GDR MASCOT-NUM 2019

Brief overview on

Stepwise Uncertainty Reduction for computer experiments

Emmanuel Vazquez

March 19, 2010

Contents

1 Stepwise Uncertainty Reduction

- Framework: Bayesian sequential design of computer experiments
- Let ϕ be a quantity of interest that depends on a simulator with inputs in $\mathbb{X}$ and outputs in $\mathbb{Z}$
- Objective: make inference about ϕ from a finite set of experiments on the simulator (runs of the computer model)
- Usual Bayesian approach to describe such experiments: sequence of random variables $Z_1, Z_2, \ldots \in \mathbb{Z}$ modeling the outcomes of the experiments at points $X_1, X_2, \ldots \in \mathbb{X}$, using a prior distribution reflecting our belief about the simulator

- X_1 may be chosen according to an arbitrary distribution
- For all $n = 1, 2, \ldots$, X_{n+1} is a decision rule, which may depend on the information

$$I_n = \{(X_1, Z_1), \ldots, (X_n, Z_n)\}$$

  collected after n experiments

- The Z_is have a prior distribution defined through one or several random processes
- Particular case of a deterministic computer model $f : \mathbb{X} \to \mathbb{Z}$: we assume $Z_i = \xi(X_i)$, for all $i = 1, 2, \ldots$, where ξ is a random process that models f
- The sequence of decision rules $\underline{X} = (X_1, X_2, \ldots)$ is the sampling strategy

- Given an estimator $\widehat{\phi}_n$ of ϕ that depends on I_n, how to choose $\underline{X}$?
- Starting point of the construction of a SUR strategy:

  $\rightsquigarrow$ a statistic H_n based on $\widehat{\phi}_n$ measuring (residual) uncertainty about ϕ given I_n (DeGroot, 1962, Uncertainty, information and sequential experiments)
- H_n may be viewed as a "distance" between ϕ and $\widehat{\phi}_n$
- Residual uncertainty H_n is non-negative; $H_n = 0$ corresponds to the absence of uncertainty

Illustrative examples

1. Approximation of a deterministic computer model $f : \mathbb{X} \to \mathbb{R}$

   - f modeled using a GP ξ and $Z_i = \xi(X_i)$
   - Consider the kriging predictor $\widehat{\xi}_n = \mathsf{E}_n(\xi) = \mathsf{E}(\xi \mid I_n)$
   - Measures of uncertainty:
     - **IMSE:** $H_n^{(1)} = \mathsf{E}_n\left(\|(\xi - \widehat{\xi}_n)\|^2_{L^2(\mathbb{X},\mu)}\right) = \int_{\mathbb{X}} \sigma_n(x)^2 \mathrm{d}\mu(x)$, where $\sigma_n(x)^2$ is the kriging variance at $x \in \mathbb{X}$
     - **MMSE:** $H_n^{(2)} = \sup_{\mathbb{X}} \sigma_n \geq 1/\mu(\mathbb{X}) H_n^{(1)}$

2. Maximization of $f : \mathbb{X} \to \mathbb{R}$

   - $M = \sup_{\mathbb{X}} \xi$, $Z_i = \xi(X_i)$ and $M_n = Z_1 \vee \cdots \vee Z_n$
   - Measure of uncertainty: $H_n = \mathsf{E}_n\left(M - M_n\right)$

- A SUR strategy consists in selecting the next observation location using the rule

$$X_{n+1} = \arg\min_{x \in \mathbb{X}} J_n(x)$$

where J_n is called sampling criterion/acquisition function, and is defined by

$$J_n : x \mapsto \mathsf{E}_n\left(H_{n+1} | X_{n+1} = x\right)$$

(note that the expectation in J_n is with respect to Z_{n+1} when $X_{n+1} = x$)

- We can also define a notion of information gain function:

$$G_n : x \in \mathbb{X} \mapsto H_n - J_n(x)$$

- Bect et al. 2018 also define the notion of quasi-SUR strategies in the case of imperfect minimization of J_n: given a non-negative sequence $(u_n)_{n \in \mathbb{N}}$ with $u_n \searrow 0$, choose X_{n+1} such that $J(X_{n+1}) \leq \inf_{\mathbb{X}} J_n + u_n$

Side note: origins of SUR

- For sequential design of numerical experiments, idea proposed by myself and my co-authors for optimization and reliability problems (V. & Piera-Martinez, 2007; Villemonteix et al., 2008; V. & Bect, 2009)
- The parents: Stepwise Entropy Reduction for shape recognition (Geman & Jedynak, 1993), and active learning (MacKay 1992; Cohn et al. 1996)
- Many greedy Bayesian methods for risk minimization/utility maximization can actually be viewed as SUR strategies.

