Sequential Design of Experiments with Gaussian Processes

Dario Azzimonti (slides) and **Cédric Travelletti** (tutorial)

LIKE22 -Lifting Inference with Kernel Embeddings

10th January 2022 - Bern, Switzerland

Contact: dario.azzimonti@idsia.ch

GP regression - recall from previous talk

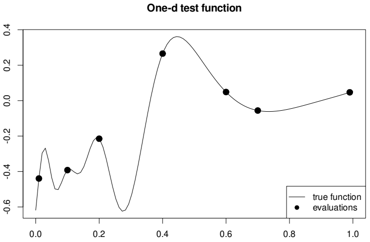
Data: $\mathcal{D} = (\mathbf{x}_i, y_i)_{i=1}^n$

$$y_i = f(\mathbf{x}_i) + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma_n^2)$$

$$\mathbf{x}_i \in D \subset \mathbb{R}^d$$

GP prior $Z \sim GP(m, k)$



GP regression - recall from previous talk

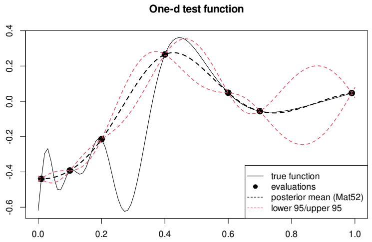
Data: $\mathcal{D} = (\mathbf{x}_i, y_i)_{i=1}^n$

$$y_i = f(\mathbf{x}_i) + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma_n^2)$$

$$\mathbf{x}_i \in D \subset \mathbb{R}^d$$

GP prior $Z \sim GP(m, k)$



GP regression:

$$m_n(\mathbf{x}) = m(\mathbf{x}) + k(\mathbf{x}, X_n)(k(X_n, X_n) + \sigma_n^2 I_n)^{-1}\mathbf{y}_n$$

$$k_n(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - k(\mathbf{x}, X_n)(k(X_n, X_n) + \sigma_n^2 I_n)^{-1}k(X_n, \mathbf{x}')$$

A GP as a surrogate for f

We observe $y_i = f(\mathbf{x}_i) + \epsilon_i$, the function f can be

- expensive computer experiment
  (e.g., PDE solver, expensive MC simulation)
- truly unknown function
  (e.g., y are real measurements of a phenomenon)

We can build a Design of Experiments (DoE) to estimate f.

Examples of fixed (non-sequential) DoE:

- space-filling sequences (e.g., Sobol');
- deterministic designs (e.g., Latin hypercube sample);
- statistically optimal designs (A-optimality, D-optimality).

What do we mean by sequential design of experiments?

Initial dataset: n observations $(\mathbf{x}_i, y_i)_{i=1}^n$, where $y_i = f(\mathbf{x}_i) + \epsilon_i$, ϵ_i realization from i.i.d. noise.

Initial GP fit: proceed as in previous talk and obtain m_n, s_n.

Sequential DoE: strategies that sequentially add new points.

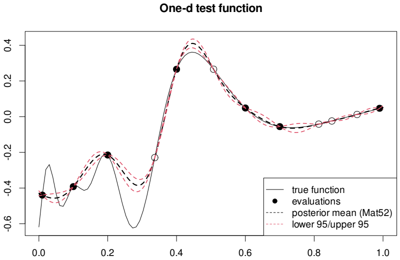
This talk

- what is the objective of adding new points?
  (learn f, finding global minimum of f, learn excursion set of f)
- strategies to achieve optimal DoE (e.g.: IMSE, EI, SUR)

What are the objectives of sequential DoE?

Today we will see

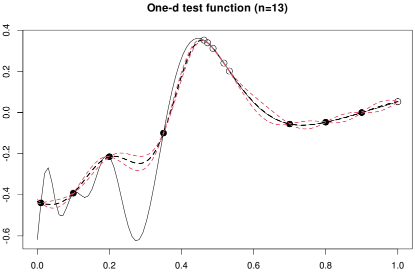
- improve the model and "learn" f,
  i.e. reduce posterior variance



What are the objectives of sequential DoE?

Today we will see

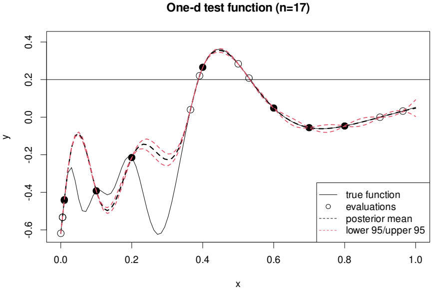
- improve the model and "learn" f
- Bayesian optimization:
  i.e. find $\mathbf{x}^* \in \arg\max_{\mathbf{x}\in D} f(\mathbf{x})$



What are the objectives of sequential DoE?

Today we will see

- improve the model and “learn” f
- Bayesian optimization
- learn level/excursion sets of f, i.e. estimate $\Gamma^\star = \{\mathbf{x} \in D : f(\mathbf{x}) \geq t\}$, fixed, known $t \in \mathbb{R}$.



Outline

Sequential DoE to improve the model
- MSE and IMSE criterion

Bayesian Optimization
- Upper Confidence Bound
- Expected Improvement
- Knowledge gradient
- Further topics

Target region estimation
- Targeted IMSE
- SUR for excursion set volume

Excursion set estimation
- Vorob'ev quantiles and Conservative estimates
- SUR strategies for conservative estimates

Sequential design of experiments

General procedure:

Start: train a surrogate model on an initial DoE;
Repeat until stopping criterion is met;

1. Use the surrogate model to define an **acquisition function**;
2. Optimize the acquisition function to find the next point(s);
3. Evaluate oracle to obtain new training data;
4. Update the surrogate model with new information;

What is the acquisition function? How to define it from a GP?

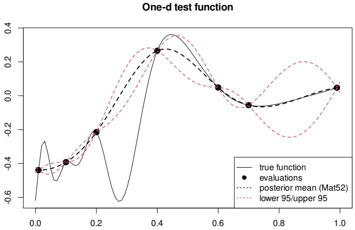
Outline

What does "improve the model" mean?

Regression data:

$\mathcal{D} = (\mathbf{x}_i, y_i)_{i=1}^n$

GP prior: $Z \sim GP(m, k)$



We compute the GP posterior $Z|\mathcal{D} \sim GP(m_n, k_n)$.

How can we improve this model? Reduce "error"

1. Mean Squared Error (MSE) criterion;
2. Integrated Mean Squared Error (IMSE) criterion.

GP regression as weighted interpolation

$Z \sim GP(m, k)$ with $m(\mathbf{x}) = 0$, $\mathbf{x} \in D$, k p.d. kernel.

Noiseless observations $Z_{\mathbf{x}_1}, \dots, Z_{\mathbf{x}_n}$ at $\mathbf{X}_n = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$.

"Best" linear predictor for $Z_{\mathbf{x}}$?

We are looking for $\widehat{Z}_{\mathbf{x}} = \sum_{i=1}^{n} \lambda_i Z_{\mathbf{x}_i}$ which is

unbiased $\mathbb{E}[\widehat{Z}_{\mathbf{x}}] = \mathbb{E}[Z_{\mathbf{x}}]$

$\lambda \in \arg\min \mathrm{MSE}(\lambda) = \mathbb{E}[(\widehat{Z}_{\mathbf{x}} - Z_{\mathbf{x}})^2]$

We obtain $\lambda(\mathbf{x}) = k(\mathbf{X}_n, \mathbf{X}_n)^{-1} k(\mathbf{X}_n, \mathbf{x})$ and

$$\begin{aligned}\mathrm{MSE}(\lambda(\mathbf{x})) &= \lambda(\mathbf{x})^T k(\mathbf{X}_n, \mathbf{X}_n)\lambda(\mathbf{x}) - 2\lambda(\mathbf{x})^T k(\mathbf{X}_n, \mathbf{x}) + k(\mathbf{x}, \mathbf{x}) \\ &= k(\mathbf{x}, \mathbf{x}) - k(\mathbf{x}, \mathbf{X}_n) k(\mathbf{X}_n, \mathbf{X}_n)^{-1} k(\mathbf{X}_n, \mathbf{x})\end{aligned}$$

SK equations

GP regression as weighted interpolation

$Z \sim GP(m, k)$ with $m(\mathbf{x}) = 0$, $\mathbf{x} \in D$, k p.d. kernel.

Noiseless observations $Z_{\mathbf{x}_1}, \ldots, Z_{\mathbf{x}_n}$ at $\mathbf{X}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$.

"Best" linear predictor for $Z_{\mathbf{x}}$?

We are looking for $\widehat{Z}_{\mathbf{x}} = \sum_{i=1}^{n} \lambda_i Z_{\mathbf{x}_i}$ which is

unbiased $\mathbb{E}[\widehat{Z}_{\mathbf{x}}] = \mathbb{E}[Z_{\mathbf{x}}]$

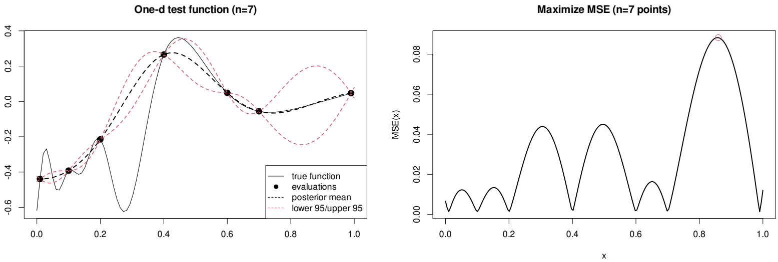
$\lambda \in \arg\min \mathrm{MSE}(\lambda) = \mathbb{E}[(\widehat{Z}_{\mathbf{x}} - Z_{\mathbf{x}})^2]$

We obtain $\lambda(\mathbf{x}) = k(\mathbf{X}_n, \mathbf{X}_n)^{-1} k(\mathbf{X}_n, \mathbf{x})$ and

$$\begin{aligned} \mathrm{MSE}(\lambda(\mathbf{x})) &= \lambda(\mathbf{x})^T k(\mathbf{X}_n, \mathbf{X}_n)\lambda(\mathbf{x}) - 2\lambda(\mathbf{x})^T k(\mathbf{X}_n, \mathbf{x}) + k(\mathbf{x}, \mathbf{x}) \\ &= k(\mathbf{x}, \mathbf{x}) - k(\mathbf{x}, \mathbf{X}_n) k(\mathbf{X}_n, \mathbf{X}_n)^{-1} k(\mathbf{X}_n, \mathbf{x}) \end{aligned}$$

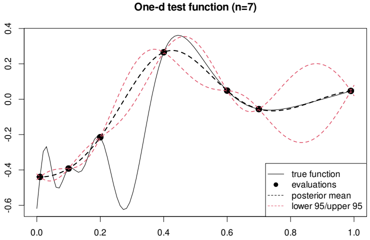
SK equations

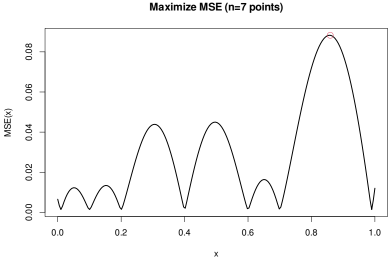
Improve the model by adding evaluations at maximum MSE points?



Mean Squared Error criterion

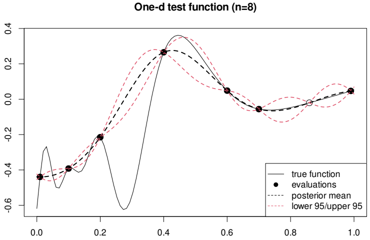
Criterion: select $\mathbf{x}_{n+1} \in \arg\max_{\mathbf{x} \in D} MSE_n(\mathbf{x})$

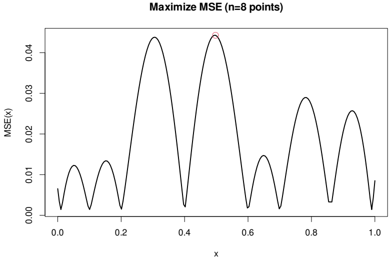




Mean Squared Error criterion

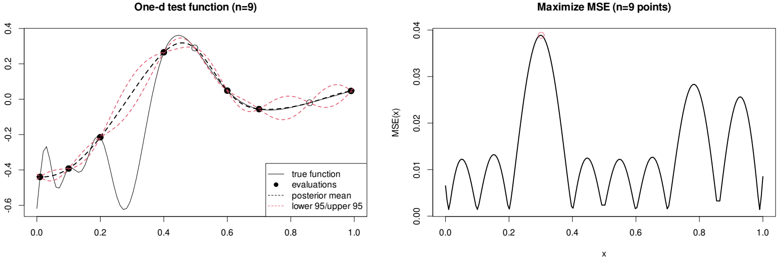
Criterion: select $\mathbf{x}_{n+1} \in \arg\max_{\mathbf{x} \in D} MSE_n(\mathbf{x})$





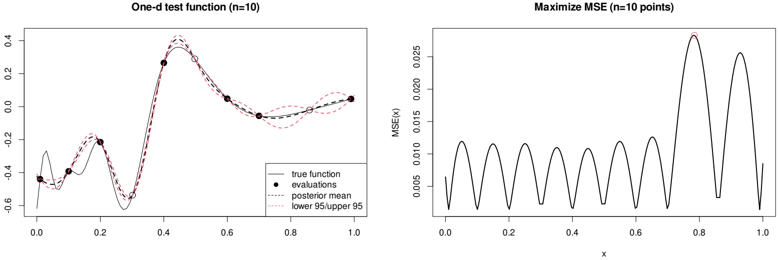
Mean Squared Error criterion

Criterion: select $\mathbf{x}_{n+1} \in \arg\max_{\mathbf{x} \in D} MSE_n(\mathbf{x})$



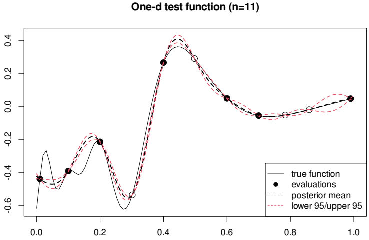
Mean Squared Error criterion

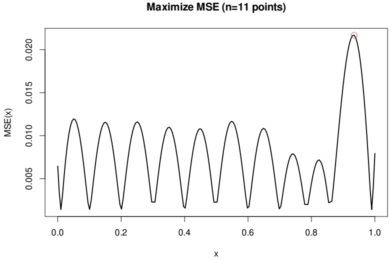
Criterion: select $\mathbf{x}_{n+1} \in \arg\max_{\mathbf{x} \in D} MSE_n(\mathbf{x})$



Mean Squared Error criterion

Criterion: select $\mathbf{x}_{n+1} \in \arg\max_{\mathbf{x} \in D} MSE_n(\mathbf{x})$





Integrated Mean Squared Error criterion

A global quantification of the uncertainty on the current model

$$H_n^{IMSE} = \int_D MSE_n(\mathbf{u}) d\mathbf{u},$$

(possible to generalize to a finite measure μ on D)

H_n^{IMSE} only considers the first n observations, can we "look-ahead"?

At step n we want $H_{n+1}^{IMSE} = \int_D MSE_{n+1}(\mathbf{u}; \mathbf{x}_{n+1}) d\mathbf{u}$,

where $MSE_{n+1}(\mathbf{u}; \mathbf{x})$ is the MSE for the model conditioned on $\{\mathbf{X}_n, \mathbf{x}\}$.

Can we compute $\mathbb{E}_n[H_{n+1}^{IMSE} \mid \mathbf{x}_{n+1} = \mathbf{x}]$?

Integrated Mean Squared Error criterion

A global quantification of the uncertainty on the current model

$$H_n^{IMSE} = \int_D MSE_n(\mathbf{u})d\mathbf{u},$$

(possible to generalize to a finite measure μ on D)

H_n^{IMSE} only considers the first n observations, can we "look-ahead"?

At step n we want $H_{n+1}^{IMSE} = \int_D MSE_{n+1}(\mathbf{u}; \mathbf{x}_{n+1})d\mathbf{u}$,

where $MSE_{n+1}(\mathbf{u}; \mathbf{x})$ is the MSE for the model conditioned on $\{\mathbf{X}_n, \mathbf{x}\}$.

Can we compute $\mathbb{E}_n[H_{n+1}^{IMSE} \mid \mathbf{x}_{n+1} = \mathbf{x}]$?

Integrated Mean Squared Error

In the GP case we can compute the criterion:

1. Compute $MSE_{n+1}(\mathbf{u})$ without the $n+1$ evaluation:

$$MSE_{n+1}(\mathbf{u};\mathbf{x}) = \lambda_{n+1}(\mathbf{u})^T k(\mathbf{X}_{n+1}(\mathbf{x}), \mathbf{X}_{n+1}(\mathbf{x}))^{-1} \lambda_{n+1}(\mathbf{u}) \\ - 2\lambda_{n+1}(\mathbf{u})^T k(\mathbf{X}_{n+1}(\mathbf{x}), \mathbf{u}) + k(\mathbf{u}, \mathbf{u}) = k_{n+1,\mathbf{x}}(\mathbf{u}, \mathbf{u})$$

where $\mathbf{X}_{n+1}(\mathbf{x}) = \{\mathbf{X}_n, \mathbf{x}\}$ is the augmented design.

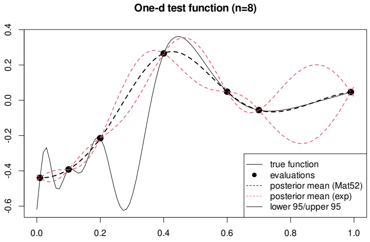
2. the criterion then is

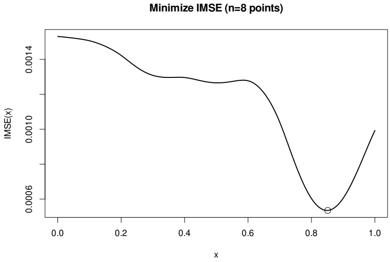
$$\mathbf{x}_{n+1} \in \arg\min_{\mathbf{x}\in D} \mathbb{E}_n[H_{n+1}^{IMSE}] = \arg\min_{\mathbf{x}\in D} \int_D k_{n+1,\mathbf{x}}(\mathbf{u}, \mathbf{u}) d\mathbf{u}$$

Sacks, J., Welch, W.J., Mitchell, T.J., Wynn, H.P. (1989). *Design and analysis of computer experiments.* Statistical Science 4(4), 409–435.

