HAL open science

Adaptive Design of Experiments for Conservative Estimation of Excursion Sets

Dario Azzimonti, David Ginsbourger, Clément Chevalier, Julien Bect, Yann Richet

To cite this version:

Dario Azzimonti, David Ginsbourger, Clément Chevalier, Julien Bect, Yann Richet. Adaptive Design of Experiments for Conservative Estimation of Excursion Sets. 2018. hal-01379642v3

HAL Id: hal-01379642

https://hal.science/hal-01379642v3

Preprint submitted on 9 Apr 2018 (v3), last revised 28 Jan 2020 (v6)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

Adaptive Design of Experiments for Conservative Estimation of Excursion Sets

Dario Azzimonti[*,†], David Ginsbourger[‡,§], Clément Chevalier[¶], Julien Bect[‖], Yann Richet[**]

Abstract

We consider a Gaussian process model trained on few evaluations of an expensive-to-evaluate deterministic function and we study the problem of estimating a fixed excursion set of this function. We focus on conservative estimates as they allow control on false positives while minimizing false negatives. We introduce adaptive strategies that sequentially selects new evaluations of the function by reducing the uncertainty on conservative estimates. Following the Stepwise Uncertainty Reduction approach we obtain new evaluations by minimizing adapted criteria. We provide tractable formulae for the conservative criteria and we benchmark the method on random functions generated under the model assumptions in two and five dimensions. Finally the method is applied to a reliability engineering test case. Overall, the proposed strategy of minimizing false negatives in conservative estimation achieves competitive performance both in terms of model based and a-posteriori indicators.

Keywords: Batch sequential strategy; Conservative estimate; Stepwise Uncertainty Reduction; Gaussian process model.

[*]The first author gratefully acknowledges the Swiss National Science Foundation, grant number 146354 and 167199 and the Hasler Foundation, grant number 16065

[†]Istituto Dalle Molle di studi sull'Intelligenza Artificiale (IDSIA), Scuola universitaria professionale della Svizzera italiana (SUPSI), Università della Svizzera italiana (USI), Via Cantonale 2c, 6928 Manno, Switzerland

[‡]Uncertainty Quantification and Optimal Design group, Idiap Research Institute, Centre du Parc, Rue Marconi 19, PO Box 592, 1920 Martigny, Switzerland.

[§]IMSV, Department of Mathematics and Statistics, University of Bern, Alpeneggstrasse 22, 3012 Bern, Switzerland.

[¶]Institute of Statistics, University of Neuchâtel, Avenue de Bellevaux 51, 2000 Neuchâtel, Switzerland.

[‖]Laboratoire des Signaux et Systèmes (UMR CNRS 8506), CentraleSupelec, CNRS, Univ Paris-Sud, Université Paris-Saclay, 91192, Gif-sur-Yvette, France.

[**]Institut de Radioprotection et de Sûreté Nucléaire (IRSN), Paris, France.

1 Introduction

The problem of estimating the set of inputs that leads a system to a particular behavior is common in many applications, notably reliability engineering (see, e.g., Bect et al., 2012; Chevalier et al., 2014a), climatology (see, e.g., French and Sain, 2013; Bolin and Lindgren, 2015) and many other fields (see, e.g., Bayarri et al., 2009; Arnaud et al., 2010; Wheeler et al., 2014). Here we consider a system modeled as a continuous expensive-to-evaluate function $f : \mathbb{X} \to \mathbb{R}$, where $\mathbb{X}$ is a compact subset of $\mathbb{R}^d$. Given few evaluations of f and a fixed closed $T \subset \mathbb{R}$, we are interested in estimates for the set

$$\Gamma^* = \{x \in \mathbb{X} : f(x) \in T\}. \tag{1}$$

The function f is modeled as a realization of a Gaussian process (GP) and, following Sacks et al. (1989), we emulate f with the posterior distribution of said process. Estimates for Γ^* can then be obtained from the posterior GP distribution.