2 Examples of SUR criteria from the literature of optimization

Expected improvement (Mockus, Zilinksas ~ 1970)

- Objective: given a compact set $\mathbb{X} \subset \mathbb{R}^d$ and a smooth function $f : \mathbb{X} \to \mathbb{R}$, estimate $M = \sup_{\mathbb{X}} f$ using the estimator $M_n = f(X_1) \vee \cdots \vee f(X_n)$
- The efficiency of the optimization strategy $\underline{X}$ at iteration n can be measured using the loss function

$$\varepsilon_n(\underline{X}, f) = M - M_n$$

- Measure of residual uncertainty about M:

$$H_n = \mathsf{E}_n(\varepsilon_n(\underline{X}, \xi)) = \mathsf{E}_n(M - M_n)$$

- SUR strategy:

$$X_{n+1} = \arg\min_{x \in \mathbb{X}} \mathsf{E}_n\left(H_{n+1} \mid X_{n+1} = x\right)$$

- In terms of expected gain/information gain:

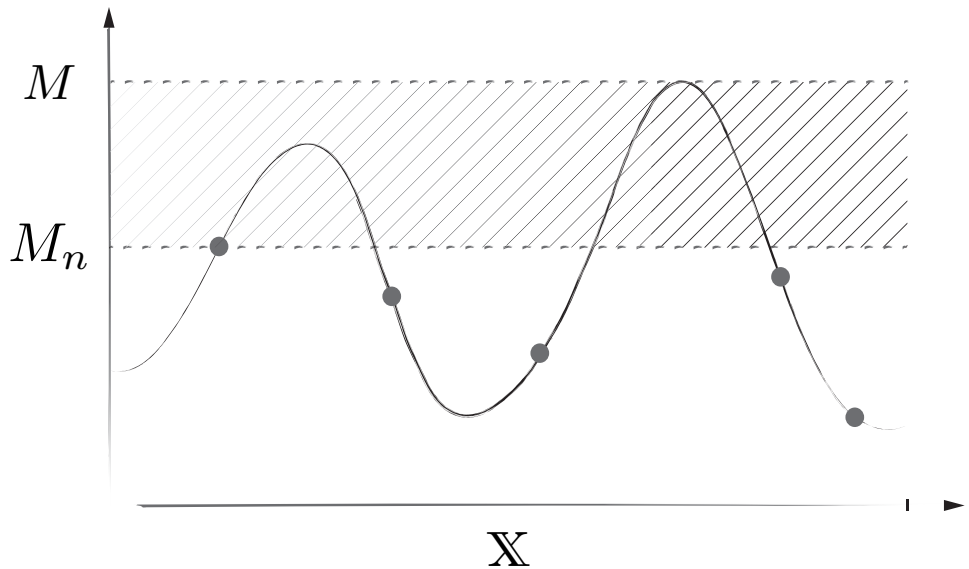
$$\begin{aligned} X_{n+1} &= \arg\max_{x \in \mathbb{X}} H_n - \mathsf{E}_n\left(H_{n+1} \,\middle|\, X_{n+1} = x\right) \\ &= \arg\max_{x \in \mathbb{X}} \mathsf{E}_n(M - M_n) - \mathsf{E}_n\left(M - M_{n+1} \,\middle|\, X_{n+1} = x\right) \\ &= \arg\max_{x \in \mathbb{X}} \mathsf{E}_n\left(M_{n+1} - M_n \,\middle|\, X_{n+1} = x\right) \end{aligned}$$

- $\rho_n(x) = \mathsf{E}_n\left(M_{n+1} - M_n \,\middle|\, X_{n+1} = x\right)$ is called the expected improvement (EI) sampling criterion (Mockus, Zilinksas, 1970-1980)

Expected integrated expected improvement

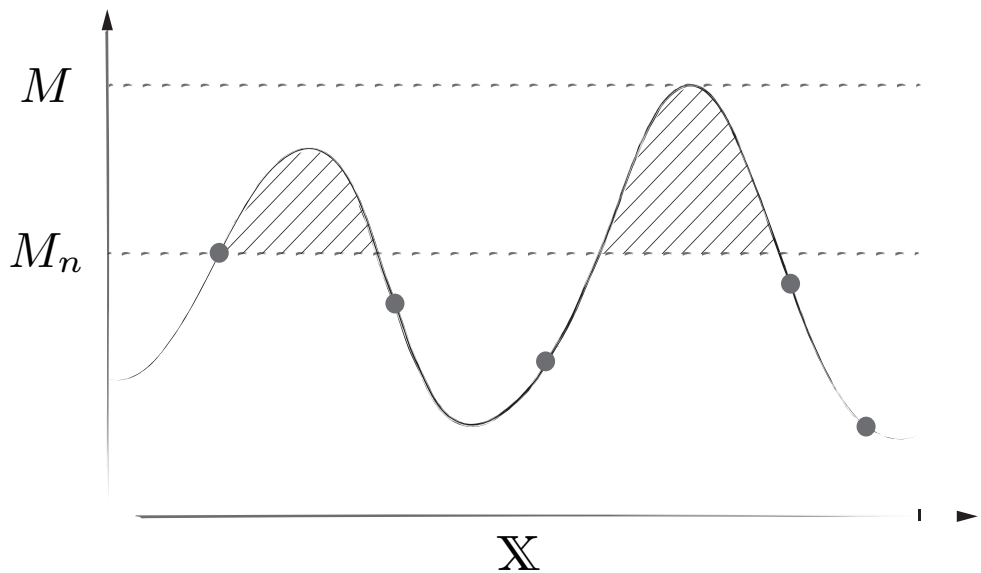
- In a global optimization problem, it is generally of interest to obtain a good approximation of both $M = \max_{\mathbb{X}} f$ and $x^\star = \arg\max_{\mathbb{X}} f$
- The loss function $\varepsilon_n(\underline{X}, f) = M - M_n$ does not measure directly the distance of $x_n^\star$ to $x^\star$ (with $x_n^\star$ such that $\xi(x_n^\star) = M_n$)