Integrated Mean Squared Error criterion

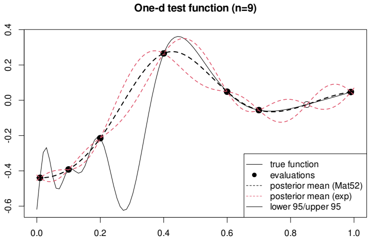
Criterion: select $\mathbf{x}_{n+1} \in \arg\min_{\mathbf{x}\in D} \mathbb{E}_n[H_{n+1}^{IMSE} \mid \mathbf{x}_{n+1} = \mathbf{x}]$

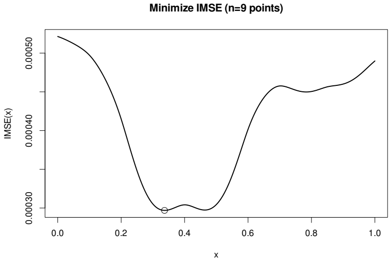




Integrated Mean Squared Error criterion

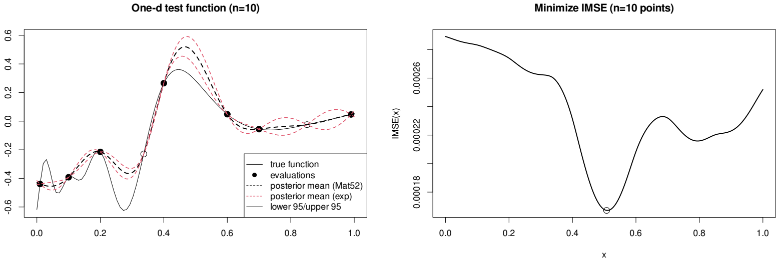
Criterion: select $\mathbf{x}_{n+1} \in \arg\min_{\mathbf{x}\in D} \mathbb{E}_n[H_{n+1}^{IMSE} \mid \mathbf{x}_{n+1} = \mathbf{x}]$





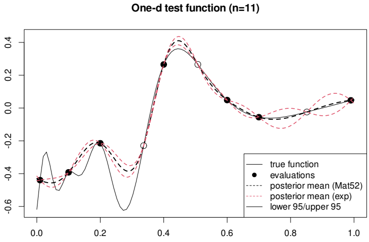
Integrated Mean Squared Error criterion

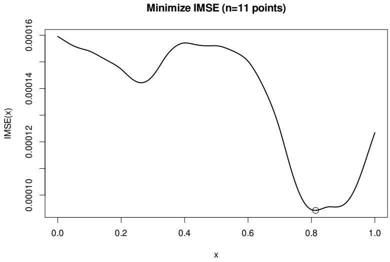
Criterion: select $\mathbf{x}_{n+1} \in \arg\min_{\mathbf{x}\in D} \mathbb{E}_n[H_{n+1}^{IMSE} \mid \mathbf{x}_{n+1} = \mathbf{x}]$



Integrated Mean Squared Error criterion

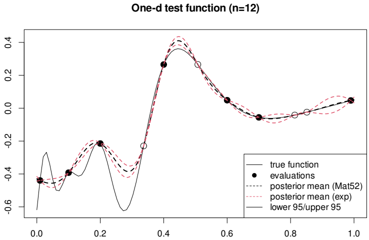
Criterion: select $\mathbf{x}_{n+1} \in \arg\min_{\mathbf{x} \in D} \mathbb{E}_n[H_{n+1}^{IMSE} \mid \mathbf{x}_{n+1} = \mathbf{x}]$

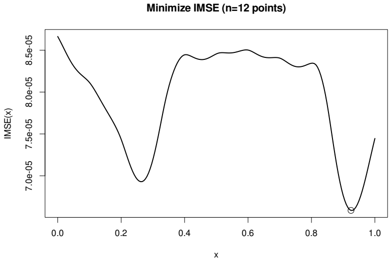




Integrated Mean Squared Error criterion

Criterion: select $\mathbf{x}_{n+1} \in \arg\min_{\mathbf{x}\in D} \mathbb{E}_n[H_{n+1}^{IMSE} \mid \mathbf{x}_{n+1} = \mathbf{x}]$





References

Sacks, J., Welch, W.J., Mitchell, T.J., Wynn, H.P. (1989). *Design and analysis of computer experiments*. Statistical Science 4(4), 409–435.

Santner, T. J., Williams, B. J., and Notz, W. I. (2018). The Design and Analysis of Computer Experiments. Springer New York.

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):509–534.

Chevalier C., Ginsbourger D., Emery X. (2014) *Corrected Kriging Update Formulae for Batch-Sequential Data Assimilation*. Lecture Notes in Earth System Sciences. Springer, Berlin, Heidelberg.

Outline

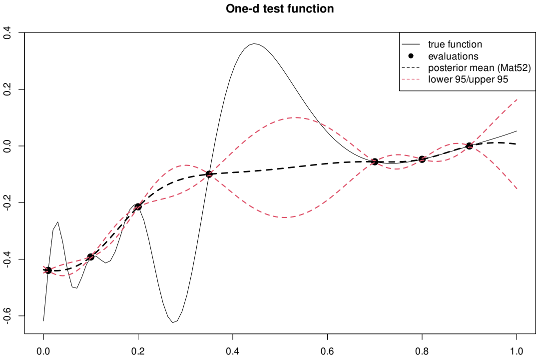
Bayesian optimization

Objective:

find global optimum

Tools:

GP predictions

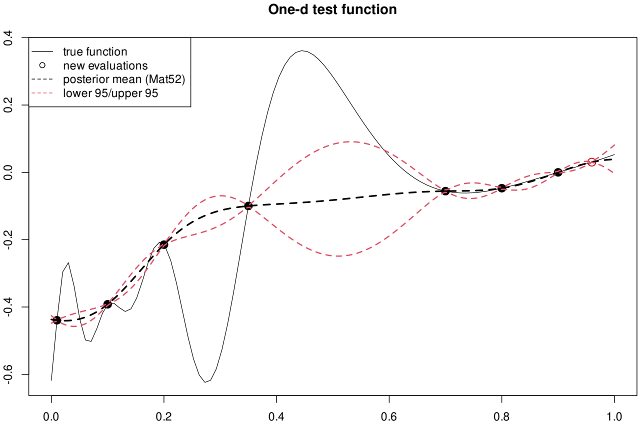


Questions:

1. where should we evaluate the function next?
2. what can we use to guide our exploration?

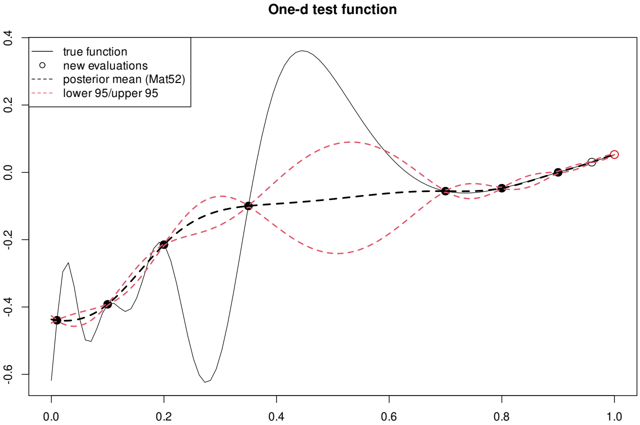
How to find the maximum?

First strategy: $\mathbf{x}_{n+1} \in \arg\max_{\mathbf{x} \in D} m_n(\mathbf{x})$ (maximizer of posterior mean)



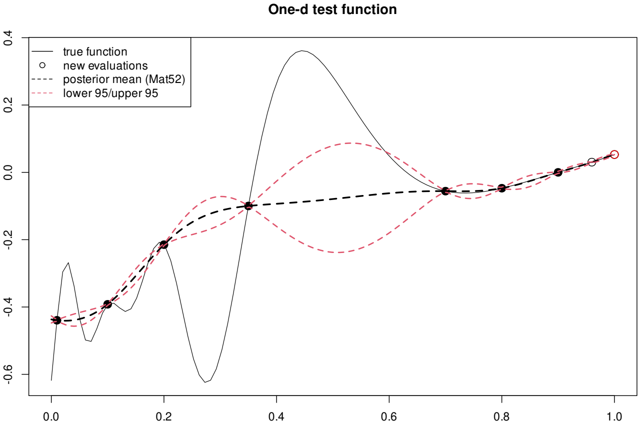
How to find the maximum?

First strategy: $\mathbf{x}_{n+1} \in \arg\max_{\mathbf{x}\in D} m_n(\mathbf{x})$ (maximizer of posterior mean)



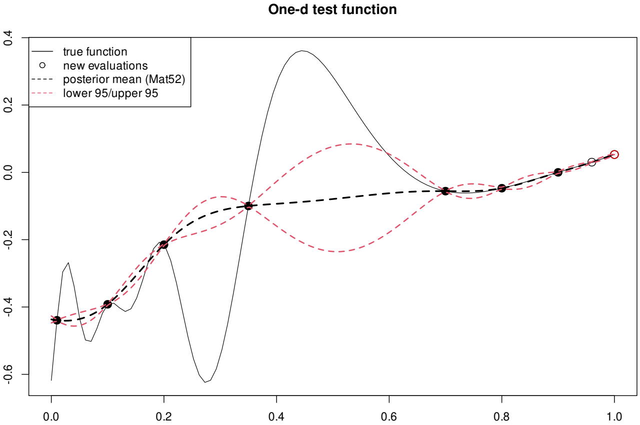
How to find the maximum?

First strategy: $\mathbf{x}_{n+1} \in \arg\max_{\mathbf{x}\in D} m_n(\mathbf{x})$ (maximizer of posterior mean)



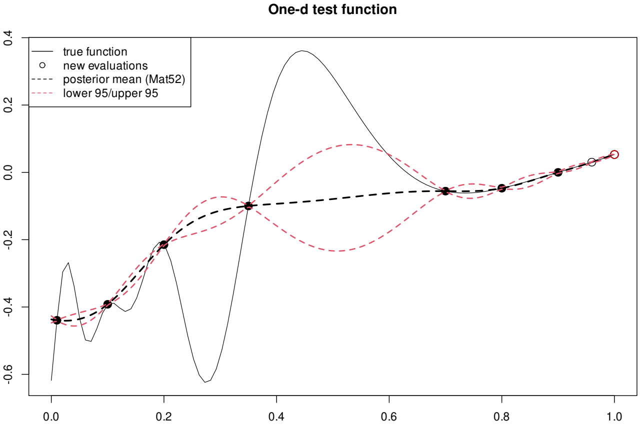
How to find the maximum?

First strategy: $\mathbf{x}_{n+1} \in \arg\max_{\mathbf{x} \in D} m_n(\mathbf{x})$ (maximizer of posterior mean)



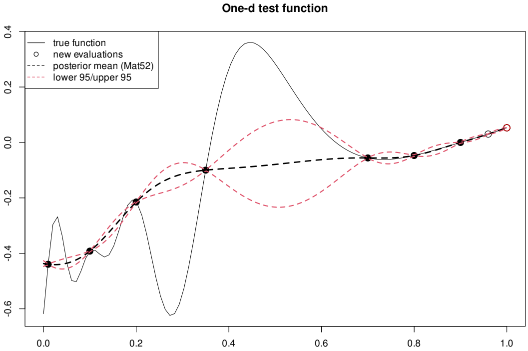
How to find the maximum?

First strategy: $\mathbf{x}_{n+1} \in \arg\max_{\mathbf{x} \in D} m_n(\mathbf{x})$ (maximizer of posterior mean)



Looking for a better strategy

Consider the final GP fit with our previous strategy



We have predictive posterior **mean** m_n and **variance** s_n, *use both?*

Looking for a better strategy

If our next evaluation is

- point with **highest variance:** MSE criterion,
  pure exploration: we will improve the model everywhere

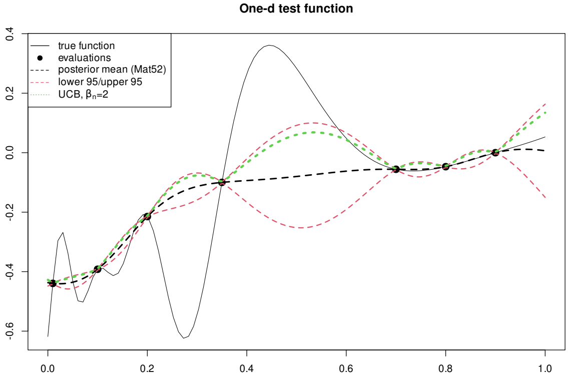
- point with **highest mean:** criterion above,
  pure exploitation: we believe that the current model is good enough

Idea: combine **exploitation** ($\max m_n$) and **exploration** ($\max s_n$)

How can we combine m_n and s_n?

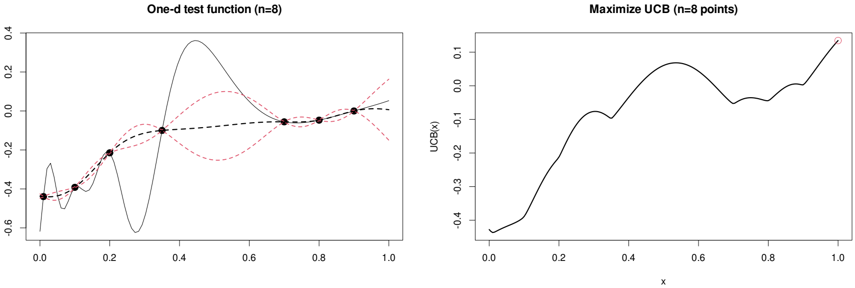
Upper Confidence Bound: $\mathbf{x}_{n+1} \in \arg\max_{\mathbf{x} \in D} UCB_n(\mathbf{x})$,

$$UCB_n(\mathbf{x}) = m_n(\mathbf{x}) + \beta_n s_n(\mathbf{x})$$



Upper Confidence Bound

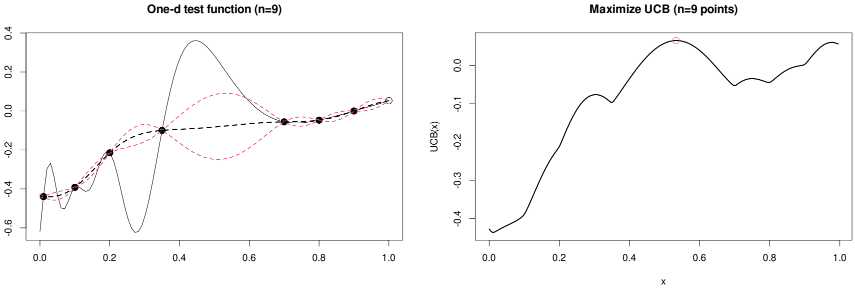
UCB: $\mathbf{x}_{n+1} \in \arg\max_{\mathbf{x} \in D} UCB_n(\mathbf{x})$, $UCB_n(\mathbf{x}) = m_n(\mathbf{x}) + \beta_n s_n(\mathbf{x})$



Here: $\beta_n = 2$, for all n.

Upper Confidence Bound

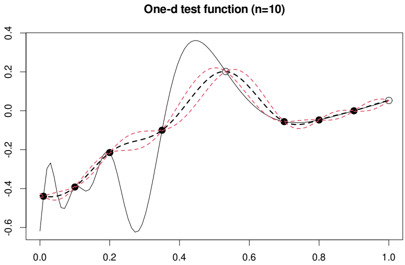
UCB: $\mathbf{x}_{n+1} \in \arg\max_{\mathbf{x}\in D} UCB_n(\mathbf{x})$, $\quad UCB_n(\mathbf{x}) = m_n(\mathbf{x}) + \beta_n s_n(\mathbf{x})$

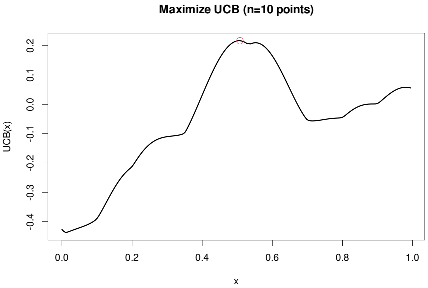


Here: $\beta_n = 2$, for all n.

Upper Confidence Bound

UCB: $\mathbf{x}_{n+1} \in \arg\max_{\mathbf{x}\in D} UCB_n(\mathbf{x})$, $\quad UCB_n(\mathbf{x}) = m_n(\mathbf{x}) + \beta_n s_n(\mathbf{x})$

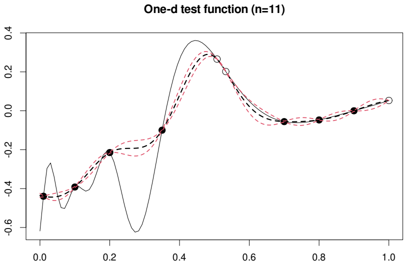


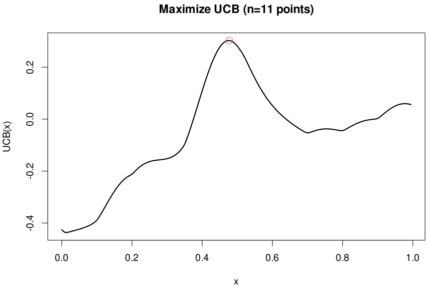


Here: $\beta_n = 2$, for all n.