A natural error measure for an estimate Γ of Γ^* is the distance in measure, i.e., $d_\mu(\Gamma^*, \Gamma) = \mu(\Gamma^* \Delta \Gamma) = \mu\big((\Gamma \setminus \Gamma^*) \ \cup \ (\Gamma^* \setminus \Gamma)\big)$ defined for $\Gamma \subset \mathbb{X}$ where μ is a finite measure on $\mathbb{X}$ such as a probability distribution or the Lebesgue measure. This metric assigns the same importance to the measure of false positives $(\Gamma \setminus \Gamma^*)$, i.e. input points estimated to be in Γ^* when in fact they are not, and false negatives $(\Gamma^* \setminus \Gamma)$, i.e. input points estimated to be in $\mathbb{X} \setminus \Gamma^*$ when in fact in Γ^*. Notably, Chevalier (2013); Chevalier et al. (2014a) estimate Γ^* with the Vorob'ev expectation which minimizes the expectation of the distance in measure among sets with measure equal to the expected measure. This estimate symmetrically minimizes (the measure of) false positive and false negative. In most applications, however, the cost of misclassification is not symmetric with higher penalties for false positives than for false negatives. For this reason, practitioners and reliability engineers are interested in estimates which would *very likely* be in the excursion set. Such property naturally gives more importance to the minimization of false positives than of false negatives.

French and Sain (2013); Bolin and Lindgren (2015) introduced the concept of conservatives estimates which select sets that are deliberately smaller – according to μ – than Γ^* and are included in the excursion set with a *large probability*, say $\alpha \approx 1$. The empty set trivially satisfies this probabilistic inclusion property, therefore conservative estimates are

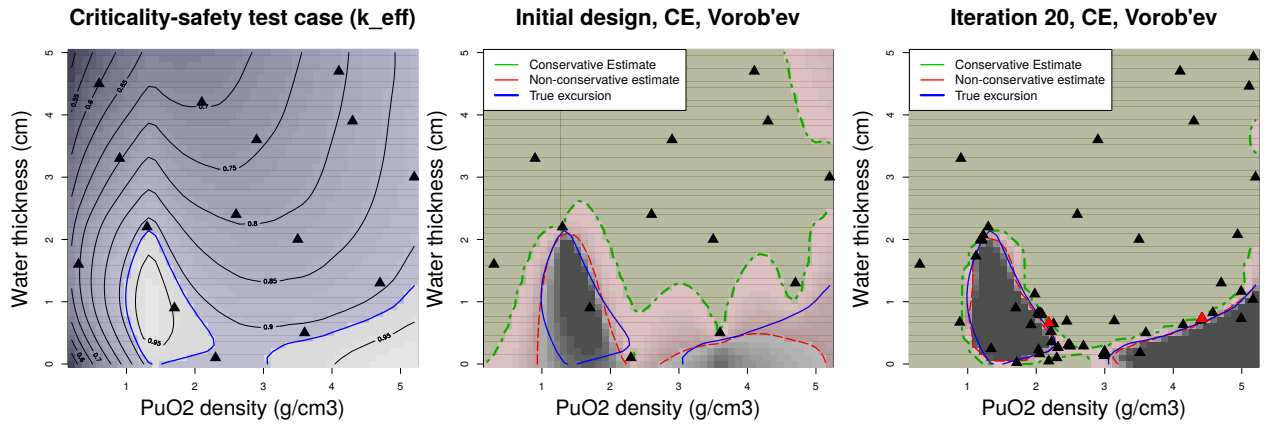
selected as the set with largest measure in this family. Conservative estimates minimize the measure of false negatives among sets with a fixed (usually low) probability of false positives. French and Sain (2013) and Bolin and Lindgren (2015) proposed an approach to compute conservative estimates for a fixed Design of Experiments (DoE), however, to the best of our knowledge, there is no study on how to reduce the uncertainty on conservative estimates with adaptive strategies. Here we focus on this problem by considering a notion of uncertainty for such sets and a sequential strategy to reduce the uncertainty.