- However, note that $\varepsilon_n(\underline{X}, f) = M - M_n \propto \lambda(\mathbb{X})\,(M - M_n)$



$\rightarrow$ coarse measure of the uncertainty about the pair $(M, x^\star)$

- Consider instead the integral loss $\varepsilon_n(\underline{X}, f) = \int_{\mathbb{X}} (f(x) - M_n)_+ \, \lambda(\mathrm{d}x)$



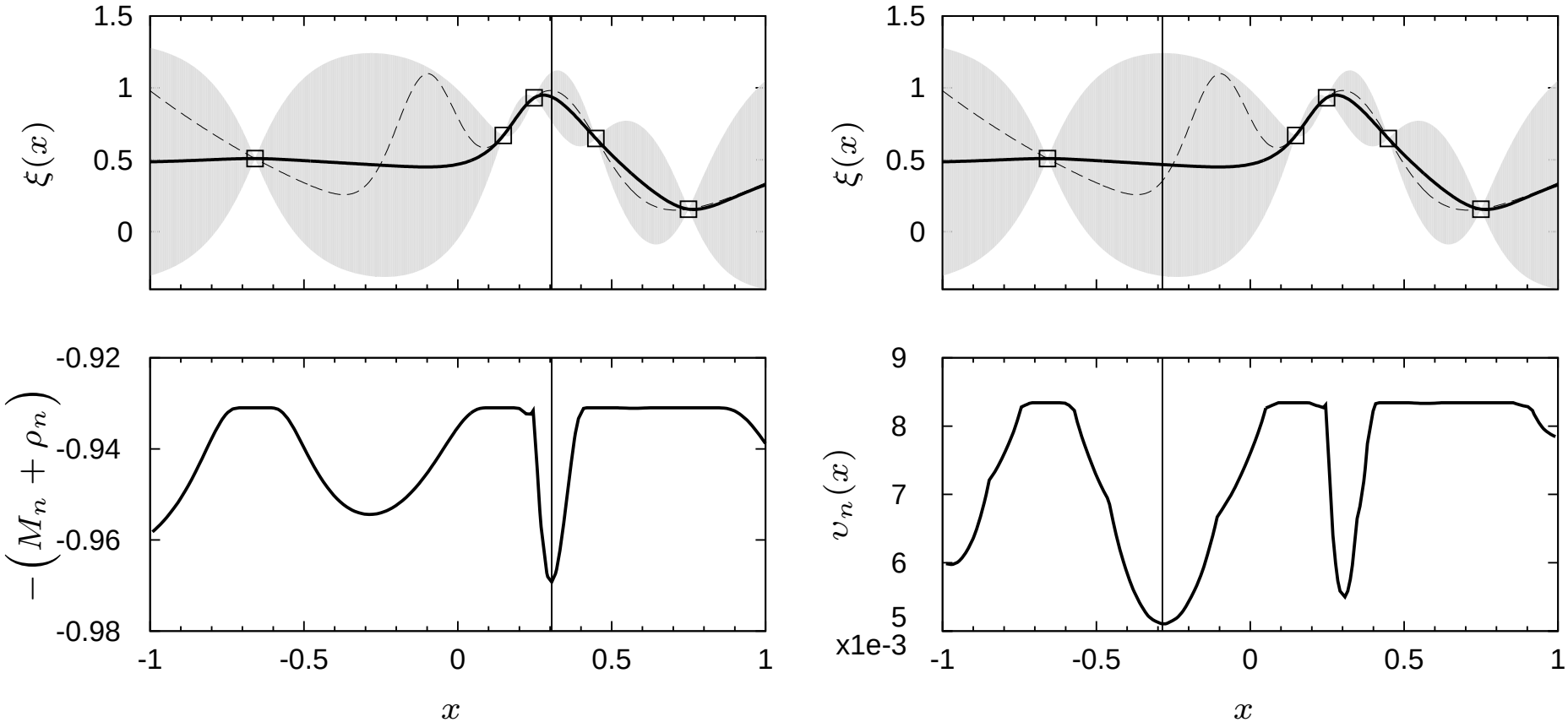
$\rightarrow$ ε_n gets smaller when uncertainty about $x^\star$ decreases

- Then, a SUR decision rule for choosing a new evaluation point X_{n+1} can be written as

$$\begin{aligned} X_{n+1} &= \arg\min_{x\in\mathbb{X}} \mathsf{E}_n\left(\int_{\mathbb{X}} (\xi(y) - M_{n+1})_+ \,\mathrm{d}y \;\middle|\; X_{n+1} = x\right) \\ &= \arg\min_{x\in\mathbb{X}} \mathsf{E}_n\left(\int_{\mathbb{X}} \mathsf{E}_{n+1}\big((\xi(y) - M_{n+1})_+\big) \,\mathrm{d}y \;\middle|\; X_{n+1} = x\right) \\ &= \arg\min_{x\in\mathbb{X}} \upsilon_n(x) := \mathsf{E}_n\left(\int_{\mathbb{X}} \rho_{n+1}(y)\mathrm{d}y \;\middle|\; X_{n+1} = x\right) \end{aligned}$$