Upper Confidence Bound

UCB: $\mathbf{x}_{n+1} \in \arg\max_{\mathbf{x} \in D} UCB_n(\mathbf{x})$, $\qquad UCB_n(\mathbf{x}) = m_n(\mathbf{x}) + \beta_n s_n(\mathbf{x})$

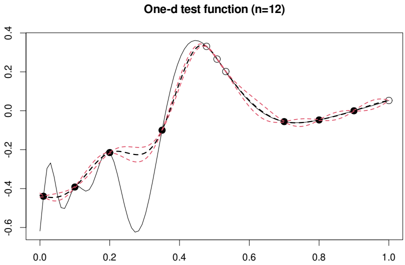


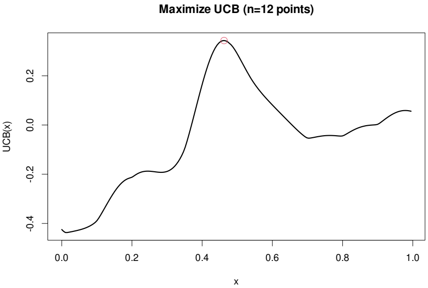


Here: $\beta_n = 2$, for all n.

Upper Confidence Bound

UCB: $\mathbf{x}_{n+1} \in \arg\max_{\mathbf{x}\in D} UCB_n(\mathbf{x})$, $\quad UCB_n(\mathbf{x}) = m_n(\mathbf{x}) + \beta_n s_n(\mathbf{x})$

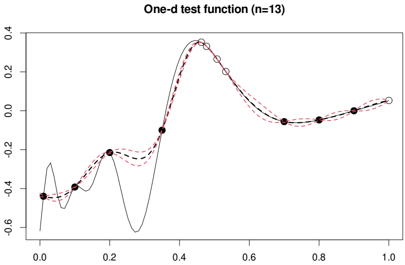


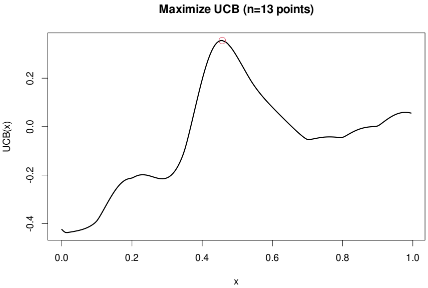


Here: $\beta_n = 2$, for all n.

Upper Confidence Bound

UCB: $\mathbf{x}_{n+1} \in \arg\max_{\mathbf{x}\in D} UCB_n(\mathbf{x})$, $\qquad UCB_n(\mathbf{x}) = m_n(\mathbf{x}) + \beta_n s_n(\mathbf{x})$





Here: $\beta_n = 2$, for all n.

Cumulative regret - a theoretical framework

- $\mathbf{x}^\star \in \arg\max_{\mathbf{x}\in D} f(\mathbf{x})$ (true maximizer of unknown function)
- **regret** at step n, $r_n = f(\mathbf{x}^\star) - f(\mathbf{x}_n)$;
  Note: $f(\mathbf{x}^\star)$ is unknown, so r_n is not computable.
- **cumulative regret:** $R_N = \sum_{n=1}^{N} r_n$

Regret and cumulative regret are tools used to

- build BO algorithms;
- analyze theoretical convergence of BO algorithms.

Upper Confidence Bound acquisition function

Properties

- Easy to implement: only requires m_n, s_n and β_n;
- Acquisition function easy to optimize;
- with variable β_n, provable convergence results
  (Simplest case: f assumed a GP sample on D,
  $\beta_n = 2\log(|D|n^2\pi^2/5\delta)$, γ_n maximum information gain
  then $R_N \leq \sqrt{C_1 N \beta_N \gamma_N}$ with high probability)

Srinivas N., Krause A., Kakade S., and Seeger M. (2010). *Gaussian process optimization in the bandit setting: no regret and experimental design.* ICML'10.

Another point of view: maximizing improvement

Denote by $t_n = \max_{i=1}^n f(\mathbf{x}_i)$, the current maximum of the evaluations.

At step n**:** we evaluate at $\mathbf{x}_{n+1}$, then

- if $f(\mathbf{x}_{n+1}) \geq t_n \Rightarrow f(\mathbf{x}_{n+1})$ is the new maximum;
- if $f(\mathbf{x}_{n+1}) \leq t_n \Rightarrow t_n$ is the maximum

The *improvement* brought by the new evaluation is
$\max(f(\mathbf{x}_{n+1}) - t_n, 0) := (f(\mathbf{x}_{n+1}) - t_n)_+$

Issue: at step n we don't know $f(\mathbf{x}_{n+1})$!

Expected improvement

Idea: use the GP assumption: $Z \sim GP(0, k)$ is a prior for f.

We obtain the random quantities $Z_{\mathbf{x}_{n+1}}$ and $T_n = \max_{i=1}^{n} Z_{\mathbf{x}_i}$ and

Expected improvement: $\mathbb{E}_n[(Z_{\mathbf{x}_{n+1}} - T_n)_+] = \mathbb{E}[(Z_{\mathbf{x}_{n+1}} - T_n)_+ | \mathcal{D}_n]$

T_n is known so we can define the acquisition function

$$EI_n(\mathbf{x}) = \mathbb{E}_n[(Z_{\mathbf{x}} - t_n)_+]$$

Močkus J. (1975) *On bayesian methods for seeking the extremum.* In: Marchuk G.I. (eds) Optimization Techniques IFIP Technical Conference Novosibirsk, July 1–7, 1974.

Jones, D. R., Schonlau, M., Welch, W. J. (1998). *Efficient global optimization of expensive black-box functions.* Journal of Global Optimization, 13(4), 455–492.

Expected improvement acquisition function

Expected improvement: $EI_n(\mathbf{x}) = \mathbb{E}_n[(Z_\mathbf{x} - t_n)_+]$.

There is analytical expression for the conditional expectation

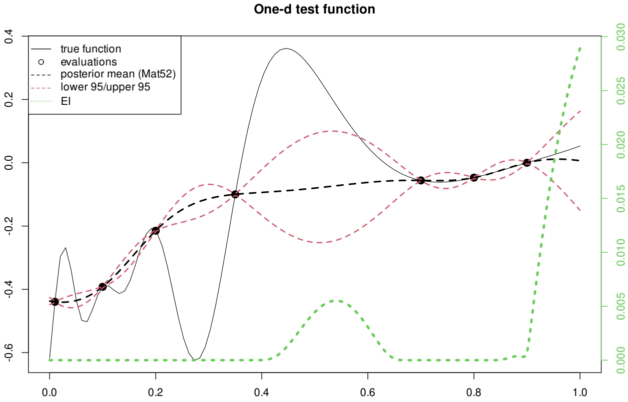
$$EI_n(\mathbf{x}) = \begin{cases} (u_n(\mathbf{x}))_+ & \text{if } s_n(\mathbf{x}) = 0 \\ (u_n(\mathbf{x}))_+ - |u_n(\mathbf{x})|\Phi\left(\frac{u_n(\mathbf{x})}{s_n(\mathbf{x})}\right) + s_n(\mathbf{x})\phi\left(\frac{u_n(\mathbf{x})}{s_n(\mathbf{x})}\right) & \text{else.} \end{cases}$$

Where $u_n(\mathbf{x}) = m_n(\mathbf{x}) - t_n$, $\Phi(\cdot)$ CDF of $N(0,1)$, $\phi(\cdot)$ PDF of $N(0,1)$.

Key take-away: EI one-step ahead can be computed with just m_n, s_n, t_n.

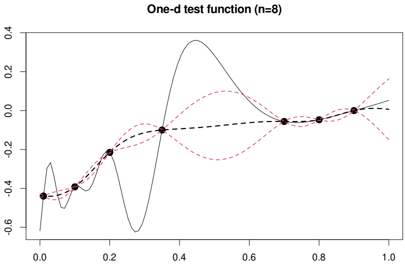
Expected improvement

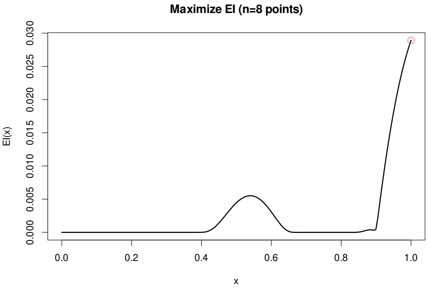
Consider our 1-d example, what is $EI(\mathbf{x})$?



Expected improvement

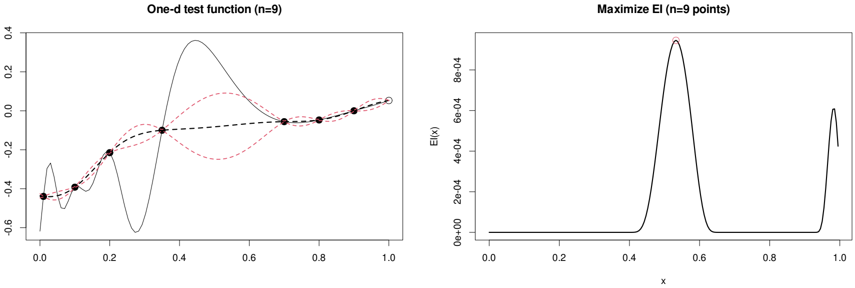
EI: $\mathbf{x}_{n+1} \in \arg\max_{\mathbf{x}\in D} EI_n(\mathbf{x}), \qquad EI_n(\mathbf{x}) = \mathbb{E}_n[(Z_{\mathbf{x}} - t_n)_+]$





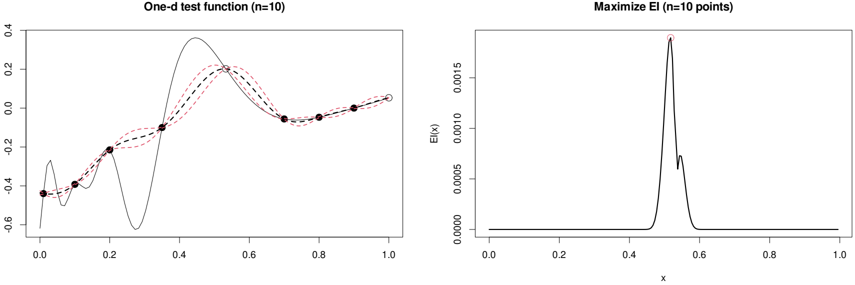
Expected improvement

EI: $\mathbf{x}_{n+1} \in \arg\max_{\mathbf{x}\in D} EI_n(\mathbf{x}), \qquad EI_n(\mathbf{x}) = \mathbb{E}_n[(Z_{\mathbf{x}} - t_n)_+]$



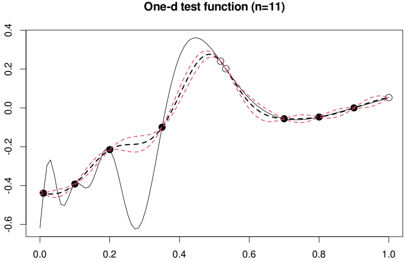
Expected improvement

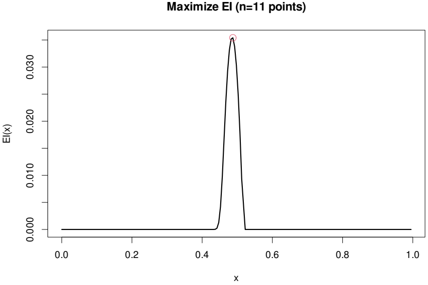
EI: $\mathbf{x}_{n+1} \in \arg\max_{\mathbf{x}\in D} EI_n(\mathbf{x})$, $\quad EI_n(\mathbf{x}) = \mathbb{E}_n[(Z_{\mathbf{x}} - t_n)_+]$



Expected improvement

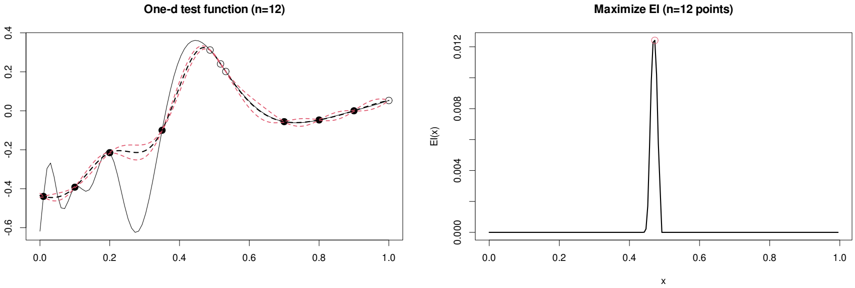
EI: $\mathbf{x}_{n+1} \in \arg\max_{\mathbf{x}\in D} EI_n(\mathbf{x})$, $\qquad EI_n(\mathbf{x}) = \mathbb{E}_n[(Z_{\mathbf{x}} - t_n)_+]$





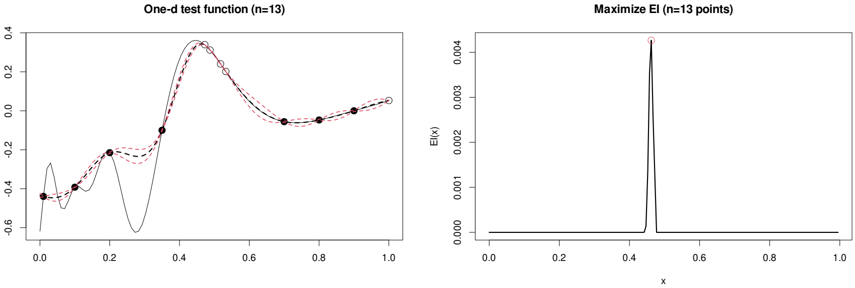
Expected improvement

EI: $\mathbf{x}_{n+1} \in \arg\max_{\mathbf{x}\in D} EI_n(\mathbf{x})$, $\quad EI_n(\mathbf{x}) = \mathbb{E}_n[(Z_{\mathbf{x}} - t_n)_+]$



Expected improvement

EI: $\mathbf{x}_{n+1} \in \arg\max_{\mathbf{x}\in D} EI_n(\mathbf{x}), \qquad EI_n(\mathbf{x}) = \mathbb{E}_n[(Z_\mathbf{x} - t_n)_+]$



Expected improvement - Properties

The acquisition function EI_n

- non-negative, vanishes at evaluations;
- generally not convex/concave and highly multi-modal;
- regularity driven by k_n;
- under technical condition on k and f, EI fills the space;
- convergence for EI recently proven in Bect et al. (2019).

Bect J., Bachoc F., Ginsbourger D. (2019) *A supermartingale approach to Gaussian process based sequential design of experiments.* Bernoulli 25 (4A) 2883 - 2919.

Expected improvement - Limits and extensions

Only one step ahead? $EI_n(\mathbf{x})$ provides the **next** evaluation $\mathbf{x}_{n+1}$.

Can we have a **batch** of q evaluations at each step?

Parallel (multi-point) EI: analytical formulae for multi-point EI (See Ginsbourger et al. (2010), Chevalier, Ginsbourger (2013), Marmin et al. (2015))

Noisy evaluations: up until now observations $y_i = f(\mathbf{x}_i) \Rightarrow$ no noise.

What if $y_i = f(\mathbf{x}_i) + \epsilon_i$?

EI and noisy observations

Data: $y_i = f(\mathbf{x}_i) + \epsilon_i$ with $\epsilon_i \sim N(0, \sigma_n^2)$, $\sigma_n^2 > 0$.

Recall:

$$EI_n(\mathbf{x}) = \mathbb{E}_n[(\overbrace{Z_{\mathbf{x}}}^{\text{not reachable}} - \overbrace{\max_{i=1}^{n} Z_{\mathbf{x}_i}}^{\text{not known}})_+]$$

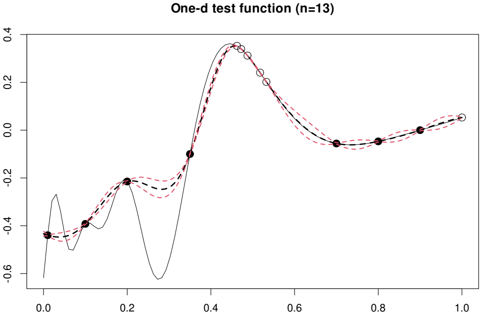
Possible fixes:

- **plug-in:** EI with highest expected value of the observed points.
- **Expected Quantile Improvement:** measure improvement on quantiles. "Best" point at highest quantile. See Picheny et al. (2013).

Acquisition function that handles well noise is **knowledge gradient**.

Knowledge gradient - a different point of view

EI: “best point” $=$ highest **evaluated** point;



Another option: return a non evaluated point

Example: $\hat{\mathbf{x}} = \arg\max_{\mathbf{x}\in D} m_n(\mathbf{x})$

Knowledge gradient - a different "improvement":

"Best" point determined as $\hat{\mathbf{x}} = \arg\max_{\mathbf{x} \in D} m_n(\mathbf{x})$

at step n "best value" $m_n^\star = m_n(\hat{\mathbf{x}})$

at step $n+1$ $m_{n+1}^\star = \max_{\mathbf{x} \in D} m_{n+1}(\mathbf{x})$

(random and unknown until $\mathbf{x}_{n+1}$ given)

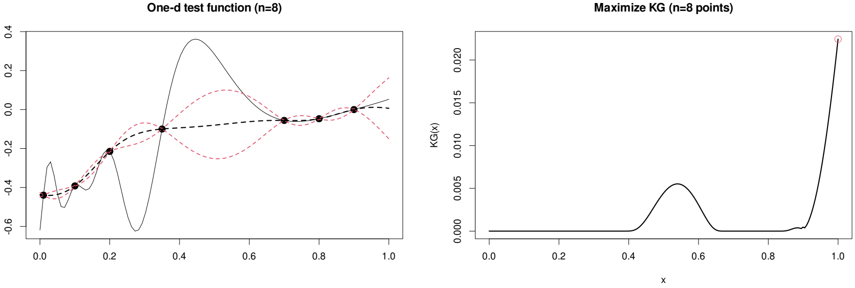
We would like $\mathbf{x}_{n+1}$ that maximizes $m_{n+1}^\star - m_n^\star$

Knowledge gradient: $KG(\mathbf{x}) = \mathbb{E}_n[m_{n+1}^\star - m_n^\star | \mathbf{x}_{n+1} = \mathbf{x}]$

Frazier P., Powell W., Dayanik S., (2009) *The Knowledge-Gradient Policy for Correlated Normal Beliefs.* INFORMS Journal on Computing 21(4):599-613.