Previous adaptive design of experiments strategies for excursion set estimation mainly focused on recovering the boundaries of the set. In particular, Picheny et al. (2010) introduced the targeted IMSE (tIMSE) criterion to add points at locations that improve the accuracy of the model at a certain level of the response variable. Bect et al. (2012) introduced the concept of Stepwise Uncertainty Reduction (SUR) strategies for GP (see also Vazquez and Bect, 2009; Chevalier et al., 2014a). Those strategies, however are not adapted to the conservative estimation case. Here, by shifting the focus on the control of false positives, we extend the conservative estimation framework introduced by Bolin and Lindgren (2015) to sequential design of experiments. We consider a definition of conservative estimates well suited to excursion sets of Gaussian processes and we provide a SUR strategy with tractable criteria to reduce the uncertainty on conservative estimates.

1.1 Motivating test case

In reliability engineering applications, the set Γ^* often represents safe inputs for a system. In such settings, it is vital to avoid flagging unsafe regions as safe. With a broad use of hypothesis testing terminology we refer to this as type I error or false positive.

Figure 1 shows an example of such reliability engineering applications: a test case from the French Institute for Radiological Protection and Nuclear Safety (IRSN). We briefly sketch the test case in this subsection, see section 5 for detailed results. The problem at hand concerns a nuclear storage facility and we are interested in estimating the set of parameters that lead to a safe storage of the material. Since this is closely linked to the production of neutrons, the safety of a system is evaluated with the neutron multiplication factor produced by fissile materials, called k-effective or k-eff : $\mathbb{X} \to [0, +\infty)$. In our appli-



(a) Function k-eff, set of interest Γ^* (shaded blue, $\mu(\Gamma^*)$=0.8816$\mu(\mathbb{X})$) and initial DoE ($n = 15$).

(b) Conservative ($\alpha = 0.95$, green) and non-conservative estimate (Vorob'ev expectation, red), initial DoE.

(c) Conservative ($\alpha = 0.95$, green) and non-conservative estimate (Vorob'ev expectation, red) after 75 evaluations.

Figure 1: Nuclear criticality safety test case. k-eff function (left), conservative and non-conservative estimates with 15 (LHS design, middle) and 75 (15+60 strategy C) evaluations.

cation $\mathbb{X} = [0.2, 5.2] \times [0, 5]$ with the two parameters representing the fissile material density, PuO_2, and the water thickness, H_2O. We are interested in the set of safe configurations

$$\Gamma^* = \{(PuO_2, H_2O) \in \mathbb{X} : \text{k-eff}(PuO_2, H_2O) \leq 0.92\}, \tag{2}$$

where the threshold $t = 0.92$ was chosen, for safety reasons, lower than the true critical case (k-eff > 1.0) when an uncontrolled chain reaction occurs. Figure 1a shows the set Γ^* shaded in blue and the contour levels for the true function computed from evaluations over a 50×50 grid, used as ground truth.

Figure 1b shows a conservative estimate at level $\alpha = 0.95$ (shaded green) and a non conservative one (Vorob'ev expectation, shaded red) computed from a GP model trained on $n = 15$ evaluations of k-eff, the true set Γ^* is delimited in blue. The DoE is a Latin Hypercube Sample (LHS). The conservative approach with $\alpha = 0.95$ provides an estimate inside the true set with high probability. Figure 1c shows that, as more evaluations are available, conservative and non-conservative estimates both get closer to the true excursion set. The estimates in this example are computed from 75 function evaluations, where the last 60 points were selected sequentially with an adaptive strategy (Strategy C in section 4).