- We call υ_n the Expected Integrated Expected Improvement (EI^2) (V. & Bect 2014)

Large expected improvement in a small region, smaller expected improvement over a large region of the search domain $\rightarrow$ here, υ_n favors better exploration than ρ_n



Knowledge Gradient (Frazier et al. 2008)

- Consider a stochastic computer model with scalar output $\rightsquigarrow$ observation model:

$$Z_i \mid \xi \overset{\text{iid}}{\sim} \mathcal{N}(\xi(X_i), \sigma^2), \quad i = 1, 2, \ldots$$

  with $\xi \mid m, k \sim \mathrm{GP}(m, k)$

- Objective: estimate $M = \sup_{\mathbb{X}} \xi$

- Which loss function for this problem?

$$\varepsilon_n(\underline{X}, \xi) = M - \cancel{M_n} \leftarrow \xi(X_n^\star)$$

  where $X_n^\star$ is an estimator of $X^\star = \arg\max_{\mathbb{X}} \xi(x)$, for instance:

$$X_n^\star = \arg\max_{\mathbb{X}} \widehat{\xi}_n(x)$$

- Measure of uncertainty: $H_n = \mathsf{E}_n(M - \xi(X_n^\star))$
- Then, a SUR strategy for this loss function may be written as

$$\begin{aligned} X_{n+1} &= \arg\max_{\mathbb{X}} H_n - \mathsf{E}_n\big(H_{n+1} \,\big|\, X_{n+1} = x\big) \\ &= \arg\max_{\mathbb{X}} \mathsf{E}_n\big(\xi(X_{n+1}^\star) - \xi(X_n^\star) \,\big|\, X_{n+1} = x\big) \\ &= \arg\max_{\mathbb{X}} \mathsf{E}_n\big(\xi(X_{n+1}^\star) \,\big|\, X_{n+1} = x\big) \\ &= \arg\max_{\mathbb{X}} \mathsf{E}_n\big(E_{n+1}\{\xi(X_{n+1}^\star)\} \,\big|\, X_{n+1} = x\big) \\ &= \arg\max \mathsf{E}_n\big(\widehat{\xi}_{n+1}(X_{n+1}^\star) \,\big|\, X_{n+1} = x\big) \end{aligned}$$

- The sampling criterion $\rho_n(x) = \mathsf{E}_n\big(\widehat{\xi}_{n+1}(X_{n+1}^\star) \,\big|\, X_{n+1} = x\big) - \widehat{\xi}_n(X_n^\star)$ is called the Knowledge Gradient (Frazier et al. 2008)

Expected hyper-volume improvement (Emmerich 2005)

- Multi-objective optimization: consider a set of functions $f_j : \mathbb{X} \to \mathbb{R}$, $j = 1, \ldots, p$, to be minimized
- Objective: build an approximation of the Pareto front and of the set of corresponding solutions

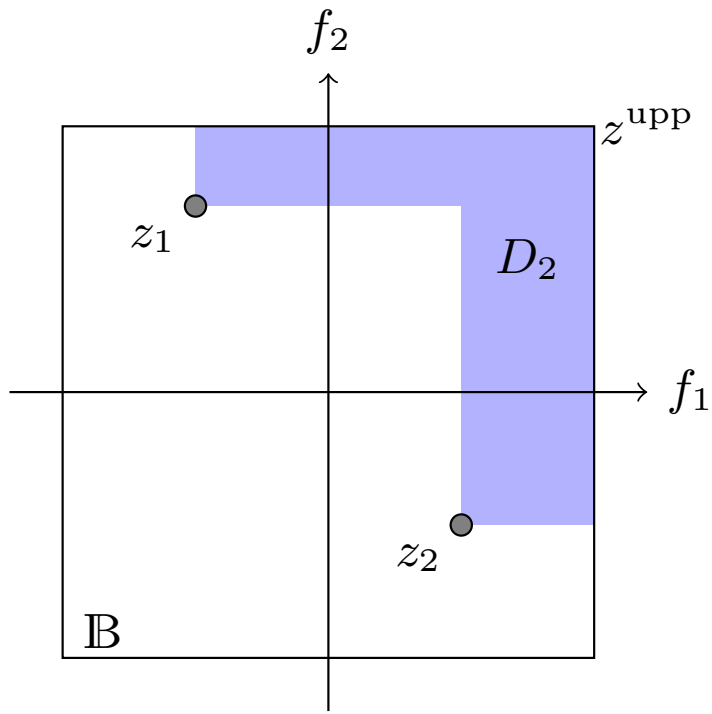
$$\Gamma = \{x \in \mathbb{X} : \nexists\, x' \in \mathbb{X} \text{ such that } \; f(x') \prec f(x)\}\,,$$

where $\prec$ stands for the Pareto domination rule defined by

$$y = (y_1, \ldots, y_p) \prec z = (z_1, \ldots, z_p) \Longleftrightarrow \left\{ \begin{array}{ll} \forall i \leq p, & y_i \leq z_i\,, \\ \exists j \leq p, & y_j < z_j\,. \end{array} \right.$$

