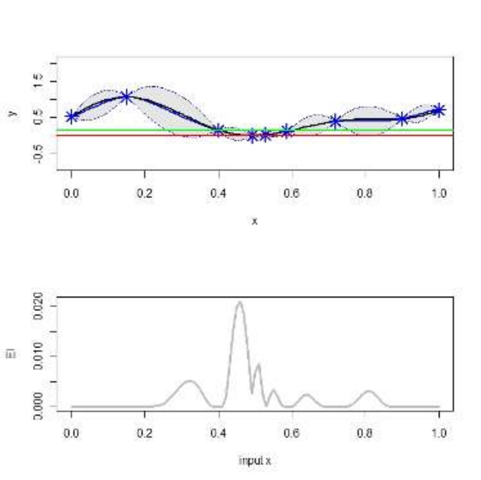
Kriging-based optimization with EGO



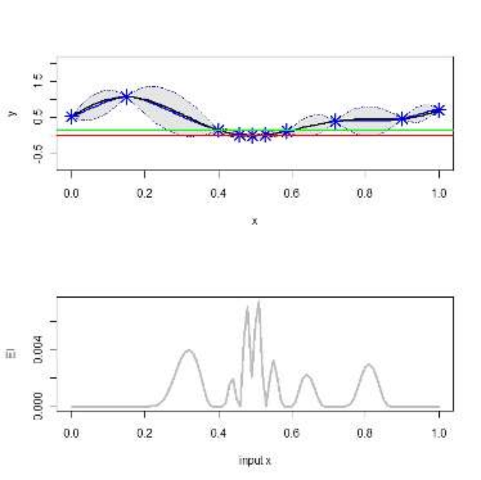
Kriging-based optimization with EGO



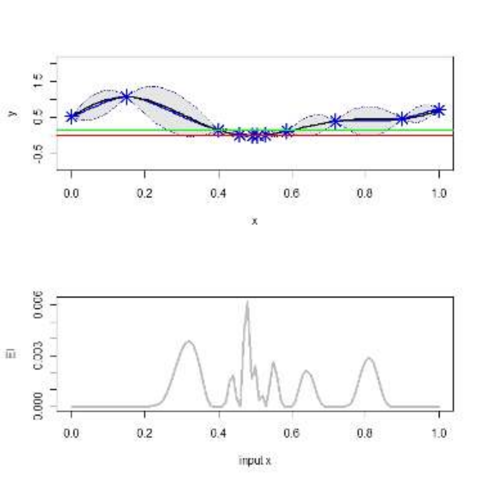
Kriging-based optimization with EGO



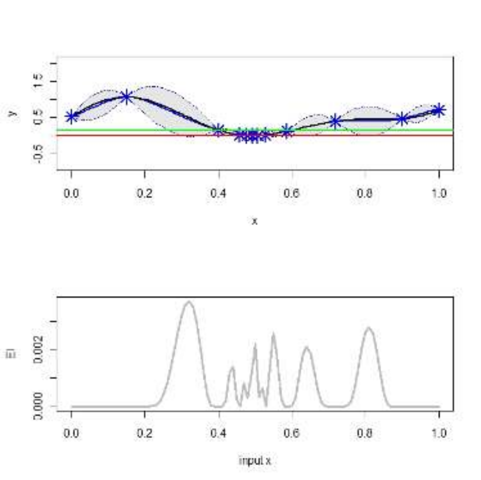
Kriging-based optimization with EGO



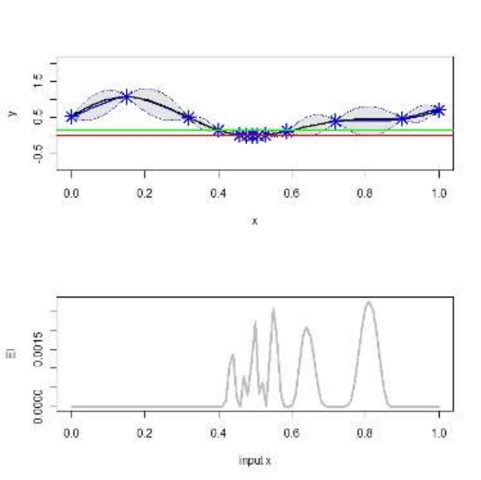
Kriging-based optimization with EGO



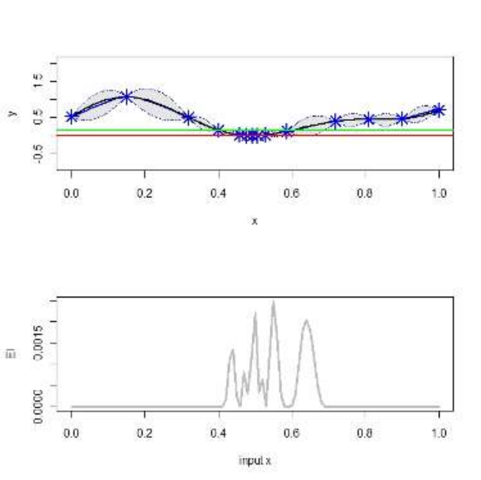
Kriging-based optimization with EGO



Kriging-based optimization with EGO



Kriging-based optimization with EGO



Outline

1. Generalities on Kriging and Kriging-based Optimization
   - From "deterministic" to "noisy" Kriging
   - Deterministic Expected Improvement and EGO