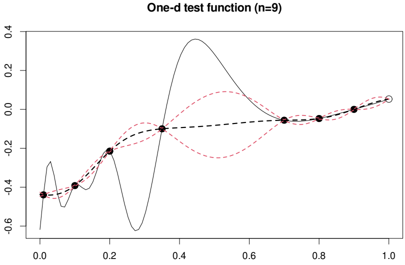
Knowledge gradient

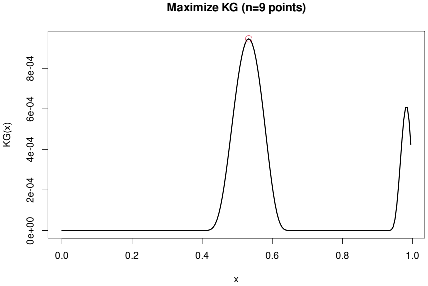
KG: $\mathbf{x}_{n+1} \in \arg\max_{\mathbf{x}\in D} KG_n(\mathbf{x}), \qquad KG_n(\mathbf{x}) = \mathbb{E}_n[m^\star_{n+1} - m^\star_n | \mathbf{x}_{n+1} = \mathbf{x}]$



Knowledge gradient

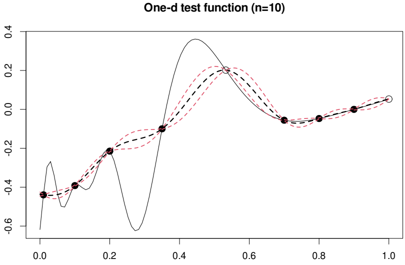
KG: $\mathbf{x}_{n+1} \in \arg\max_{\mathbf{x} \in D} KG_n(\mathbf{x})$, $\quad KG_n(\mathbf{x}) = \mathbb{E}_n[m^\star_{n+1} - m^\star_n | \mathbf{x}_{n+1} = \mathbf{x}]$

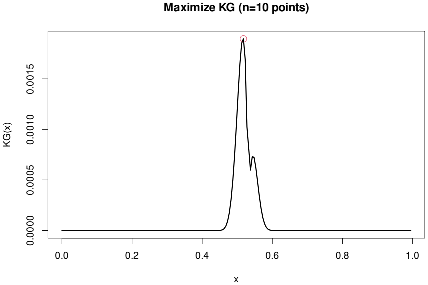




Knowledge gradient

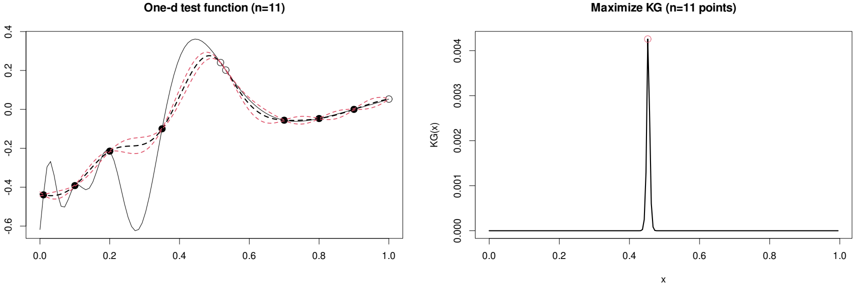
KG: $\mathbf{x}_{n+1} \in \arg\max_{\mathbf{x}\in D} KG_n(\mathbf{x})$, $\quad KG_n(\mathbf{x}) = \mathbb{E}_n[m^\star_{n+1} - m^\star_n | \mathbf{x}_{n+1} = \mathbf{x}]$





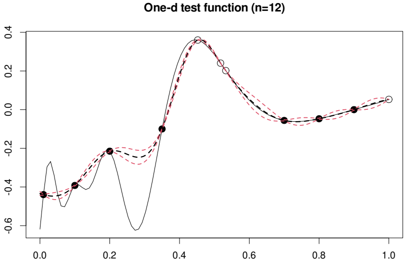
Knowledge gradient

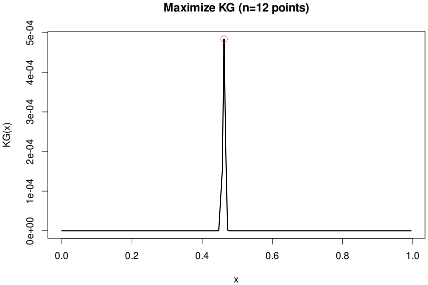
KG: $\mathbf{x}_{n+1} \in \arg\max_{\mathbf{x}\in D} KG_n(\mathbf{x}), \qquad KG_n(\mathbf{x}) = \mathbb{E}_n[m^\star_{n+1} - m^\star_n | \mathbf{x}_{n+1} = \mathbf{x}]$



Knowledge gradient

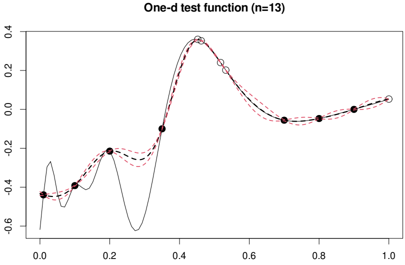
KG: $\mathbf{x}_{n+1} \in \arg\max_{\mathbf{x}\in D} KG_n(\mathbf{x}),\qquad KG_n(\mathbf{x}) = \mathbb{E}_n[m^\star_{n+1} - m^\star_n | \mathbf{x}_{n+1} = \mathbf{x}]$

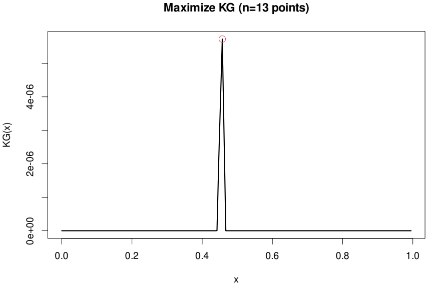




Knowledge gradient

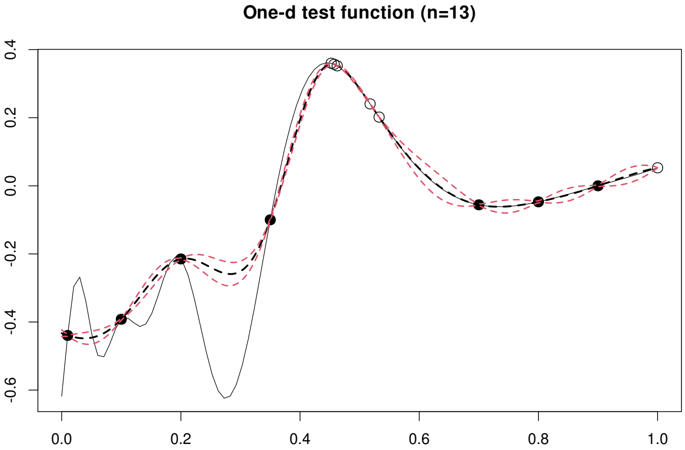
KG: $\mathbf{x}_{n+1} \in \arg\max_{\mathbf{x}\in D} KG_n(\mathbf{x})$, $\quad KG_n(\mathbf{x}) = \mathbb{E}_n[m^\star_{n+1} - m^\star_n | \mathbf{x}_{n+1} = \mathbf{x}]$





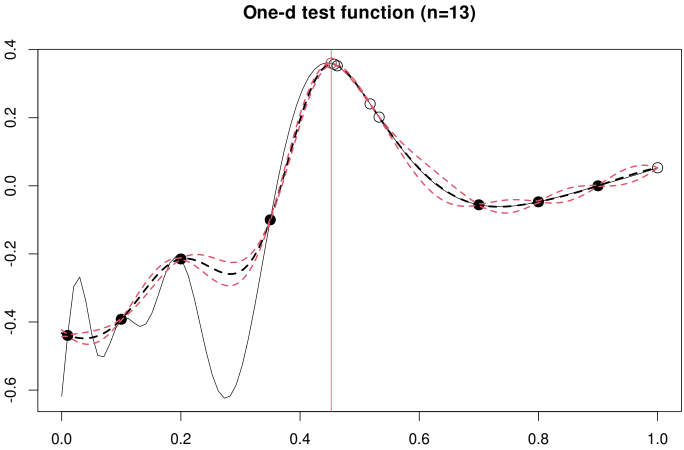
Knowledge gradient - final iteration

Where is our maximum?



Knowledge gradient - final iteration

Where is our maximum? **Recall:** best point is $\arg\max_{\mathbf{x}\in D} m_n(\mathbf{x})$



Knowledge gradient - further details

Computation: $KG_n(\mathbf{x})$ can be computed

- via simulation (Frazier et al. 2009)
- multi-start stochastic gradient ascent (more efficient, see Wu and Frazier (2016))
- parallelizable (see Wu and Frazier (2016))

Noisy evaluations

- *KG* does not use directly function evaluations
- on noisy functions showed good performances

Wu, J. and Frazier, P. (2016). *The parallel knowledge gradient method for batch Bayesian optimization.* In Advances in Neural Information Processing Systems.

Quick and partial summary on BO

	Pros	Cons
UCB	+ fast to implement	- no one-step optimality
	+ convergence guarantees	
EI	+ one-step optimal	- not adapted to noisy data
	+ convergence guarantees	
KG	+ one-step optimal	- no convergence guarantee
	+ handle noisy data	

Further topics not covered here

Other notable strategies: Thomson sampling, Entropy sampling, Predictive entropy sampling, ...

Constrained optimization

Multi-fidelity optimization

Non-myopic strategies

High-dimensional BO

Gray-box (using derivative information on f)

A very nice tutorial

Frazier P. (2018). *A Tutorial on Bayesian Optimization.* arXiv:1807.02811 [stat.ML]

Partial list of software

Python: BOtorch (pytorch), GPflowOpt (tensorflow), Cornell-MOE, GPyOpt(end of life),Spearmint (no longer actively maintained)

R: DiceOptim, laGP

Matlab: DACE (no longer actively maintained)

References (1)

Srinivas N., Krause A., Kakade S., and Seeger M. (2010). *Gaussian process optimization in the bandit setting: no regret and experimental design.* ICML'10.

Močkus J. (1975) *On bayesian methods for seeking the extremum.* In: Marchuk G.I. (eds) Optimization Techniques IFIP Technical Conference Novosibirsk, July 1–7, 1974.

Jones, D. R., Schonlau, M., Welch, W. J. (1998). *Efficient global optimization of expensive black-box functions.* Journal of Global Optimization, 13(4), 455–492.

Ginsbourger D., Le Riche R., Carraro L. (2010) *Kriging Is Well-Suited to Parallelize Optimization.* In: Computational Intelligence in Expensive Optimization Problems. Adaptation Learning and Optimization, vol 2.

Chevalier C., Ginsbourger D. (2013) *Fast Computation of the Multi-Points Expected Improvement with Applications in Batch Selection.* In: Learning and Intelligent Optimization. LION 2013.

Bect J., Bachoc F., Ginsbourger D. (2019) *A supermartingale approach to Gaussian process based sequential design of experiments.* Bernoulli 25 (4A) 2883 - 2919.

References (2)

Marmin S., Chevalier C., Ginsbourger D. (2015) *Differentiating the Multipoint Expected Improvement for Optimal Batch Design.* In: Machine Learning, Optimization, and Big Data. MOD 2015.

Picheny V., Ginsbourger D., Richet Y., Caplin G. (2013) *Quantile-Based Optimization of Noisy Computer Experiments With Tunable Precision.* Technometrics, 55:1, 2-13

Frazier P., Powell W., Dayanik S., (2009) *The Knowledge-Gradient Policy for Correlated Normal Beliefs.* INFORMS Journal on Computing 21(4):599-613.

Wu, J. and Frazier, P. (2016). *The parallel knowledge gradient method for batch Bayesian optimization.* In Advances in Neural Information Processing Systems.

Outline

The framework

Consider now $\Gamma^\star = \{\mathbf{x} \in D : f(\mathbf{x}) \in T\} = f^{-1}(T)$
where $D \subset \mathbb{R}^d$ is compact, $f : D \longrightarrow \mathbb{R}$ is continuous, $T \subset \mathbb{R}$.

Particular case: $T = (-\infty, t]$ for a fixed $t \in \mathbb{R}$.
$\Gamma^\star = \{\mathbf{x} \in D : f(\mathbf{x}) \leq t\}$ is the excursion set of f below t.

Objectives

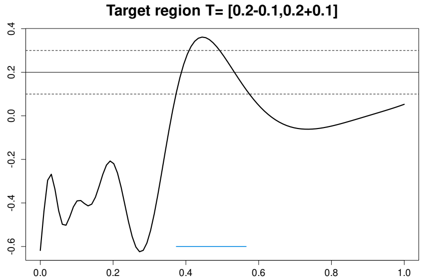
Starting from evaluations of f at $\mathbf{X}_n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset D$

- Obtain a good regression model for values of f in $[t - \epsilon, t + \epsilon]$
- Estimate the volume of $\Gamma^\star$, $\alpha_\mu = \mu(\Gamma^\star)$ and reduce its uncertainty.
- Estimate $\Gamma^\star$ and quantify uncertainty on it

Example

Consider the function f and

$$\Gamma^\star = \{\mathbf{x} \in D : f(\mathbf{x}) \in T\} = f^{-1}(T)$$



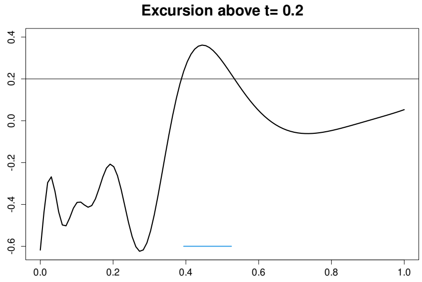
- target region: $T = [t - \epsilon, t + \epsilon]$, $t = 0.2$, $\epsilon = 0.1$;
- excursion region

  $T = [t, +\infty)$, $t = 0.2$.

Example

Consider the function f and

$$\Gamma^\star = \{\mathbf{x} \in D : f(\mathbf{x}) \in T\} = f^{-1}(T)$$



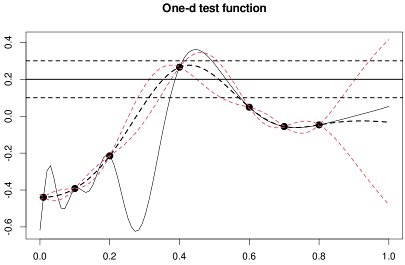
- target region: $T = [t - \epsilon, t + \epsilon]$, $t = 0.2$, $\epsilon = 0.1$;
- excursion region

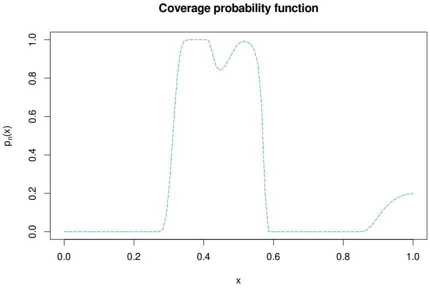
  $T = [t, +\infty)$, $t = 0.2$.

The coverage probability function - target region case

GP model $(Z_{\mathbf{x}})_{\mathbf{x}\in D} \sim GP(m, k)$ and n evaluations $\mathbf{y}_n$ at points $\mathbf{X}_n$.
The **(posterior) coverage probability function** is

$$p_{n,t,\epsilon}(\mathbf{x}) = P(\mathbf{x} \in \Gamma | Z_{\mathbf{X}_n} = \mathbf{y}_n) = \Phi\left(\frac{t+\epsilon-m_n(\mathbf{x})}{\sqrt{k_n(\mathbf{x},\mathbf{x})}}\right) - \Phi\left(\frac{t-\epsilon-m_n(\mathbf{x})}{\sqrt{k_n(\mathbf{x},\mathbf{x})}}\right)$$

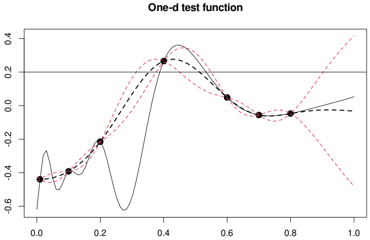


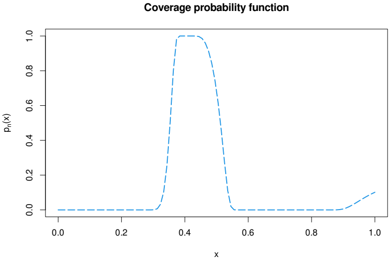


The coverage probability function - excursion case

GP model $(Z_{\mathbf{x}})_{\mathbf{x}\in D} \sim GP(m,k)$ and n evaluations $\mathbf{y}_n$ at points $\mathbf{X}_n$.
The **(posterior) coverage probability function** is

$$p_n(\mathbf{x}) = P(\mathbf{x} \in \Gamma | Z_{\mathbf{X}_n} = \mathbf{y}_n) = \Phi\left(\frac{m_n(\mathbf{x}) - t}{\sqrt{k_n(\mathbf{x},\mathbf{x})}}\right)$$





Stepwise Uncertainty Reduction (SUR) framework

Start: initial design $\mathbf{X}_n$, $\mathbf{y}_n = f(\mathbf{X}_n)$, prior model $(Z_\mathbf{x})_{\mathbf{x} \in D} \sim GP(m, k)$.