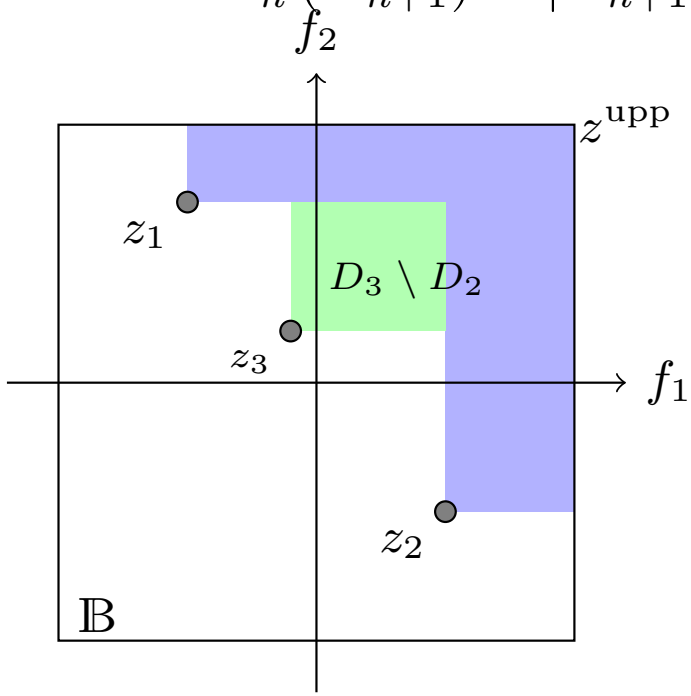
- Given evaluation results $z_1 = f(X_1), \ldots, z_n = f(X_n) \in \mathbb{R}^p$, define a set of dominated solutions:

$$D_n = \{y \in \mathbb{B}; \exists i \leq n,\ f(X_i) \prec y\}\,,$$



- Use the increase of the volume of the dominated region as information gain

$$G_n\left(X_{n+1}\right) = \left|D_{n+1} \setminus D_n\right| = \left|D_{n+1}\right| - \left|D_n\right|,$$



$\rightsquigarrow$ expected hyper-volume improvement (Emmerich 2005)

- Extended in (Feliot et al. 2015) to deal with constrained multi-objective problems and in (Feliot et al. 2018) to deal with user preferences for the exploration of a Pareto front

3 Examples from the literature of reliability

- Let $f : \mathbb{X} \to \mathbb{R}$ be a deterministic computer model, $u \in \mathbb{R}$, and consider the excursion set

$$\Gamma = \{x \in \mathbb{X} : f(x) > u\}$$

- An estimator $\widehat{\Gamma}_n$ of Γ can be obtained by building an approximation $\eta_n : \mathbb{X} \to \{0, 1\}$ of the excess indicator $\mathbb{1}_{f>u}$ from I_n
- For instance, we could use a support vector machine to build η_n (SMART method, Deheeger and Lemaire, 2007)

- Or we can use a GP ξ as a model of f and set

$$\eta_n(x) = \mathbb{1}_{p_n(x)>1/2} = \mathbb{1}_{\widehat{\xi}_n(x)>u}$$

where p_n is the posterior excursion probability:

$$p_n : x \in \mathbb{X} \mapsto \mathsf{P}_n\{\xi(x) > u\}$$

- Or we can use the notion of conservative estimates (Azzimonti et al. 2018): for a high probability β, define

$$\widehat{\Gamma}_n^\beta = \operatorname*{arg\,min}_{Q \in \mathcal{Q}_{n,\beta}} \mu(Q)$$

where $\mathcal{Q}_{n,\beta}$ is the family of Vorob'ev quantiles $Q_{n,\rho} = \{p_n \geq \rho\}$, $\rho \in [0,1]$, such that $\mathsf{P}_n(Q_{n,\rho} \supset \Gamma) \geq \beta$

- Related objectives:
  - given a measure/probability μ over $\mathbb{X}$, estimate the volume of the excursion set/probability of failure $\alpha = \mu(\Gamma)$, in which case a natural estimator for α is the posterior mean

$$\widehat{\alpha}_n = \mathsf{E}_n(\alpha) = \mathsf{E}_n\left(\int_{\mathbb{X}} \mathbb{1}_{\xi > u} \mathrm{d}\mu(x) \right) = \int_{\mathbb{X}} p_n \mathrm{d}\mu$$

  - quantile estimation: given a random input $X \in \mathbb{X}$ and $\alpha \in [0,1]$ estimate

$$q_\alpha = \inf \left\{ z \in \mathbb{R}; \mathsf{P}(f(X) \leq z) \geq \alpha \right\},$$

  - Given a stochastic simulator $x \mapsto \mathsf{P}_x = \mathcal{N}(\xi(x), \sigma^2)$, estimate $\alpha = \int_{\mathbb{X}} \mathsf{P}_x(]u, \infty[) \mathrm{d}\mu(x)$ (Stroh et al. 2017)