2. Noisy EI and random metamodels
   - Noisy EI: past and future noises
   - Where random metamodels emerge

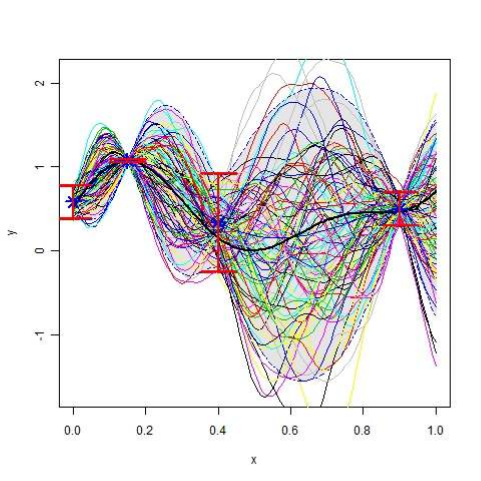
1. Conclusion and perspectives

What's wrong with EI and noisy observations?

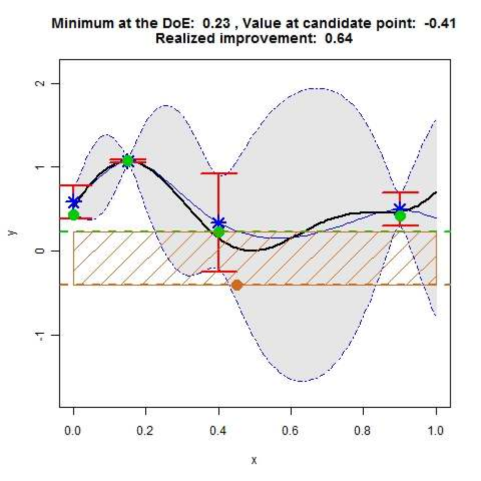
$$EI_n(\mathbf{x}) = \mathbb{E}\left[\left(\underbrace{\min(Y(\mathbf{X}^n))}_{unknown} - Y(\mathbf{x})\right)^+ \middle| Y(\mathbf{X}^n) + \varepsilon = \mathbf{Y^n} + \epsilon\right]$$

1. $\min_{\mathbf{x}\in\mathbf{X}^n}(Y(\mathbf{x}))$ is unfortunately not known conditionally on $Y(\mathbf{X}^n) + \varepsilon = \mathbf{Y^n} + \epsilon$ (it is not $(Y(\mathbf{X}^n) + \varepsilon)$-measurable)

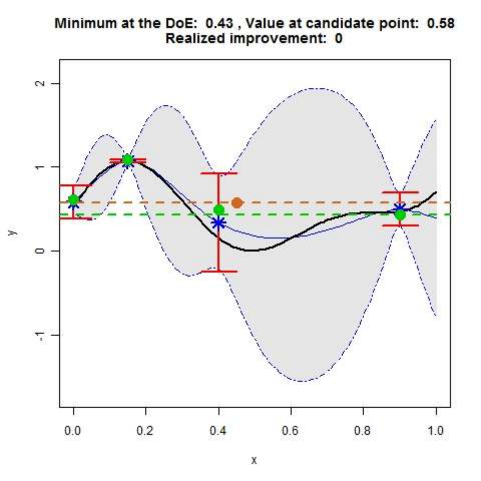
Monte Carlo Simulation. Example.