In this paper we propose a method to compute conservative estimates well suited to examples such as figure 1. This method was used to compute the conservative estimates in figures 1b and 1c and to quantify the uncertainty on conservative estimates. The definition of an uncertainty allows us to define natural SUR strategies, which require the, possibly computationally challenging, evaluation of criteria. Here we provide tractable formulae for such criteria in order to mitigate the strategies' computational burden.

1.2 Outline of the paper

The remainder of the paper is structured as follows. In the next section we briefly recall some background material necessary to follow the contributions. In particular, section 2.1 reviews set estimates preliminary to this work and section 2.2 recalls the concept of SUR strategies. In section 3 we motivate our definition of conservative estimates and we introduce the metrics used to quantify the uncertainty on such estimates. In section 4, we detail the proposed sequential strategies, we derive closed-form formulae for the criteria and we illustrate their implementation. Section 5 presents the results obtained on the IRSN test case introduced in this section and section 6 shows a benchmark study on Gaussian process realizations. In supplementary material we further apply the proposed strategies on a coastal flood problem. All proofs are in appendices A and B.

2 Background

In this work we assume that the function f is observed with measurement noise, i.e.

$$y_i = f(x_i) + \tau(x_i)\varepsilon_i \qquad x_i \in \mathbb{X}, \quad \varepsilon_i \sim N(0,1), \quad i = 1, \dots, n$$

with ε_i independent measurement noise and τ^2 a known heterogeneous noise variance.

In a Bayesian framework (see, e.g., Chilès and Delfiner, 2012, and references therein) we consider f as a realization of an a.s. continuous Gaussian process (GP) $\xi \sim GP(\mathfrak{m}^\xi, \mathfrak{K}^\xi)$, with mean function $\mathfrak{m}^\xi(x) := \mathbb{E}[\xi_x]$ and covariance function $\mathfrak{K}^\xi(x, x') := \mathrm{Cov}(\xi_x, \xi_{x'}),\ x, x' \in \mathbb{X}$. We consider the process $Z_x = \xi_x + \tau(x)\varepsilon_x$, defined for $x \in \mathbb{X}$ with $\varepsilon_x \sim N(0,1)$ i.i.d. and independent from ξ for each $x \in \mathbb{X}$. The process Z is Gaussian and we denote by

$\mathfrak{m}$ and with $\mathfrak{K}$ its mean and covariance functions respectively. For $n > 0$, we denote by $\mathbf{y}_n = (y_1, \ldots, y_n) \in \mathbb{R}^n$ the observations at an initial design of experiments (DoE) $\mathbf{X}_n = (x_1, \ldots, x_n) \in \mathbb{X}^n$. The posterior distribution of the process is Gaussian with mean and covariance computed as the conditional mean $\mathfrak{m}_n$ and conditional covariance $\mathfrak{K}_n$ given the observations, see, e.g., Santner et al. (2003) for the closed form formulae.

2.1 Vorob'ev expectation and conservative estimates

There exists different approaches to estimate Γ^* from the posterior distribution of Z and to quantify the uncertainty on such estimates, see, e.g. Chevalier et al. (2014a); Bolin and Lindgren (2015); Azzimonti (2016) for more details. In this section we briefly review the Vorob'ev expectation and the conservative estimate approach.

The prior distribution on Z induces a distribution on the (random) excursion set $\Gamma = \{x \in \mathbb{X} : \xi_x \in T\}$. Summaries of the posterior distribution of the random closed set Γ, provide estimates for Γ^*. A central tool for the approach presented here is the *coverage probability function* of a random closed set Γ, defined as

$$p_\Gamma(x) = P(x \in \Gamma), \; x \in \mathbb{X}.$$

In our case we consider the posterior coverage function $p_{\Gamma,n} = p_n$, defined with the posterior probability $P_n(\cdot) = P(\cdot \mid Z_{\mathbf{X}_n} = \mathbf{y}_n)$, where $Z_{\mathbf{X}_n} = (Z_{x_1}, \ldots, Z_{x_n})$ and we drop the subscript Γ as the set is clear from the context. If $T = (-\infty, t]$, then $p_n(x) = \Phi\left(\frac{t - \mathfrak{m}_n(x)}{\mathfrak{s}_n(x)}\right)$, where $\Phi(\cdot)$ is the CDF of a standard Normal random variable and $\mathfrak{s}_n(x) = \sqrt{\mathfrak{K}_n(x,x)}$. The coverage function defines the family of *Vorob'ev quantiles*