- Uncertainty measures?
- Assuming f is modeled using a random process ξ, V. and Piera-Martinez 2007, Bect et al 2012, Chevalier et al. 2014:
  - $H_n^{(1)} = \mathrm{var}_n(\alpha) = \mathsf{E}_n\big((\alpha - \widehat{\alpha}_n)^2\big)$
  - $H_n^{(2)} = \int_{\mathbb{X}} \mathsf{E}_n\left[(\mathbb{1}_{\xi>u} - \mathbb{1}_{\widehat{\xi}_n>u})^2\right]^{1/2} \mathrm{d}\mu$
  - $H_n^{(3)} = \mathsf{E}_n\Big(\mu\big(\Gamma \Delta \widehat{\Gamma}_n\big)\Big) = \mathsf{E}_n\left(\int_{\mathbb{X}} (\mathbb{1}_{\xi>u} - \mathbb{1}_{\widehat{\xi}_n>u})^2 \mathrm{d}\mu\right) = \int_{\mathbb{X}} p_n(1-p_n)\mathrm{d}\mu$
  - ...
- Note that $\left(H_n^{(1)}\right)^{1/2} \leq H_n^{(2)} \leq \left(H_n^{(3)}\right)^{1/2}$ by Minkowski and Cauchy-Schwarz inequalities
- If flagging "unsafe" regions as "safe" is bad $\rightsquigarrow H_n^{(4)} = \mathsf{E}\Big(\mu(\Gamma \setminus \widehat{\Gamma}_n)\Big)$ (Azzimonti et al. 2018)

4 Convergence results

- V. & Bect 2011 prove the consistency of the EI algorithm using two properties:

  1. $\operatorname{liminf}_{n\to\infty} \sup_{\mathbb{X}} G_n = 0$
  2. $\inf H_n > 0 \implies \operatorname{liminf}_{n\to\infty} \sup_{\mathbb{X}} G_n > 0$

- Bect et al. 2018 improve the result and show the consistency of several other SUR strategies. In particular, if $\forall x \in \mathbb{X}$

$$J_n(x) = \mathsf{E}_n(H_{n+1} \mid X_{n+1} = x) \leq H_n$$

  then $\sup_{\mathbb{X}} G_n \to 0$ a.s.

Rates?

- Almost unknown!
- For the optimization problem, Bull (2011) constructs an upper-bound of the convergence rate of the expected improvement strategy:

$$\sup_{\|f\|_{\mathcal{R}} \leq 1} M - M_n = O(n^{-(\nu \wedge 1)/d} (\log n)^{\beta})$$

- For the optimization of analytic functions using expected improvement, Yarotsky (2012) prove exponential convergence rates
- Open question: could we recover the optimal convergence rate of approximation for SUR strategies?

- In more details, let $\xi \sim \mathrm{GP}(0, k)$ and assume there exists $\Phi : \mathbb{R}^d \to \mathbb{R}$ with Fourier transform $\tilde{\Phi}$ satisfying

$$c_1(1 + \|u\|_2^2)^{-\nu - d/2} \leq \tilde{\Phi}(u) \leq c_2(1 + \|u\|_2^2)^{-\nu - d/2}\,, \quad u \in \mathbb{R}^d\,,$$

  such that $k(x, y) = \Phi(x - y)$ ($\nu > 0$, $0 < c_1 < c_2$)

- Then, Ritter (2010) shows that $e_n = \sup_{\mathbb{X}} \sigma_n(x) \geq C n^{-2\nu/d}$

- Moreover, V. & Bect (2011) show that the classical sequential non-adaptive strategy

$$X_{n+1} = \arg\max_{\mathbb{X}} \sigma_n(x)$$

  achieves $\sup_{\mathbb{X}} \sigma_n(x) = O(n^{2\nu/d})$

5 Concluding remarks

Computational cost

- SUR sampling criteria mostly often based on the computation of many posterior distributions $\rightsquigarrow$ non negligible computational additional cost
- Alleviate computational costs:
  - Bettinger et al. 2009 $\rightsquigarrow$ Tsallis entropy to maximize response diversity
  - Chevalier et al. 2014, Stroh et al. 2017 $\rightsquigarrow$ fast formulaes for uncertainty reduction about excursion sets
  - Hernandez-Lobato et al. 2014 $\rightsquigarrow$ predictive entropy search for optimization
  - Couckuyt et al. 2014, Hupkens et al. 2015, Zhao et al. 2019: fast computation of the EHVI

  - Labopin-Richard & Picheny, 2017 $\rightsquigarrow$ quantile estimation
  - ...
  - Batch evaluations: Ginsbourger et al. 2010, Chevalier et al. 2014, Dutrieux et al. 2015...

About priors

- Non stationary models?
  - Gaussian trees (Gramacy 2007)
  - Warped GP (Snelson et al 2004)...
  - Deep GP (Snoek et al. 2015, Hebbal et al. 2019...)
  - ...