1. First model for f from posterior mean of $Z_\mathbf{x}$ given $Z_{\mathbf{X}_n} = \mathbf{y}_n$
2. define uncertainty measure H_n
3. next evaluation $\mathbf{x}_{n+1}$ as the minimizer of expected future uncertainty ($\mathbf{x}_{n+1} \in \arg\min_\mathbf{x} \mathbb{E}_n[H_{n+1} \mid \mathbf{x}_{n+1} = \mathbf{x}]$)
4. update the model
5. repeat steps 2-4 until convergence or until computational budget is exhausted.

Target region exploration - targeted IMSE

Define the targeted IMSE with target t and tolerance ϵ as

$$H_n^{tIMSE} = \mathbb{E}_n[\int_D MSE_n(\mathbf{u})\mathbf{1}_{[t-\epsilon,t+\epsilon]}(Z_{\mathbf{u}})\mu(d\mathbf{u})]$$
$$= \int_D k_n(\mathbf{u},\mathbf{u})p_{n,t,\epsilon}(\mathbf{u})\mu(d\mathbf{u})$$

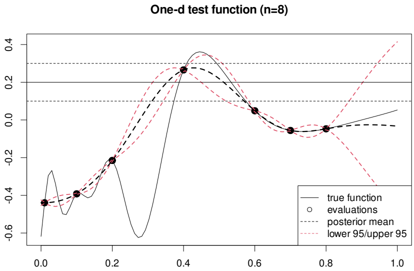
The corresponding criterion is

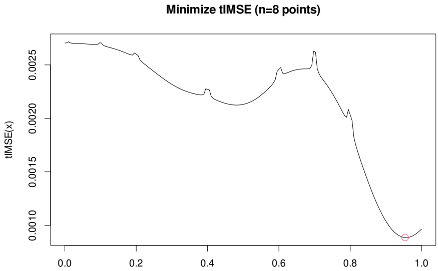
$$J_n^{tIMSE}(\mathbf{x}) = \mathbb{E}_n[\int_D k_{n+1,\mathbf{x}_{n+1}}(\mathbf{u},\mathbf{u})p_{n,t,\epsilon}(\mathbf{u})\mu(d\mathbf{u}) \mid \mathbf{x}_{n+1} = \mathbf{x}]$$

Picheny V., Ginsbourger D., Roustant O., Haftka R.T., Kim N. (2010). *Adaptive Designs of Experiments for Accurate Approximation of a Target Region.* ASME. J. Mech. Des. 132(7):071008-071008-9.

targeted IMSE - Example

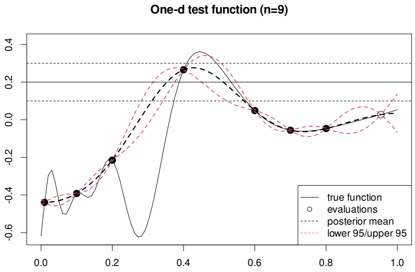
Initial design $n = 7$, add 10 points.

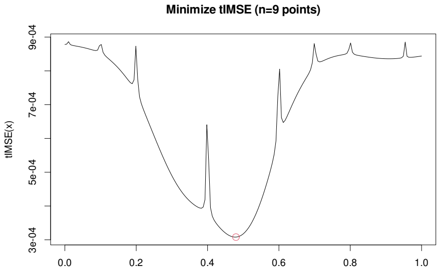




targeted IMSE - Example

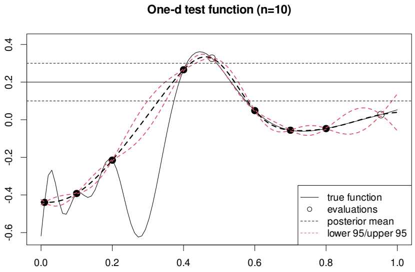
Initial design $n = 7$, add 10 points.

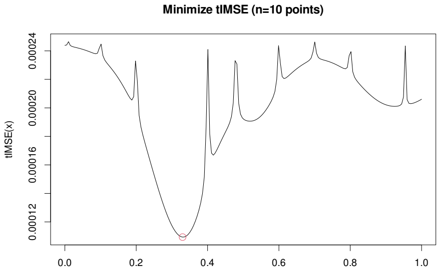




targeted IMSE - Example

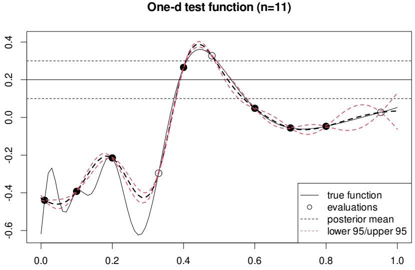
Initial design $n = 7$, add 10 points.

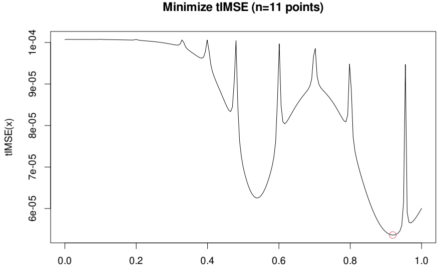




targeted IMSE - Example

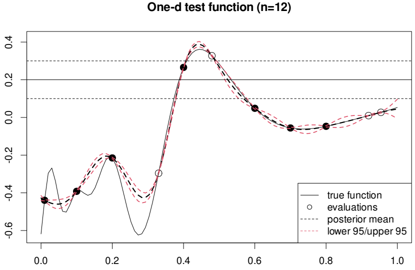
Initial design $n = 7$, add 10 points.

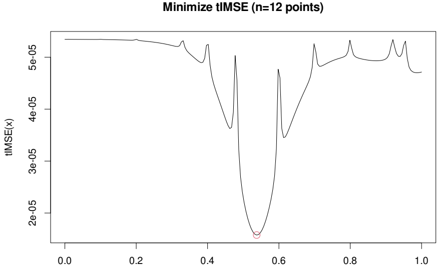




targeted IMSE - Example

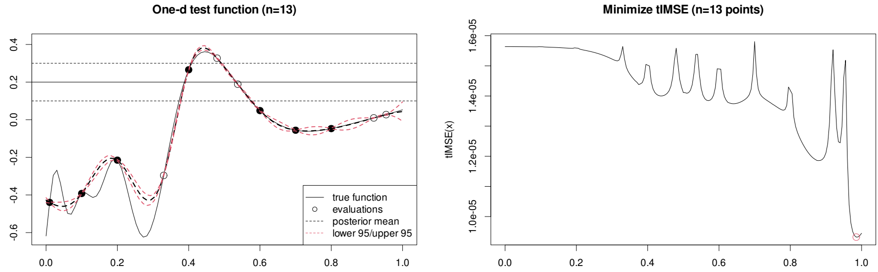
Initial design $n = 7$, add 10 points.





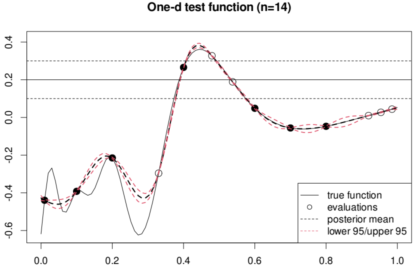
targeted IMSE - Example

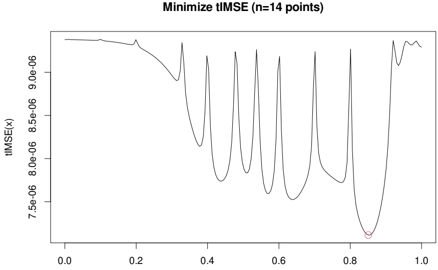
Initial design $n = 7$, add 10 points.



targeted IMSE - Example

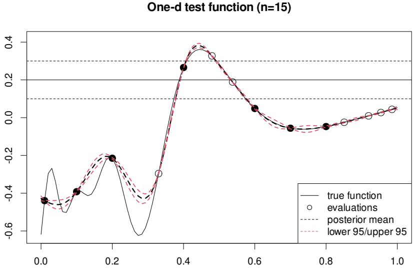
Initial design $n = 7$, add 10 points.

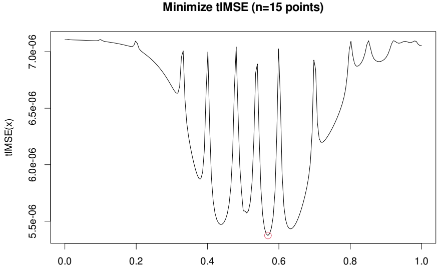




targeted IMSE - Example

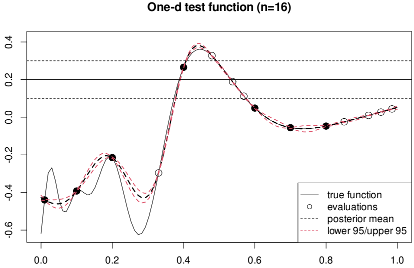
Initial design $n = 7$, add 10 points.

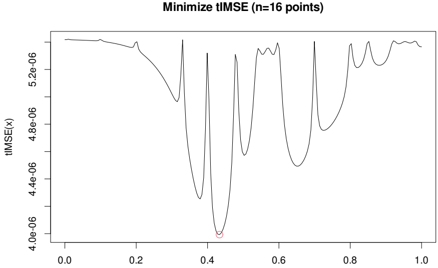




targeted IMSE - Example

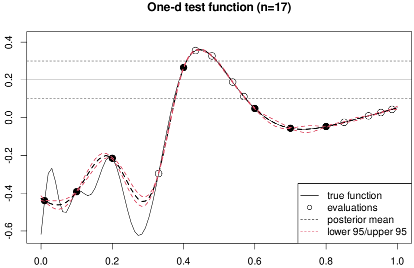
Initial design $n = 7$, add 10 points.

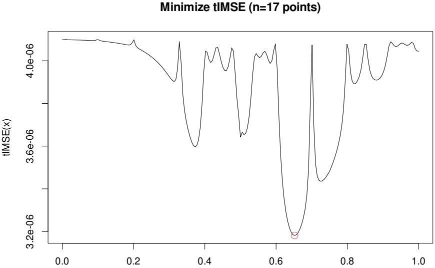




targeted IMSE - Example

Initial design $n = 7$, add 10 points.





The measure of excursion as proxy

Consider now the problem of estimating the set

$$\Gamma^\star = \{\mathbf{x} \in D : f(\mathbf{x}) \geq t\}.$$

Bayesian approach: f is a realization of a GP $(Z_\mathbf{x})_{\mathbf{x} \in D}$, then $\Gamma^\star$ is a realization of $\Gamma = \{\mathbf{x} \in D : Z_\mathbf{x} \geq t\}$

Interesting feature: measure of the set

Finite measure μ on D, a proxy for the set is $\alpha = \mu(\Gamma^\star)$,
For Γ (random), estimate it with

$$\widehat{\alpha} = \mathbb{E}_n[\mu(\Gamma)] = \int_D p_n(\mathbf{u})\mu(d\mathbf{u})$$

Uncertainty: variance of excursion volume

Uncertainty on $\widehat{\alpha} = \mathbb{E}_n[\mu(\Gamma)]$ can be measured with $\mathrm{var}_n[\mu(\Gamma)]$

$$\mathrm{var}_n[\mu(\Gamma)] = \mathbb{E}_n[\left(\int_D \mathbf{1}_\Gamma(\mathbf{u}) - p_n(\mathbf{u})\mu(d\mathbf{u})\right)^2]$$

$$\leq \mu(D)\mathbb{E}_n[\int_D (\mathbf{1}_\Gamma(\mathbf{u}) - p_n(\mathbf{u}))^2 \mu(d\mathbf{u})]$$

$$\leq \mu(D)\int_D p_n(\mathbf{u})(1 - p_n(\mathbf{u}))\mu(d\mathbf{u})$$

Uncertainty measures

$H_n^\Gamma = \int_D p_n(\mathbf{u})(1 - p_n(\mathbf{u}))\mu(d\mathbf{u})$, integrated Bernoulli variance;

$H_n^{(\alpha)} = \mathrm{var}_n[\mu(\Gamma)]$, variance of excursion volume.

Criterion: integrated Bernoulli variance

The integrated Bernoulli variance uncertainty naturally provides a criterion for a sequential design

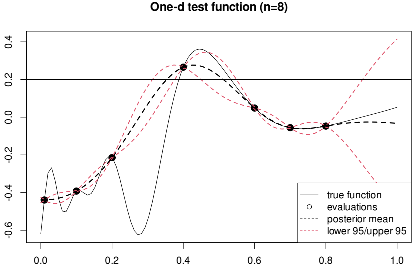
$$
\begin{aligned}
J_n^{\Gamma}(\mathbf{x}) &= \mathbb{E}_n[H_{n+1}^{\Gamma} | \mathbf{x}_{n+1} = \mathbf{x}] \\
&= \mathbb{E}_n\left[\int_D p_{n+1,\mathbf{x}_{n+1}}(\mathbf{u})(1 - p_{n+1,\mathbf{x}_{n+1}}(\mathbf{u}))\mu(d\mathbf{u}) \mid \mathbf{x}_{n+1} = \mathbf{x}\right] \\
&= \int_D \Phi_2\left(\begin{pmatrix} a(\mathbf{u}) \\ -a(\mathbf{u}) \end{pmatrix}; \begin{pmatrix} c(\mathbf{u}) & 1 - c(\mathbf{u}) \\ 1 - c(\mathbf{u}) & c(\mathbf{u}) \end{pmatrix}\right)\mu(d\mathbf{u})
\end{aligned}
$$

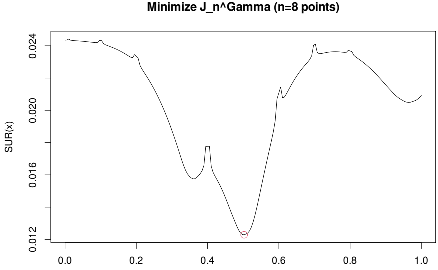
where $a(\mathbf{u}) = \frac{m_n(\mathbf{u}) - t}{\sqrt{k_{n+1}(\mathbf{u})}}$, $c(\mathbf{u}) = k_n(\mathbf{u})/k_{n+1}(\mathbf{u})$.

Bect, J., Ginsbourger, D., Li, L., Picheny, V., and Vazquez, E. (2012). *Sequential design of computer experiments for the estimation of a probability of failure.* Stat. Comput., 22 (3):773–793.

Integrated Bernoulli variance

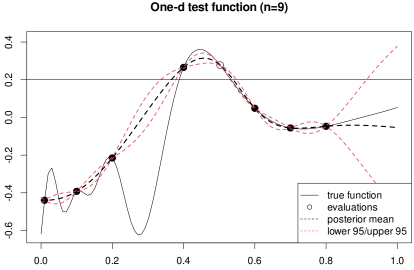
Initial design $n = 7$, add 10 points.

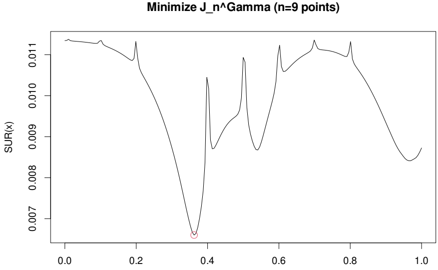




Integrated Bernoulli variance

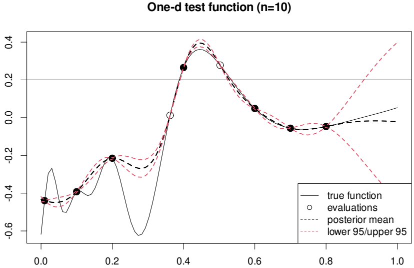
Initial design $n = 7$, add 10 points.

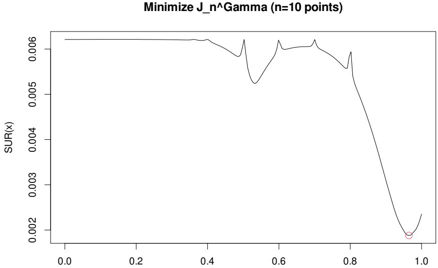




Integrated Bernoulli variance

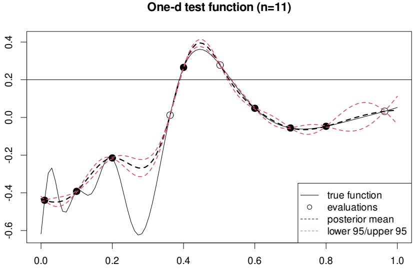
Initial design $n = 7$, add 10 points.

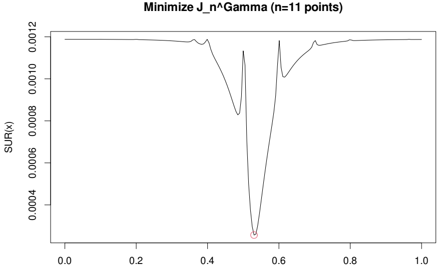




Integrated Bernoulli variance

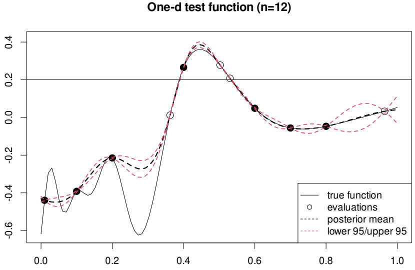
Initial design $n = 7$, add 10 points.