$$Q_{n,\rho} = \{x \in \mathbb{X} : p_n(x) \geq \rho\}, \tag{3}$$

with $\rho \in [0,1]$. These sets are closed for each $\rho \in [0,1]$ as the coverage function is upper semi-continuous, see Molchanov (2005), Proposition 1.34.

The level ρ can be selected in different ways. A non-conservative choice with $\rho = 0.5$ leads to the *Vorob'ev median*. The *Vorob'ev expectation* (Vorob'ev, 1984; Molchanov, 2005; Chevalier et al., 2013) is defined through a finite measure μ on $\mathbb{X}$ as the quantile Q_{n,ρ_V} such that the measure $\mu(Q_{n,\rho_V})$ is as close as possible to $\mathbb{E}[\mu(\Gamma)]$, the expected measure

Table 1: Summary values for example in figure 2, estimated from 100 GP realizations.

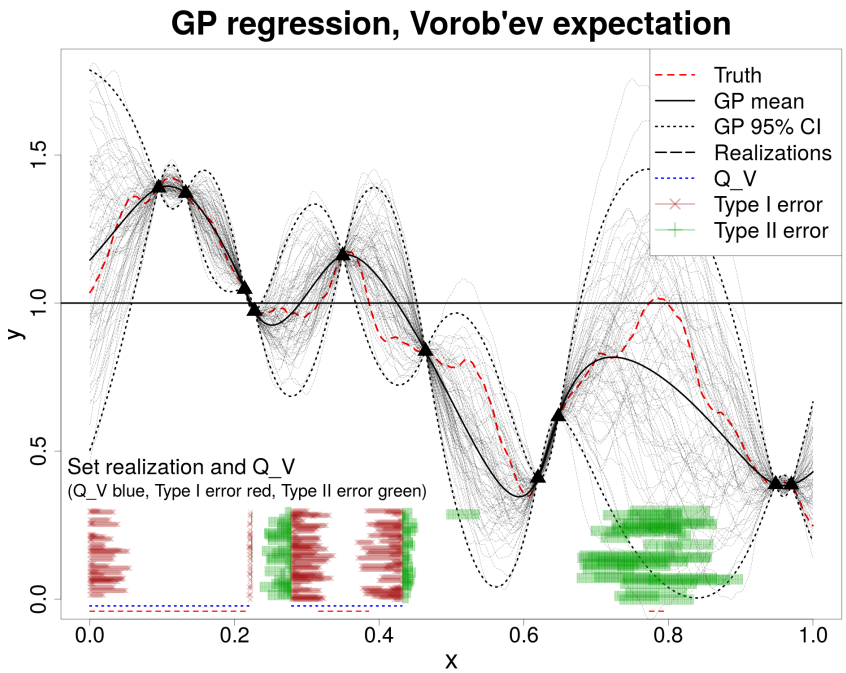
	ρ	Type I error (mean $\pm$ sd)	Type II error (mean $\pm$ sd)	$\hat{P}(Q_\rho \subset \Gamma)$
Q_{ρ_V}	0.393	0.046 ± 0.029	0.053 ± 0.058	0.02
$Q_{0.5}$	0.500	0.035 ± 0.026	0.061 ± 0.058	0.07
$Q_{0.95}$	0.950	$5.7 \times 10^{-4} \pm 1.9 \times 10^{-3}$	0.168 ± 0.063	0.87
$\mathrm{CE}_{0.95}$	0.987	$9.0 \times 10^{-5} \pm 4.7 \times 10^{-4}$	0.187 ± 0.063	0.95