Fully Bayesian vs empirical Bayes vs alternatives

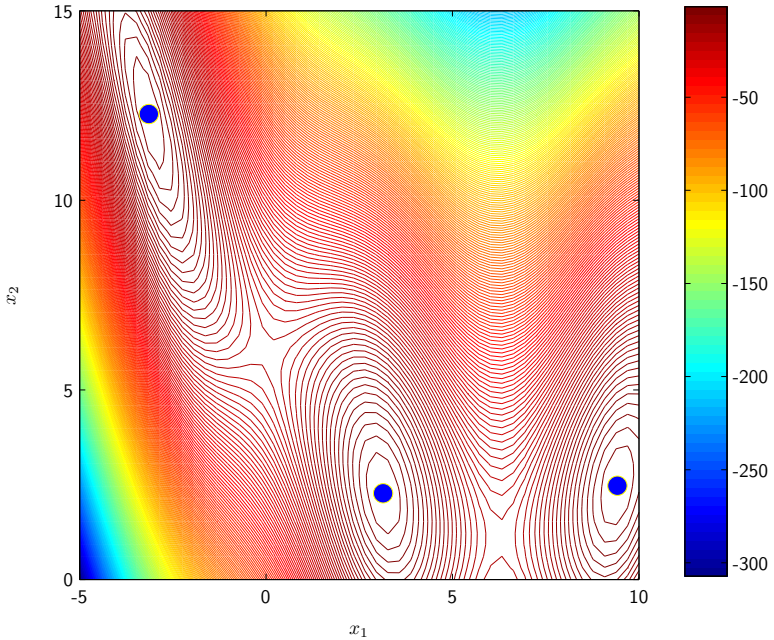
$\rightsquigarrow$ use a prior

$$\xi : \quad \begin{cases} \xi \mid \beta_1, \ldots, \beta_q, \theta \sim \mathrm{GP}\left(\sum_{i=1}^{q} \beta_i p_i,\, k_\theta\right), \\ \beta_1, \ldots, \beta_q \sim \mathrm{N}(0, \infty), \\ \theta \sim \pi_0, \end{cases}$$

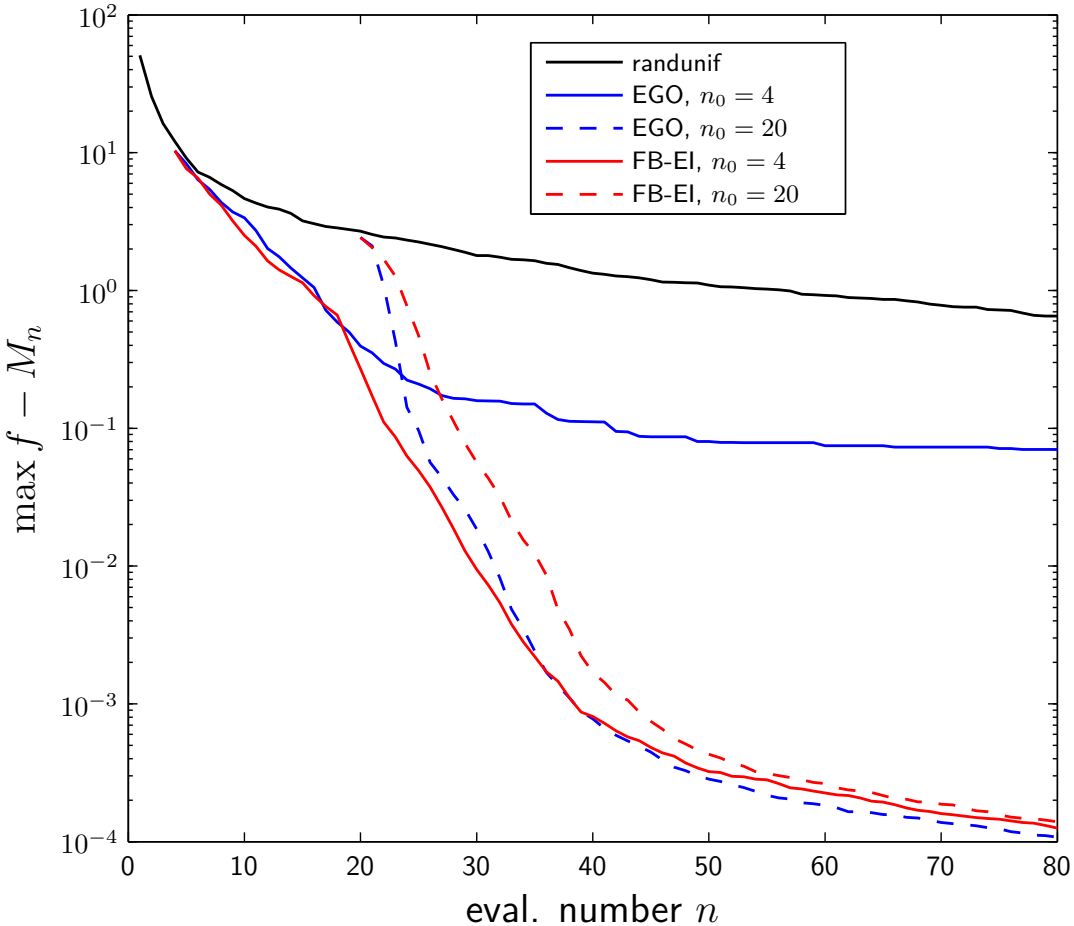
or estimate the prior from data inside a family of prior processes $\{\xi_\theta; \theta \in \Theta\}$ of the type

$$\xi_\theta : \quad \begin{cases} \xi_\theta \mid \beta_1, \ldots, \beta_q \sim \mathrm{GP}\left(\sum_{i=1}^{q} \beta_i p_i,\, k_\theta\right), \\ \beta_1, \ldots, \beta_q \sim \mathrm{N}(0, \infty) \end{cases} \qquad ?$$