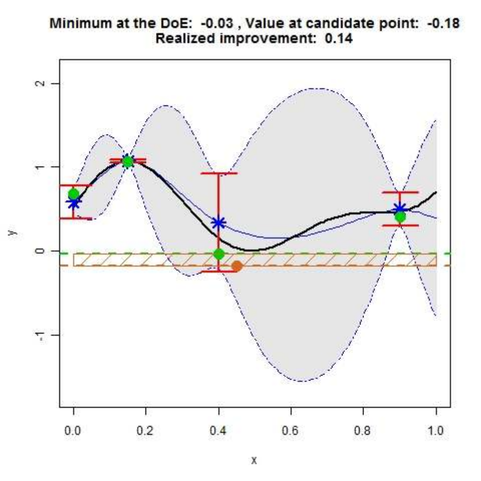
Monte Carlo Simulation. Example.



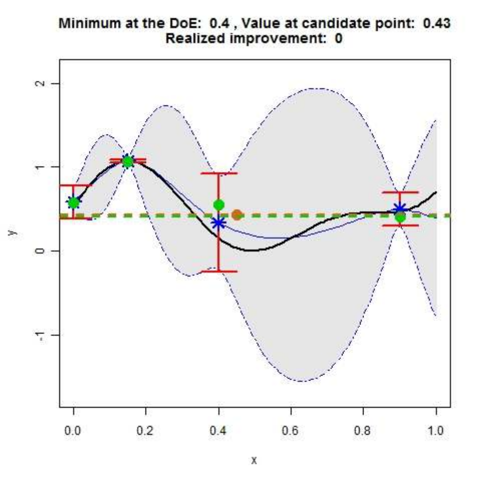
Monte Carlo Simulation. Example.



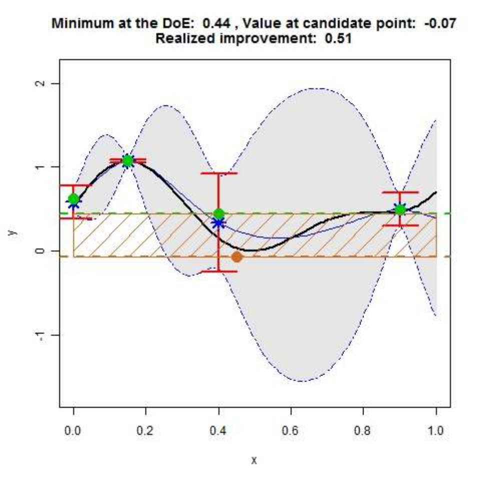
Monte Carlo Simulation. Example.



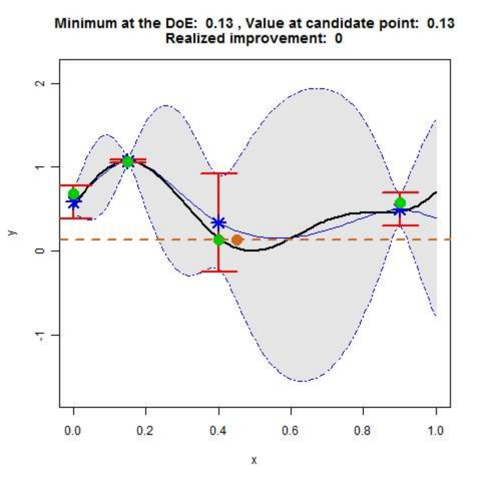
Monte Carlo Simulation. Example.