of Γ. The measure μ is often a probability measure or the Lebesgue measure on $\mathbb{X}$. The set Q_{n,ρ_V} is also the minimizer of $\mathbb{E}[\mu(\Gamma \Delta M)]$ among all sets such that $\mu(M) = \mathbb{E}[\mu(\Gamma)]$, see, e.g., Molchanov (2005, Theorem 2.3, Chapter 2). The Vorob'ev expectation minimizes symmetrically the expected measure of false positives ($\mathbb{E}[\mu(M \setminus \Gamma)]$) and false negatives ($\mathbb{E}[\mu(\Gamma \setminus M)]$) among sets with measure equal to the expected measure. In section 3 we prove a similar result for generic Vorob'ev quantiles. The quantity $\mathbb{E}[\mu(\Gamma_1 \Delta \Gamma_2)]$, for two random sets $\Gamma_1, \Gamma_2 \subset \mathbb{X}$, is the probabilistic equivalent of the distance in measure introduced in section 1 and it is often called *expected distance in measure*. In the Ph.D. thesis of Chevalier (2013), this notion was introduced to adaptively reduce the uncertainty on Vorob'ev expectations. Here, it is used in section 3.2 to provide uncertainty functions for conservative estimates.

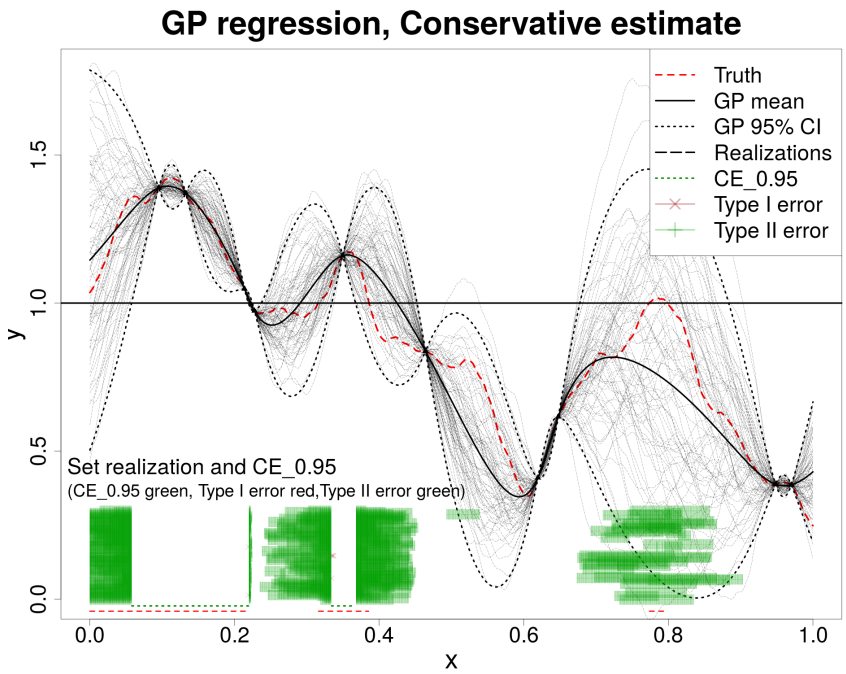
Conservative estimates (Bolin and Lindgren, 2015; French and Sain, 2013) are a different type of set estimate that embed a probabilistic control on false positives in the estimator. Denote with $\mathfrak{C}$ a family of closed subsets in $\mathbb{X}$. A *conservative estimate at level* α for Γ^* is a set $\mathrm{CE}_{\alpha,n}$ defined as

$$\mathrm{CE}_{\alpha,n} \in \arg\max_{C \in \mathfrak{C}} \{\mu(C) : P_n(C \subset \Gamma) \geq \alpha\}. \tag{4}$$

The condition $P_n(C \subset \Gamma) = P_n(C \setminus \Gamma = \emptyset) \geq \alpha$ controls the probability of false positives. Figure 2 shows estimates for the excursion set of a one-dimensional example generated as a realization of a mean zero GP with Matérn ($\nu = 3/2$, $l = 0.3$, $\sigma^2 = 0.3$) covariance. In particular, figure 2a shows how false positive (type I error, red) and false negatives (type II error, green) are symmetrically minimized by the Vorob'ev expectation. Figure 2b instead shows that a conservative estimate with $\alpha = 0.95$ has low probability of false



(a) Vorob'ev expectation (Q_V, blue dotted).