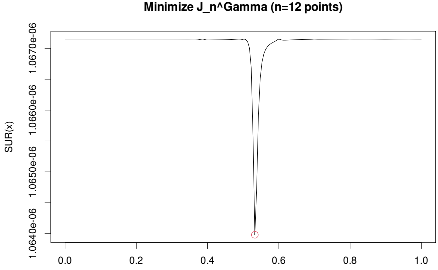




Integrated Bernoulli variance

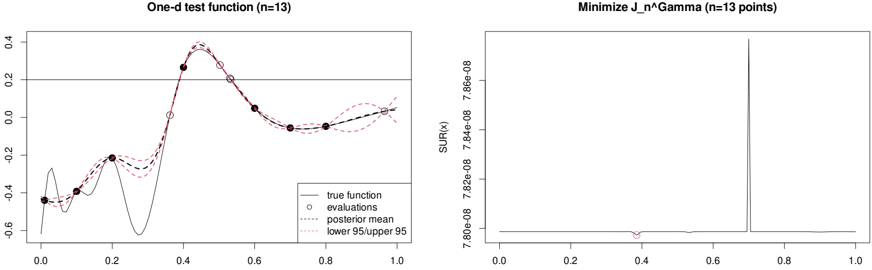
Initial design $n = 7$, add 10 points.





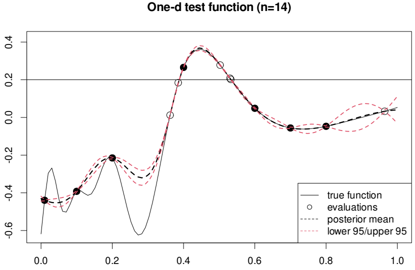
Integrated Bernoulli variance

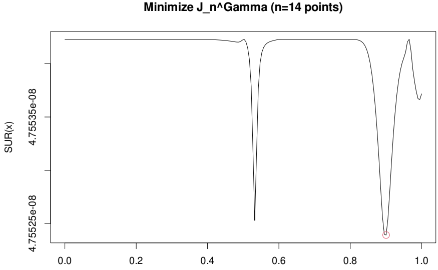
Initial design $n = 7$, add 10 points.



Integrated Bernoulli variance

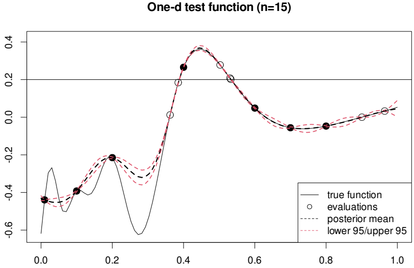
Initial design $n = 7$, add 10 points.

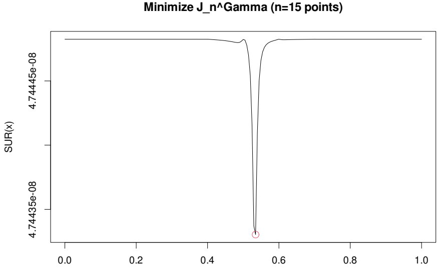




Integrated Bernoulli variance

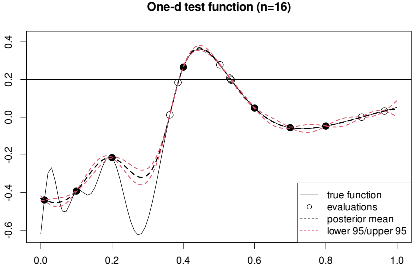
Initial design $n = 7$, add 10 points.

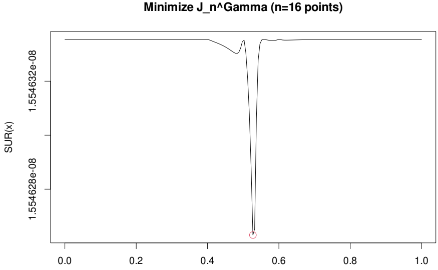




Integrated Bernoulli variance

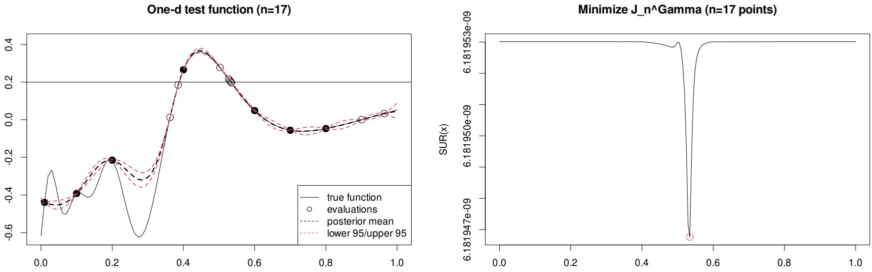
Initial design $n = 7$, add 10 points.





Integrated Bernoulli variance

Initial design $n = 7$, add 10 points.



Criterion: variance of excursion volume

The variance of the excursion volume also defines a criterion

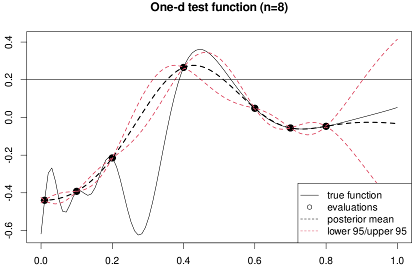
$$
\begin{aligned}
J_n^{(\alpha)}(\mathbf{x}) &= \mathbb{E}_n[H_{n+1}^{(\alpha)} | \mathbf{x}_{n+1} = \mathbf{x}] \\
&= \mathbb{E}_n\left[\mathrm{var}_{n+1}[\mu(\Gamma)] \mid \mathbf{x}_{n+1} = \mathbf{x}\right] \\
&= \gamma_n - \int_{D\times D} \Phi_2\left(\begin{pmatrix} a(\mathbf{u}_1) \\ a(\mathbf{u}_2) \end{pmatrix}; \begin{pmatrix} c(\mathbf{u}_1) & d(\mathbf{u}_1, \mathbf{u}_2) \\ d(\mathbf{u}_1, \mathbf{u}_2) & c(\mathbf{u}_2) \end{pmatrix}\right) \mu(d\mathbf{u}_1)\mu(d\mathbf{u}_2)
\end{aligned}
$$

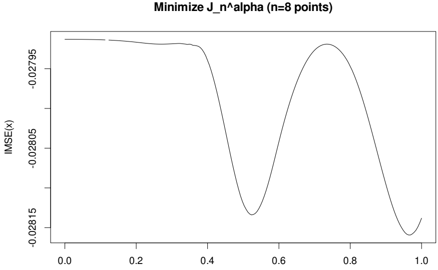
where $a(\mathbf{u}) = \frac{m_n(\mathbf{u}) - t}{\sqrt{k_{n+1}(\mathbf{u})}}$, $c(\mathbf{u}) = k_n(\mathbf{u})/k_{n+1}(\mathbf{u})$,
$d(\mathbf{u}_1, \mathbf{u}_2) = k_n(\mathbf{u}_1, \mathbf{x}_{n+1}) k_n(\mathbf{u}_2, \mathbf{x}_{n+1}) / k_n(\mathbf{x}_{n+1}, \mathbf{x}_{n+1})$

Chevalier, C., Bect, J., Ginsbourger, D., Vazquez, E., Picheny, V. and Richet, Y. (2013). *Fast parallel kriging-based stepwise uncertainty reduction with application to the identification of an excursion set.* Technometrics, 56 (4):455–465.

Variance of excursion volume

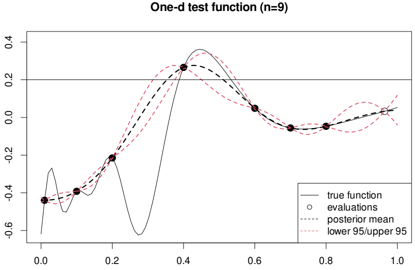
Initial design $n = 7$, add 5 points.

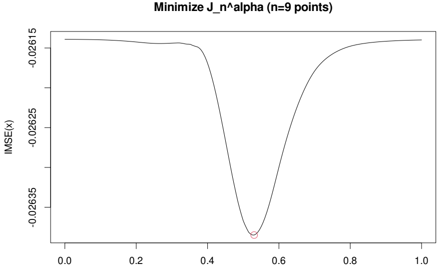




Variance of excursion volume

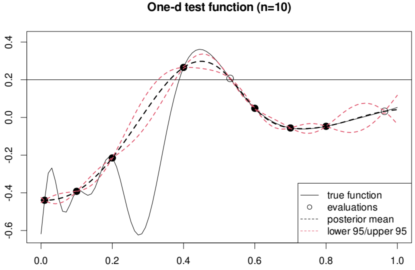
Initial design $n = 7$, add 5 points.

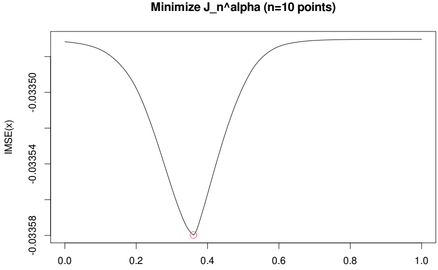




Variance of excursion volume

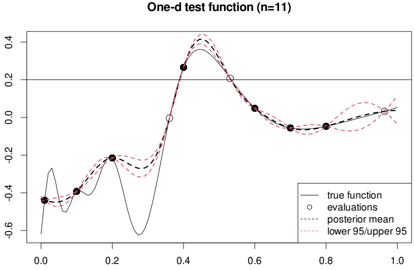
Initial design $n = 7$, add 5 points.

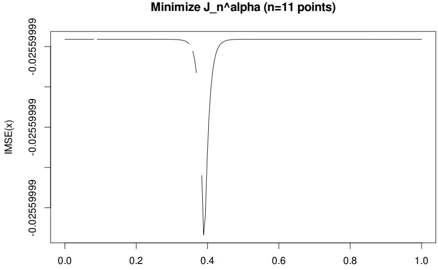




Variance of excursion volume

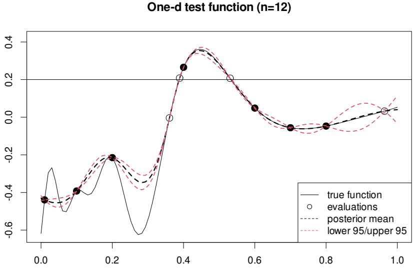
Initial design $n = 7$, add 5 points.

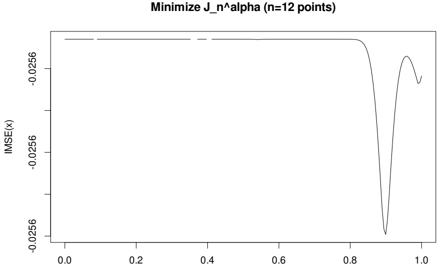




Variance of excursion volume

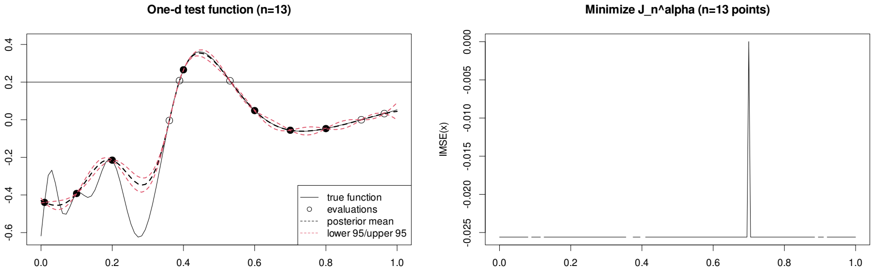
Initial design $n = 7$, add 5 points.





Variance of excursion volume

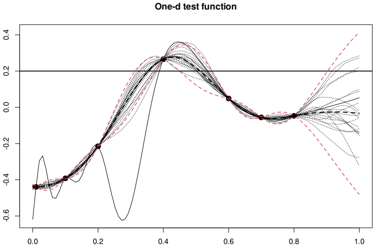
Initial design $n = 7$, add 5 points.



Outline

A prior on the space of functions

Assume: f realization of $(Z_{\mathbf{x}})_{\mathbf{x}\in D}$, Gaussian Random Field (GRF)



Prior: $(Z_{\mathbf{x}})_{\mathbf{x}\in D}$ with

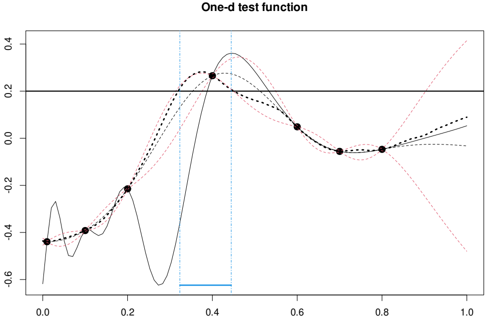
- a.s. continuous paths;
- Matérn covariance kernel k ($\nu = 5/2$);
- constant mean function m.

Given $n = 7$ evaluations $\mathbf{z}_n$ at $\mathbf{X}_n$

We can generate realizations of the **posterior field**.

Distribution of excursion sets

The posterior field induces a posterior distribution on excursion sets.



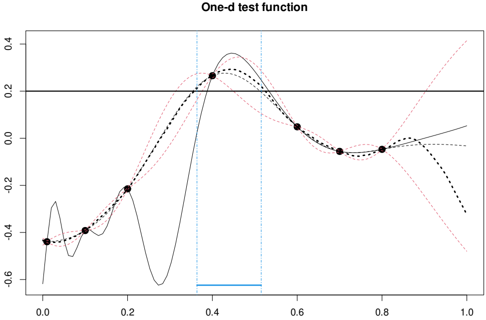
Note that

- Z continuous paths;
- $[t, +\infty)$ is a closed set.

The set $\Gamma = \{\mathbf{x} \in D : Z_{\mathbf{x}} \geq t\}$ is a random closed set.

Distribution of excursion sets

The posterior field induces a posterior distribution on excursion sets.



Note that

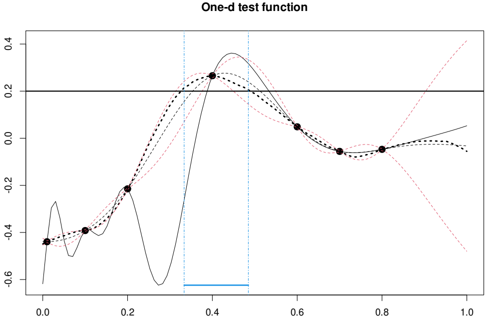
Z continuous paths;

$[t, +\infty)$ is a closed set.

The set $\Gamma = \{\mathbf{x} \in D : Z_{\mathbf{x}} \geq t\}$ is a random closed set.

Distribution of excursion sets

The posterior field induces a posterior distribution on excursion sets.



Note that

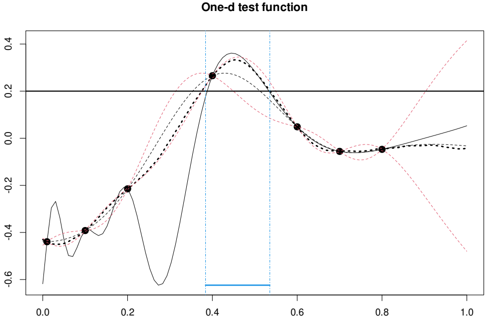
Z continuous paths;

$[t, +\infty)$ is a closed set.

The set $\Gamma = \{\mathbf{x} \in D : Z_{\mathbf{x}} \geq t\}$ is a random closed set.

Distribution of excursion sets

The posterior field induces a posterior distribution on excursion sets.



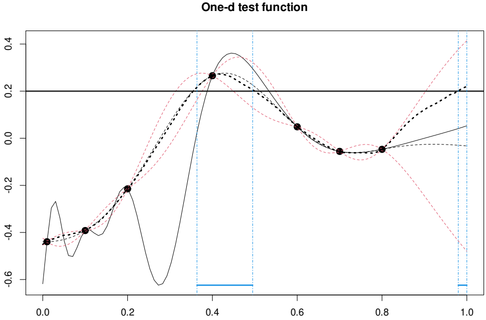
Note that

- Z continuous paths;
- $[t, +\infty)$ is a closed set.

The set $\Gamma = \{\mathbf{x} \in D : Z_{\mathbf{x}} \geq t\}$ is a random closed set.

Distribution of excursion sets

The posterior field induces a posterior distribution on excursion sets.



Note that

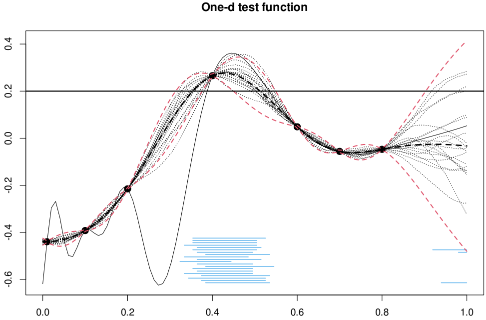
Z continuous paths;

$[t, +\infty)$ is a closed set.

The set $\Gamma = \{\mathbf{x} \in D : Z_{\mathbf{x}} \geq t\}$ is a random closed set.

Distribution of excursion sets

The posterior field induces a posterior distribution on excursion sets.



Note that

- Z continuous paths;
- $[t, +\infty)$ is a closed set.

The set $\Gamma = \{\mathbf{x} \in D : Z_{\mathbf{x}} \geq t\}$ is a random closed set.

How to summarize the distribution on sets?

Here we consider **estimates** for Γ^* with

Expectations of random closed sets[1],
in particular *Vorob'ev expectation*.

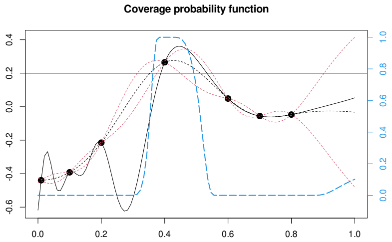
Conservative estimates, based on Vorob'ev quantiles.

1. for more definitions of expectation see Molchanov, I. (2017). Theory of Random Sets. Second edition. Springer.

Vorob'ev quantiles

Recall: posterior **coverage probability function of** Γ

$$p_n(\mathbf{x}) = P_n(\mathbf{x} \in \Gamma) = \Phi\left(\frac{m_n(\mathbf{x}) - t}{\sqrt{k_n(\mathbf{x},\mathbf{x})}}\right)$$



used in volume computation, tIMSE criterion;

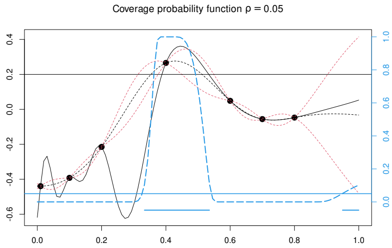
p_n creates a family of set estimates $Q_\rho = \{\mathbf{x} \in D : p_n(\mathbf{x}) \geq \rho\}$

high ρ, then points in Q_ρ have high marginal probability.