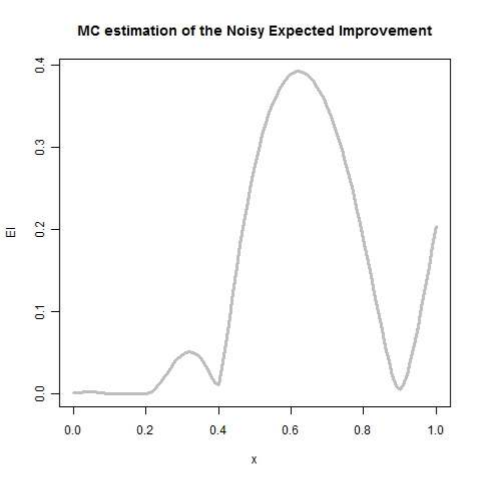
Monte Carlo Simulation. Example.



Monte Carlo Simulation. Example.

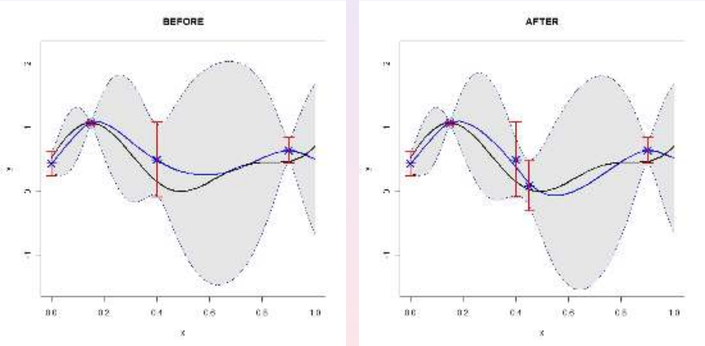


Monte Carlo Simulation. Example.



Why random metamodels?

What does ”improvement” mean when starting from a noisy metamodel and finishing with another one?



What's wrong with EI and noisy observations (bis)?

$$EI_n(\mathbf{x}) = \mathbb{E}\left[\left(\underbrace{\min(Y(\mathbf{X}^n))}_{\textit{unknown}} - \underbrace{Y(\mathbf{x})}_{\textit{unreachable}} \right)^+ \middle| Y(\mathbf{X}^n) + \varepsilon = \mathbf{Y^n} + \epsilon \right]$$

1. $\min_{\mathbf{x} \in \mathbf{X}^n}(Y(\mathbf{x}))$ is not known conditionally on $Y(\mathbf{X}^n) + \varepsilon$
2. $Y(\mathbf{x})$ will remain non-exactly known ...

A bit of formalism

Let $\mathbf{x}^{n+1}$ and τ_{n+1}^2 be the point visited at the n^{th} iteration and the corresponding noise variance. We consider them now as fixed.

Let $M_{n+1}(\mathbf{x}) := \mathbb{E}[Y_{\mathbf{x}}|A_n, Y_{n+1}]$ be the kriging mean function at the $(n+1)^{\text{th}}$ iteration and $V_{n+1}(\mathbf{x}) := Var[Y_{\mathbf{x}}|A_n, Y_{n+1}]$ the corresponding variance.

Conditional on A_n, $\begin{cases} M_{n+1}(.) \text{ is a } (Y_{n+1}\text{-measurable) random function} \\ V_{n+1}(.) \text{ is a deterministic function} \end{cases}$

Cool result

Conditional on A_n, the kriging mean at the $(n+1)^{\text{th}}$ iteration $M_{n+1}(.)$ is a GP

Cool result

Conditional on A_n, the kriging mean at the $(n+1)^{\text{th}}$ iteration $M_{n+1}(.)$ is a GP

Proof (Simple Kriging case for convenience).

$$M_{n+1}(\mathbf{x}) := \mathbb{E}[Y_{\mathbf{x}}|A_n, Y_{n+1}] = \overbrace{\mathbf{k}_{n+1}(\mathbf{x})^T (K_{n+1} + \Delta_{n+1})^{-1}}^{:=\alpha(\mathbf{x})=(\alpha_1(\mathbf{x}),\dots,\alpha_{n+1}(\mathbf{x}))} (Y_1, \dots, Y_n, Y_{n+1})^T$$

$$= \underbrace{\sum_{j=1}^{n} \alpha_j Y_j}_{A_n\text{-measurable}} + \underbrace{\alpha_{n+1} Y_{n+1}}_{\text{Gaussian conditional on } A_n}$$

□

Cool(er) result(s)

Let $Q_{n+1}(\mathbf{x})$ be the β-quantile ($\beta \in]0,1[$) of the conditional distribution of $Y_{\mathbf{x}}$ knowing A_n and Y_{n+1}.