(b) Conservative estimate ($\alpha = 0.95$, green dotted).

Figure 2: 1-dimensional example: true function and set (red, dashed), posterior GP mean (black) and posterior realizations (black, dashed) with $n = 10$ evaluations. Type I/II errors for Vorob'ev expectation and conservative estimate.

positives, however results in much larger false negatives. Table 1 reports the values for the expected volume of type I and II errors and the estimated probability of inclusion. This example shows that the Vorob'ev expectation might be closer to the truth than conservative estimates, however $\mathrm{CE}_{\alpha,n}$ gives control on the probability of false positives. Table 1 also reports the values for the Vorob'ev quantile $Q_{0.95}$; in particular note that $P(Q_{0.95} \subset \Gamma) < 0.95$. This is a consequence of the quantile's definition based on the marginal probability $p_n(x) \geq 0.95$, $x \in \mathbb{X}$, which does imply any joint probability statement.

The computation of $\mathrm{CE}_{\alpha,n}$ in equation (4) requires a set of maximum measure which is included in the random set Γ with probability at least α. Such an optimization procedure can be very challenging and crucially depends on the choice of the family $\mathfrak{C}$. Often (see, e.g., French and Sain, 2013; Bolin and Lindgren, 2015; Azzimonti and Ginsbourger, 2017) $\mathfrak{C}$ is a parametric family of nested sets depending on a one dimensional parameter. Here we rely on the family of Vorob'ev quantiles $\{Q_\rho : \rho \in [0,1]\}$ and, in section 3, we provide motivation for this choice.

2.2 SUR strategies

Sequential design of experiments adaptively chooses the next evaluations according to a strategy with the aim of improving the final estimate. We follow the Stepwise Uncertainty Reduction (SUR, see, e.g., Fleuret and Geman, 1999; Bect et al., 2012; Chevalier et al., 2014a; Bect et al., 2017) approach and we select a sequence of points $X_1, X_2, \ldots, X_n$ in order to reduce the uncertainty on selected quantities of interest. In the remainder of the paper we denote by $\mathbb{E}_n[\cdot] = \mathbb{E}[\cdot \mid Z_{\mathbf{X}_n} = \mathbf{y}_n]$ the expectation conditional on $Z_{\mathbf{X}_n} = \mathbf{y}_n$. Moreover in a sequential setting the locations $X_1, \ldots, X_n$ and the evaluations at those points are random, therefore we denote by $\mathcal{A}_n$ the σ-algebra generated by the couples $X_1, Z_{X_1}, \ldots, X_n, Z_{X_n}$. Finally $\mathbb{E}_{n,\mathbf{x}^{(q)}}[\cdot]$ denotes the conditional expectation with respect to $\mathcal{A}_{n+q}$ given the first n evaluations and with $(X_{n+1}, \ldots, X_{n+q}) = \mathbf{x}^{(q)}$ for a fixed $\mathbf{x}^{(q)} = (x_{n+1}, \ldots, x_{n+q}) \in \mathbb{X}^q$.