Vorob'ev quantiles

Recall: posterior **coverage probability function of** Γ

$$p_n(\mathbf{x}) = P_n(\mathbf{x} \in \Gamma) = \Phi\left(\frac{m_n(\mathbf{x}) - t}{\sqrt{k_n(\mathbf{x},\mathbf{x})}}\right)$$



used in volume computation, tIMSE criterion;

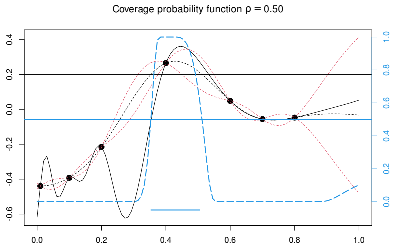
p_n creates a family of set estimates $Q_\rho = \{\mathbf{x} \in D : p_n(\mathbf{x}) \geq \rho\}$

high ρ, then points in Q_ρ have high marginal probability.

Vorob'ev quantiles

Recall: posterior **coverage probability function of** Γ

$$p_n(\mathbf{x}) = P_n(\mathbf{x} \in \Gamma) = \Phi\left(\frac{m_n(\mathbf{x}) - t}{\sqrt{k_n(\mathbf{x}, \mathbf{x})}}\right)$$



used in volume computation, tIMSE criterion;

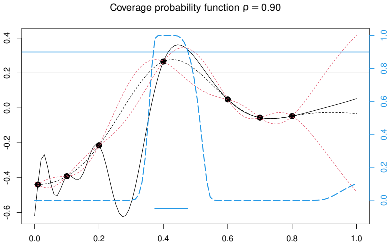
p_n creates a family of set estimates $Q_\rho = \{\mathbf{x} \in D : p_n(\mathbf{x}) \geq \rho\}$

high ρ, then points in Q_ρ have high marginal probability.

Vorob'ev quantiles

Recall: posterior **coverage probability function of** Γ

$$p_n(\mathbf{x}) = P_n(\mathbf{x} \in \Gamma) = \Phi\left(\frac{m_n(\mathbf{x}) - t}{\sqrt{k_n(\mathbf{x}, \mathbf{x})}}\right)$$



used in volume computation, tIMSE criterion;

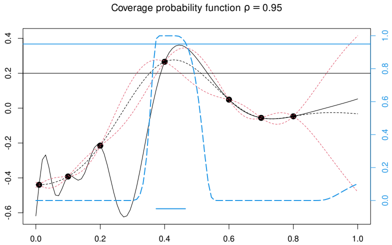
p_n creates a family of set estimates $Q_\rho = \{\mathbf{x} \in D : p_n(\mathbf{x}) \geq \rho\}$

high ρ, then points in Q_ρ have high marginal probability.

Vorob'ev quantiles

Recall: posterior **coverage probability function of** Γ

$$p_n(\mathbf{x}) = P_n(\mathbf{x} \in \Gamma) = \Phi\left(\frac{m_n(\mathbf{x}) - t}{\sqrt{k_n(\mathbf{x}, \mathbf{x})}}\right)$$



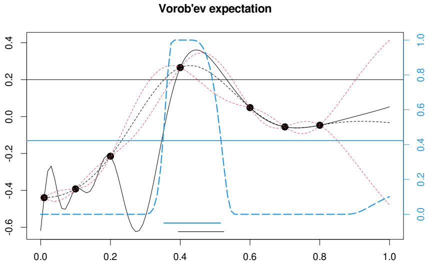
used in volume computation, tIMSE criterion;

p_n creates a family of set estimates $Q_\rho = \{\mathbf{x} \in D : p_n(\mathbf{x}) \geq \rho\}$

high ρ, then points in Q_ρ have high marginal probability.

Vorob'ev expectation

From Q_ρ, choose $Q_{\widetilde{\rho}}$ such that $\mu(Q_{\widetilde{\rho}}) = \mathbb{E}[\mu(\Gamma)]$.



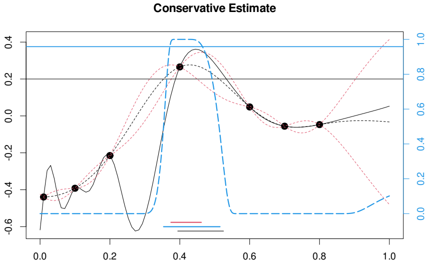
Properties:

- depends on (finite) measure μ;
- generally fast to compute;

Chevalier, C., Ginsbourger, D., Bect, J., and Molchanov, I. (2013). *Estimating and quantifying uncertainties on level sets using the Vorob'ev expectation and deviation with Gaussian process models.* mODa 10.

A **conservative estimate** of Γ^* is

$$C_{\Gamma,n} = Q_{\rho^*} \text{ where } \rho^* \in \arg\max_{\rho \in [0,1]} \{\mu(Q_\rho) : P_n(Q_\rho \subset \Gamma) \geq \alpha\}$$



- joint confidence statement on the set estimate;
- general method based on high dimensional Gaussian approximation

More

A., D., Ginsbourger, D. (2018) *Estimating Orthant Probabilities of High-Dimensional Gaussian Vectors with An Application to Set Estimation*, Journal of Computational and Graphical Statistics, 27:2, 255-267.

How to reduce the uncertainty on set estimate?

Uncertainty function(s):

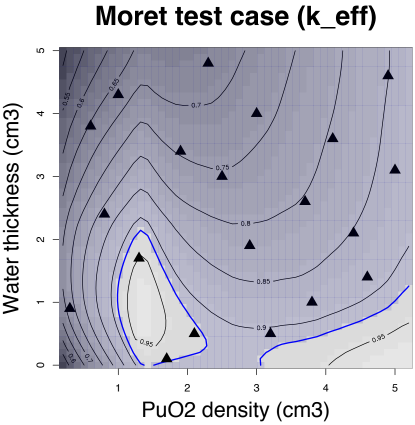
$$H_n(\rho) = \mathbb{E}_n[\mu(\Gamma \Delta Q_{n,\rho})], \quad \Gamma \Delta Q_{n,\rho} = \Gamma \setminus Q_{n,\rho} \cup Q_{n,\rho} \setminus \Gamma,$$
$$H_n^{\mathrm{T2}}(\rho_n^\alpha) = \mathbb{E}_n[\mu(\Gamma \setminus Q_{n,\rho_n^\alpha})],$$

For each uncertainty function, we define the following **SUR criteria**

$$J_n(\mathbf{x}) = \mathbb{E}_n[H_{n+1}(\rho) \mid X_{n+1} = \mathbf{x}] = \mathbb{E}_{n+1}[\mu(\Gamma \Delta Q_{n+1,\rho}) \mid X_{n+1} = \mathbf{x}]$$
$$J_n^{\mathrm{T2}}(\mathbf{x}) = \mathbb{E}_n[H_{n+1}^{\mathrm{T2}}(\rho_n^\alpha) \mid X_{n+1} = \mathbf{x}] = \mathbb{E}_{n+1}[\mu(\Gamma \setminus Q_{n+1,\rho_n^\alpha}) \mid X_{n+1} = \mathbf{x}],$$

More

Example: IRSN test case



Test case:

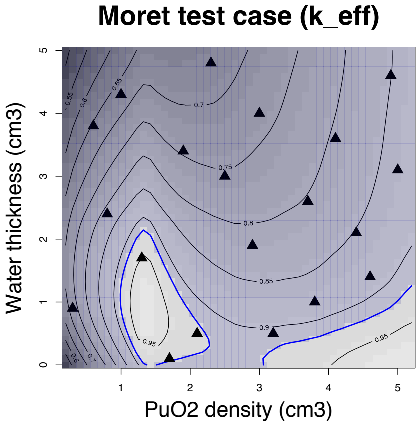
- $\mathrm{k_{eff}}$ function of $\mathrm{PuO_2}$ density and $\mathrm{H_2O}$ thickness, $D = [0.2, 5.2] \times [0, 5]$;
- $\mathrm{k_{eff}}$ continuous, expensive to evaluate;
- $n = 20$ observations (black triangles);

Objective: estimate $\Gamma^\star = \{\mathbf{x} \in D : f(\mathbf{x}) \leq t\}$ and evaluate the uncertainty of the estimate.

GRF model: m constant, k Matérn ($\nu = 5/2$), MLE for hyper-parameters.

Acknowledgements: Yann Richet, Institut de Radioprotection et de Sûreté Nucleaire.

Example: IRSN test case



Test case:

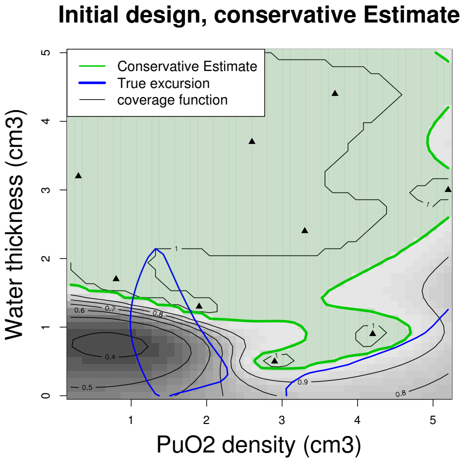
- $\mathrm{k_{eff}}$ function of $\mathrm{PuO_2}$ density and $\mathrm{H_2O}$ thickness, $D = [0.2, 5.2] \times [0, 5]$;
- $\mathrm{k_{eff}}$ continuous, expensive to evaluate;
- $n = 20$ observations (black triangles);

Objective: estimate $\Gamma^\star = \{\mathbf{x} \in D : f(\mathbf{x}) \leq t\}$ and evaluate the uncertainty of the estimate.

GRF model: m constant, k Matérn ($\nu = 5/2$), MLE for hyper-parameters.

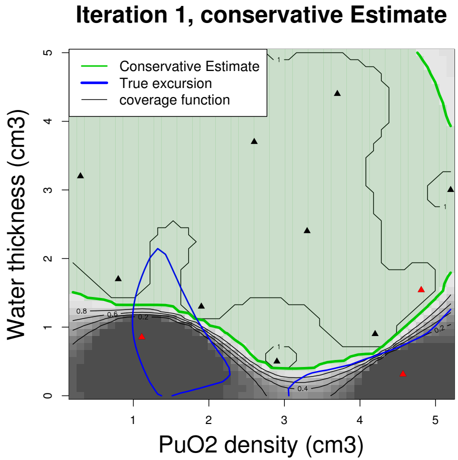
Acknowledgements: Yann Richet, Institut de Radioprotection et de Sûreté Nucleaire.

Example: type II uncertainty



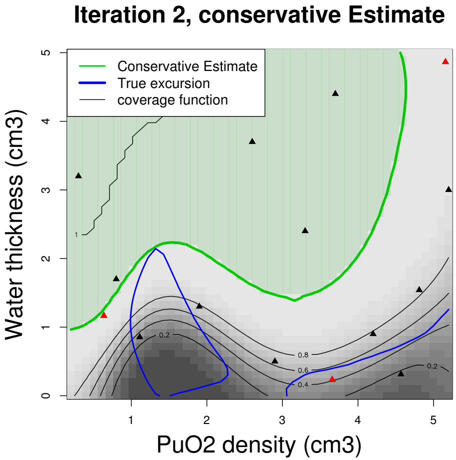
Criterion: $J^{\mathrm{T2}}_{n+q}(\mathbf{x}_q) = \mathbb{E}_n[\mu(\Gamma \setminus Q_{n+1,\rho_n^\alpha}) \mid \mathbf{X}_{n+q} = \mathbf{x}_q]$,

Example: type II uncertainty



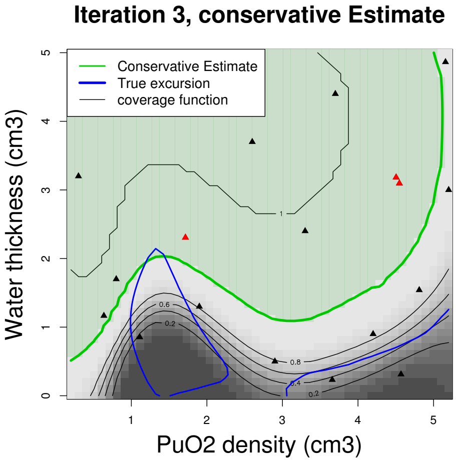
Criterion: $J^{\mathrm{T2}}_{n+q}(\mathbf{x}_q) = \mathbb{E}_n[\mu(\Gamma \setminus Q_{n+1,\rho_n^\alpha}) \mid \mathbf{X}_{n+q} = \mathbf{x}_q],$

Example: type II uncertainty



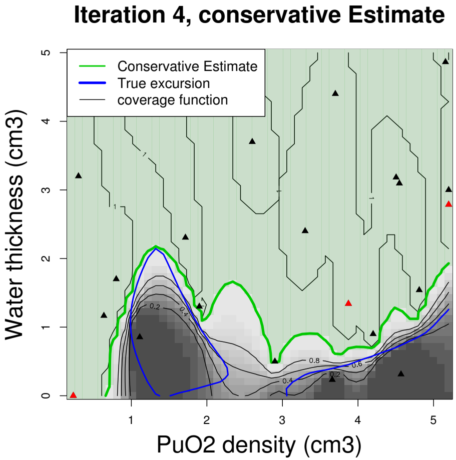
Criterion: $J^{T2}_{n+q}(\mathbf{x}_q) = \mathbb{E}_n[\mu(\Gamma \setminus Q_{n+1,\rho_n^\alpha}) \mid \mathbf{X}_{n+q} = \mathbf{x}_q],$

Example: type II uncertainty



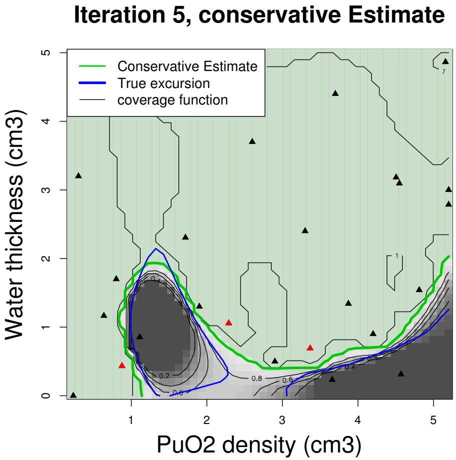
Criterion: $J^{T2}_{n+q}(\mathbf{x}_q) = \mathbb{E}_n[\mu(\Gamma \setminus Q_{n+1,\rho_n^\alpha}) \mid \mathbf{X}_{n+q} = \mathbf{x}_q]$,

Example: type II uncertainty



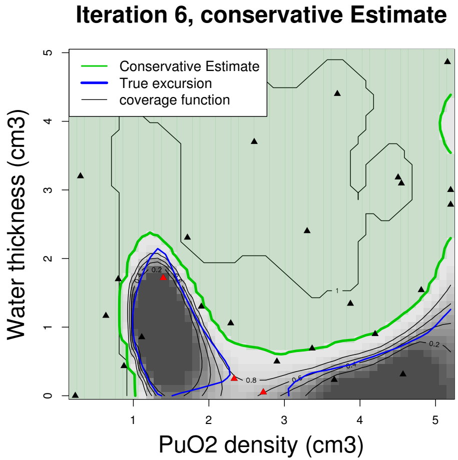
Criterion: $J^{T2}_{n+q}(\mathbf{x}_q) = \mathbb{E}_n[\mu(\Gamma \setminus Q_{n+1,\rho_n^\alpha}) \mid \mathbf{X}_{n+q} = \mathbf{x}_q],$

Example: type II uncertainty



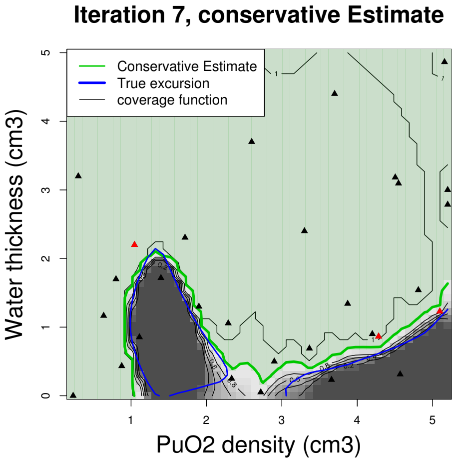
Criterion: $J^{T2}_{n+q}(\mathbf{x}_q) = \mathbb{E}_n[\mu(\Gamma \setminus Q_{n+1,\rho_n^\alpha}) \mid \mathbf{X}_{n+q} = \mathbf{x}_q]$,

Example: type II uncertainty



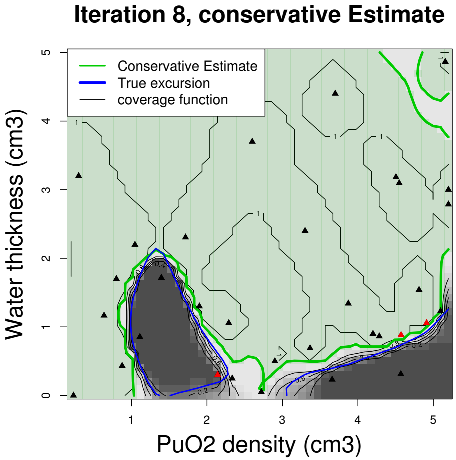
Criterion: $J^{\mathrm{T2}}_{n+q}(\mathbf{x}_q) = \mathbb{E}_n[\mu(\Gamma \setminus Q_{n+1,\rho_n^\alpha}) \mid \mathbf{X}_{n+q} = \mathbf{x}_q]$,