Consequence of M_{n+1} being a GP and V_{n+1} being deterministic $|A_n$:

$$Q_{n+1}(.) = M_{n+1} + \Phi^{-1}(\beta)\sqrt{V_{n+1}} \text{ is a GP too}$$

Conclusion: we get an analytical quantile EI

$$EI_{Q_n}(\mathbf{x}) := \mathbb{E}\left[\left(\min_{i \leq n}(Q_n(\mathbf{x}^i)) - Q_{n+1}(\mathbf{x})\right)^+ \middle| A_n\right]$$

Conclusion: we get an analytical quantile EI

$$EI_{Q_n}(\mathbf{x}) := \mathbb{E}\left[\left(\min_{i \leq n}(Q_n(\mathbf{x}^i)) - Q_{n+1}(\mathbf{x})\right)^+ \Big| A_n\right]$$

Remark: τ_{n+1} is implicitely fixed. It might be a second argument of EI_{Q_n}...

Appendix 1: derivation of the quantile EI

$$EI_{Q_n}(\mathbf{x}) = \left(\min(\mathbf{q}_n) - m_{Q_{n+1}}(\mathbf{x})\right) \Phi\left(\frac{\min(\mathbf{q}_n) - m_{Q_{n+1}}(\mathbf{x})}{s_{Q_{n+1}}(\mathbf{x})}\right) + s_{Q_{n+1}}(\mathbf{x})\phi\left(\frac{\min(\mathbf{q}_n) - m_{Q_{n+1}}(\mathbf{x})}{s_{Q_{n+1}}(\mathbf{x})}\right)$$

where $m_{Q_{n+1}}(\mathbf{x}) = \mathbb{E}[Q_{n+1}(\mathbf{x})|A_n]$, $s_{Q_{n+1}}(\mathbf{x}) = \sqrt{Var[Q_{n+1}(\mathbf{x})|A_n]}$, and $\mathbf{q}_n = \{q_n(\mathbf{x}^i), i \leq n\}$ is the set of current quantile values at the already visited points.

Appendix 2: cool results (detail)

Conditional mean and covariance of $M_{n+1}(\mathbf{x})$

$$\mathbb{E}[M_{n+1}(\mathbf{x})|A_n] = \mathbf{k}_{n+1}(\mathbf{x})^T(K_{n+1} + \Delta_{n+1})^{-1}(y_1, \ldots, y_n, m_n(\mathbf{x}^{n+1}))^T$$

$$Cov[M_{n+1}(\mathbf{x}), M_{n+1}(\mathbf{x}')|A_n] = Cov[\alpha_{n+1}(\mathbf{x})Y_{n+1}, \alpha_{n+1}(\mathbf{x}')Y_{n+1}|A_n]$$

$$= \alpha_{n+1}(\mathbf{x})\alpha_{n+1}(\mathbf{x}')(v_n(\mathbf{x}^{n+1}) + \tau^{n+1})$$

Conditional mean and covariance of $Q_{n+1}(\mathbf{x})$ (see appendix)

$$\mathbb{E}[Q_{n+1}(\mathbf{x})|A_n] = \mathbb{E}[M_{n+1}(\mathbf{x})|A_n] + \Phi^{-1}(\beta)\sqrt{v_{n+1}(\mathbf{x})}$$

$$Cov[Q_{n+1}(\mathbf{x}), Q_{n+1}(\mathbf{x}')|A_n] = \alpha_{n+1}(\mathbf{x})\alpha_{n+1}(\mathbf{x}')(v_n(\mathbf{x}^{n+1}) + \tau^{n+1})$$