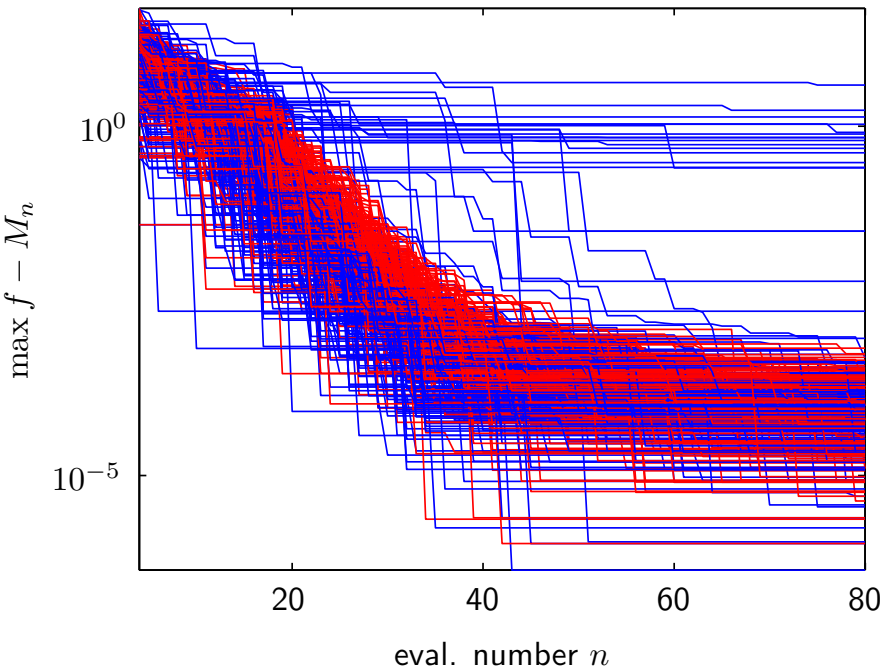
Possible shortcoming of the empirical Bayes approach (Benassi et al. 2011)



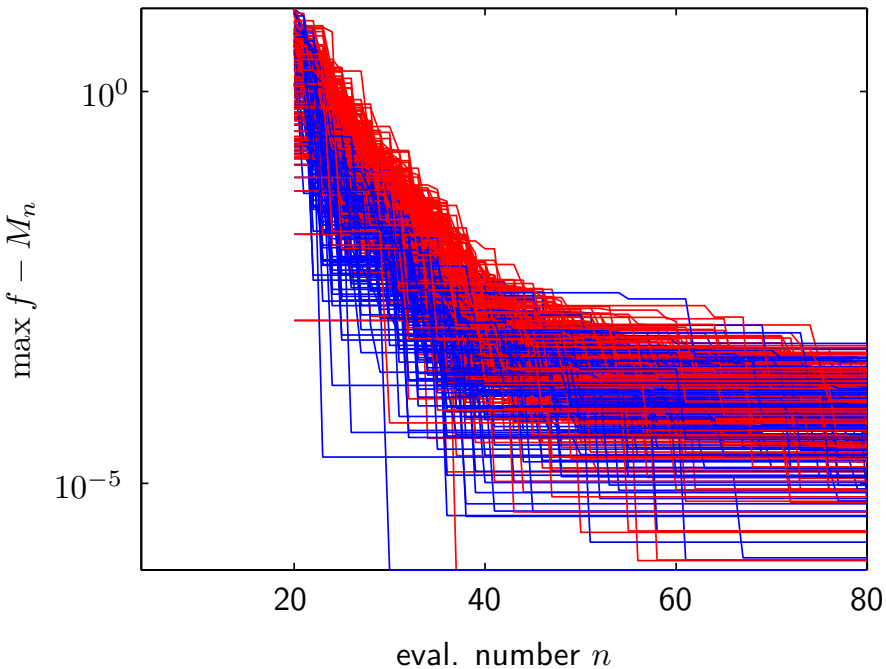
Branin function (three local maxima)



Average error (200 optim.) with random init. designs



(a) $n_0 = 4$



(b) $n_0 = 20$

Errors on each optimization run
(EGO in blue, FB-EI in red)

- Conclusion:
  - use a plugin method with a sufficiently large initial design
  - or a fully Bayesian approach
  - or maybe "modify" the kriging variance (Harville & Jeske, 1992, Abt 1999, Zimmerman 2006, Muller et al. 2010, Pronzato & Muller 2012...)

Future work?

- Research on SUR-like methods is under active development in the DACE/UQ community and also in machine learning
- Development of uncertainty measures for specific tasks (robust optimization, multi-fidelity...)
- Development of "good" models
- Better understanding of the properties of SUR strategies

GDR MASCOT-NUM 2019

Brief overview on

Stepwise Uncertainty Reduction for computer experiments

Emmanuel Vazquez

March 19, 2010