Example: type II uncertainty



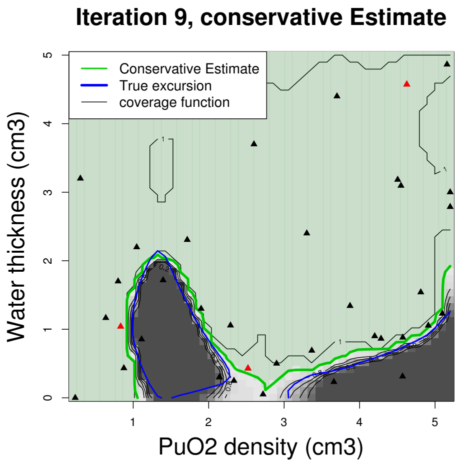
Criterion: $J^{T2}_{n+q}(\mathbf{x}_q) = \mathbb{E}_n[\mu(\Gamma \setminus Q_{n+1,\rho_n^\alpha}) \mid \mathbf{X}_{n+q} = \mathbf{x}_q],$

Example: type II uncertainty



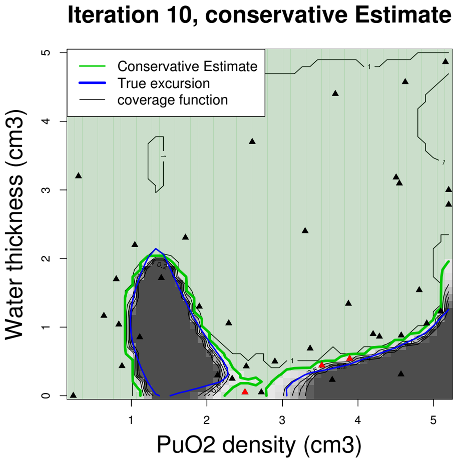
Criterion: $J^{\mathrm{T2}}_{n+q}(\mathbf{x}_q) = \mathbb{E}_n[\mu(\Gamma \setminus Q_{n+1,\rho_n^\alpha}) \mid \mathbf{X}_{n+q} = \mathbf{x}_q],$

Example: type II uncertainty



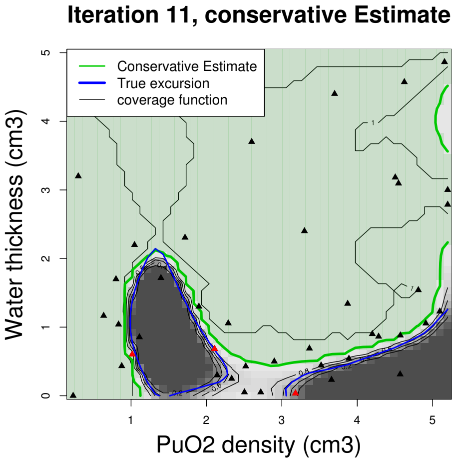
Criterion: $J^{\mathrm{T2}}_{n+q}(\mathbf{x}_q) = \mathbb{E}_n[\mu(\Gamma \setminus Q_{n+1,\rho_n^\alpha}) \mid \mathbf{X}_{n+q} = \mathbf{x}_q]$,

Example: type II uncertainty



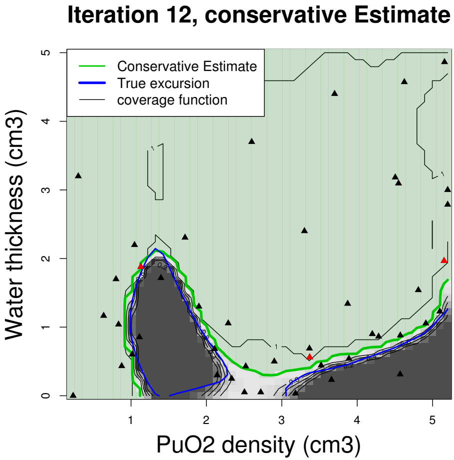
Criterion: $J^{T2}_{n+q}(\mathbf{x}_q) = \mathbb{E}_n[\mu(\Gamma \setminus Q_{n+1,\rho_n^\alpha}) \mid \mathbf{X}_{n+q} = \mathbf{x}_q]$,

Example: type II uncertainty



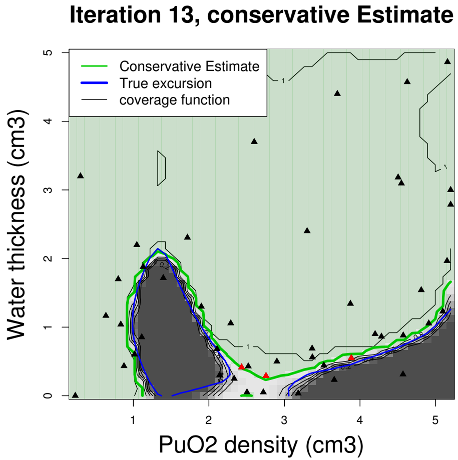
Criterion: $J^{T2}_{n+q}(\mathbf{x}_q) = \mathbb{E}_n[\mu(\Gamma \setminus Q_{n+1,\rho_n^\alpha}) \mid \mathbf{X}_{n+q} = \mathbf{x}_q]$,

Example: type II uncertainty



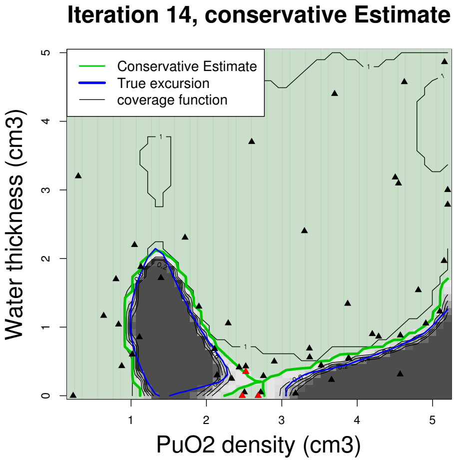
Criterion: $J^{T2}_{n+q}(\mathbf{x}_q) = \mathbb{E}_n[\mu(\Gamma \setminus Q_{n+1,\rho_n^\alpha}) \mid \mathbf{X}_{n+q} = \mathbf{x}_q]$,

Example: type II uncertainty



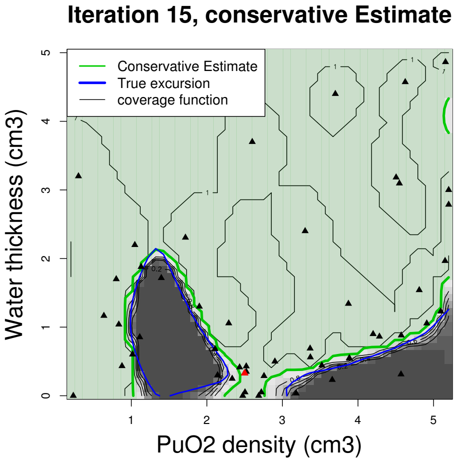
Criterion: $J^{\mathrm{T2}}_{n+q}(\mathbf{x}_q) = \mathbb{E}_n[\mu(\Gamma \setminus Q_{n+1,\rho_n^\alpha}) \mid \mathbf{X}_{n+q} = \mathbf{x}_q]$,

Example: type II uncertainty



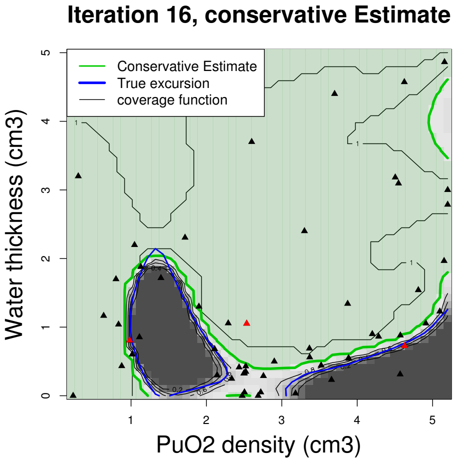
Criterion: $J^{T2}_{n+q}(\mathbf{x}_q) = \mathbb{E}_n[\mu(\Gamma \setminus Q_{n+1,\rho_n^\alpha}) \mid \mathbf{X}_{n+q} = \mathbf{x}_q],$

Example: type II uncertainty



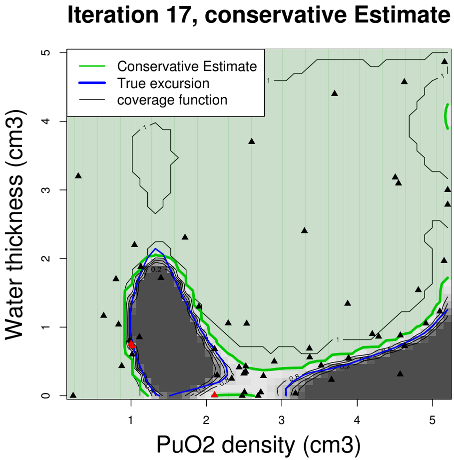
Criterion: $J^{T2}_{n+q}(\mathbf{x}_q) = \mathbb{E}_n[\mu(\Gamma \setminus Q_{n+1,\rho_n^\alpha}) \mid \mathbf{X}_{n+q} = \mathbf{x}_q]$,

Example: type II uncertainty



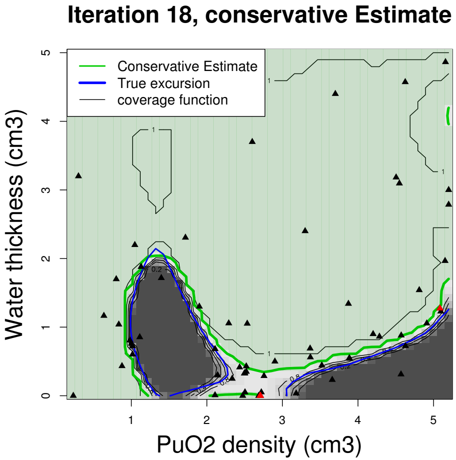
Criterion: $J^{T2}_{n+q}(\mathbf{x}_q) = \mathbb{E}_n[\mu(\Gamma \setminus Q_{n+1,\rho_n^\alpha}) \mid \mathbf{X}_{n+q} = \mathbf{x}_q],$

Example: type II uncertainty



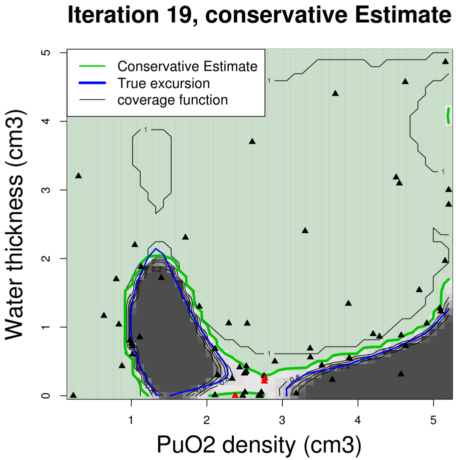
Criterion: $J^{\mathrm{T2}}_{n+q}(\mathbf{x}_q) = \mathbb{E}_n[\mu(\Gamma \setminus Q_{n+1,\rho_n^\alpha}) \mid \mathbf{X}_{n+q} = \mathbf{x}_q]$,

Example: type II uncertainty



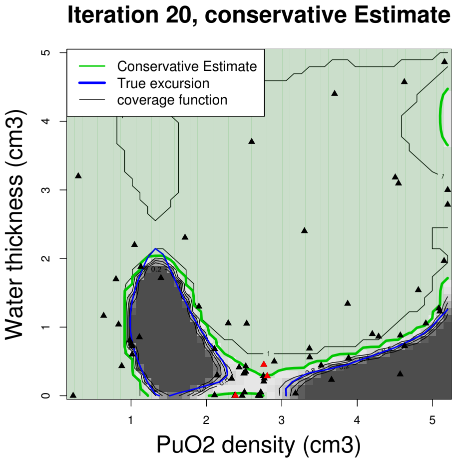
Criterion: $J^{T2}_{n+q}(\mathbf{x}_q) = \mathbb{E}_n[\mu(\Gamma \setminus Q_{n+1,\rho_n^\alpha}) \mid \mathbf{X}_{n+q} = \mathbf{x}_q],$

Example: type II uncertainty



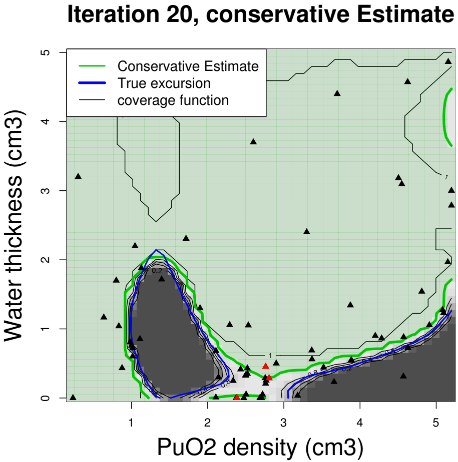
Criterion: $J^{T2}_{n+q}(\mathbf{x}_q) = \mathbb{E}_n[\mu(\Gamma \setminus Q_{n+1,\rho_n^\alpha}) \mid \mathbf{X}_{n+q} = \mathbf{x}_q]$,

Example: type II uncertainty



Criterion: $J^{T2}_{n+q}(\mathbf{x}_q) = \mathbb{E}_n[\mu(\Gamma \setminus Q_{n+1,\rho_n^\alpha}) \mid \mathbf{X}_{n+q} = \mathbf{x}_q],$

Example: Type II uncertainty



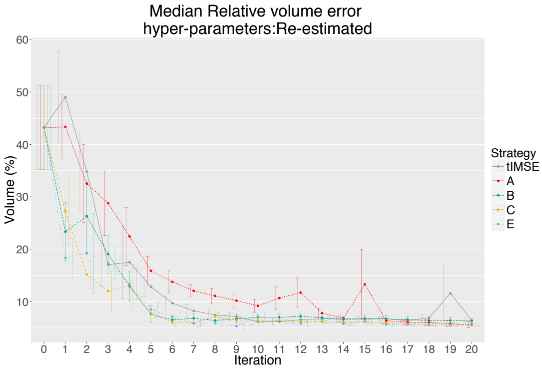
$n = 60$ new evaluations;

Next evaluation chosen in order to minimize criterion $J_n^{\mathrm{T}2}$;

Volume of updated CE: 21.30 (true excursion: 22.04, initial estimate: 18.10)

Comparison with state-of-the-art

IRSN test case: m constant, k Matérn ($\nu = 5/2$).



$n_0 = 10$ initial points;

$q = 3$ points per iteration;

Relative error with respect to 50×50 evaluation grid.

Strategy	Criterion
tIMSE	target=0.92
A	$J_n(\cdot; \rho_n)$, $\rho_n = 0.5$
B	$J_n(\cdot; \rho_n^*)$, $\alpha = 0.95$
C	$J_n^{\mathrm{MEAS}}(\cdot; \rho_n^*)$, $\alpha = 0.95$
E	$J_n^{\mathrm{T2}}(\cdot; \rho_n^*)$, $\alpha = 0.95$

A, D., Ginsbourger, D., Chevalier, C., Bect, J. and Richet, Y. (2021). *Adaptive Design of Experiments for Conservative Estimation of Excursion Sets.* Technometrics, 63:1, 13-26.

Thanks for your attention!

Contact: dario.azzimonti@idsia.ch

References

A., D., Ginsbourger, D. (2018) *Estimating Orthant Probabilities of High-Dimensional Gaussian Vectors with An Application to Set Estimation*, Journal of Computational and Graphical Statistics, 27:2, 255-267.

A, D., Ginsbourger, D., Chevalier, C., Bect, J. and Richet, Y. (2021). *Adaptive Design of Experiments for Conservative Estimation of Excursion Sets.* Submitted. Technometrics, 63:1, 13-26.

Azzimonti D. (2016) *Contributions to Bayesian set estimation relying on random field priors.* Ph.D. thesis, University of Bern.

Bect, J., Ginsbourger, D., Li, L., Picheny, V., and Vazquez, E. (2012). *Sequential design of computer experiments for the estimation of a probability of failure.* Stat. Comput., 22 (3):773—793.

Bect J., Bachoc F., Ginsbourger D. (2019) *A supermartingale approach to Gaussian process based sequential design of experiments.* Bernoulli 25 (4A) 2883 - 2919.

References

Bolin, D. and Lindgren, F. (2015). *Excursion and contour uncertainty regions for latent Gaussian models.* JRSS: B, 77(1):85—106.

Chevalier, C., Ginsbourger, D., Bect, J., and Molchanov, I. (2013). *Estimating and quantifying uncertainties on level sets using the Vorob'ev expectation and deviation with Gaussian process models.* mODa 10.

Chevalier, C., Bect, J., Ginsbourger, D., Vazquez, E., Picheny, V. and Richet, Y. (2013). *Fast parallel kriging-based stepwise uncertainty reduction with application to the identification of an excursion set.* Technometrics, 56 (4):455–465.

Chevalier C., Ginsbourger D., Emery X. (2014) *Corrected Kriging Update Formulae for Batch-Sequential Data Assimilation*. Lecture Notes in Earth System Sciences. Springer, Berlin, Heidelberg.

Chevalier C., (2013) *Fast uncertainty reduction strategies relying on Gaussian process models*. Ph.D. thesis, University of Bern.

References

Molchanov I. (2017). Theory of Random Sets. Second edition. Springer, London.

Picheny V., Ginsbourger D., Roustant O., Haftka R.T., Kim N. (2010). *Adaptive Designs of Experiments for Accurate Approximation of a Target Region.* ASME. J. Mech. Des. 132(7):071008-071008-9.

Ranjan, P., Bingham, D., and Michailidis, G. (2008). *Sequential experiment design for contour estimation from complex computer codes.*Technometrics, 50(4):527—541.

Sacks, J., Welch, W.J., Mitchell, T.J., Wynn, H.P. (1989). *Design and analysis of computer experiments.* Statistical Science 4(4), 409–435.

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):509–534.