For a specific problem, we define a measure of residual uncertainty at step n, denoted by H_n, a $\mathcal{A}_n$-measurable random variable. If the first n locations and evaluations are known, then H_n is a (deterministic) real number quantifying the residual uncertainty on

the estimate. A SUR strategy selects the next locations

$$\mathbf{X}^*_{n+q} \in \arg\min_{\mathbf{x}^{(q)} \in \mathbb{X}^q} \mathbb{E}_{n,\mathbf{x}^{(q)}}[H_{n+q}], \tag{5}$$

i.e. a minimizer of the future uncertainty in expectation. For a more complete and theoretical perspective on SUR strategies see, e.g., Bect et al. (2017) and references therein. There are many ways to proceed with the minimization introduced above, see, e.g., Osborne et al. (2009); Ginsbourger and Le Riche (2010); Bect et al. (2012); González et al. (2016) and references therein. Here we focus on batch-sequential sub-optimal strategies, also called *one-step lookahead strategies*, that select the next batch of locations by greedily minimizing a one-step lookahead sampling criterion. The objective function in equation (5) is called *batch sequential one-step lookahead sampling criterion* and is denoted by $J_n : \mathbf{x}^{(q)} \in \mathbb{X}^q \mapsto \mathbb{E}_{n,\mathbf{x}^{(q)}}[H_{n+q}] \in \mathbb{R}$. By minimizing this criterion we obtain the new locations where to evaluate f. Batch sequential sampling criteria are often used in practice because parallel function evaluations can save user time.

In sections 3.2 and 4, we introduce uncertainty functions and SUR criteria tailored for conservative estimates. Let us first specify our definition of conservative estimates.

3 Static properties of conservative estimate

3.1 Conservative estimates with Vorob'ev quantiles

The conservative estimate definition in equation (4) requires a family $\mathfrak{C}$ where to search for the optimal set $\mathrm{CE}_{\alpha,n}$. In practice, it is convenient to choose a parametric family indexed by a real parameter. Here we choose $\mathfrak{C}_\rho = \{Q_\rho : \rho \in [0,1]\}$, i.e., the Vorob'ev quantiles. This is a nested family indexed by $\rho \in [0,1]$ where $Q_0 = \mathbb{X} \in \mathfrak{C}_\rho$ and, for each $\rho_1 > \rho_2$,

$$Q_{\rho_1} \subset Q_{\rho_2}, \qquad Q_{\rho_1}, Q_{\rho_2} \in \mathfrak{C}_\rho. \tag{6}$$

We now detail how to compute $\mathrm{CE}_{\alpha,n}$ based on $\mathfrak{C}_\rho$, for a fixed $\alpha \in [0,1]$ from n observations. For each $\rho \in [0,1]$, we define the function $\psi_\Gamma : [0,1] \to [0,1]$ that associates to each ρ the probability $\psi_\Gamma(\rho) := P_n(Q_\rho \subset \Gamma)$. The function ψ_Γ is non decreasing due to the nested property in equation (6). Moreover, $\mu(Q_{\rho_1}) \leq \mu(Q_{\rho_2})$ for $\rho_1 \geq \rho_2$. The

computation of $\mathrm{CE}_{\alpha,n}$ amounts to finding the smallest $\rho = \rho^\star$ such that $\psi_\Gamma(\rho^\star) \geq \alpha$, which is achievable, for example, with a simple dichotomic search. For $T = (-\infty, t]$, we approximate $\psi_\Gamma(\rho) \approx P_n(Z_{q_1} \leq t, \ldots, Z_{q_\ell} \leq t)$, where $\{q_1, \ldots, q_\ell\} \subset Q_\rho$ is a set of ℓ points in Q_ρ, with ℓ large, and use the computationally efficient integration proposed by Azzimonti and Ginsbourger (2017). The procedure above is valid for any nested family of sets indexed by a real parameter, however, the Vorob'ev quantiles, in addition, have the following property.

Proposition 1. *Consider a measure μ such that $\mu(\mathbb{X}) < \infty$ and an arbitrary $\rho \in [0, 1]$. A Vorob'ev quantile Q_ρ minimizes the expected distance in measure with Γ among all measurable M such that $\mu(M) = \mu(Q_\rho)$.*

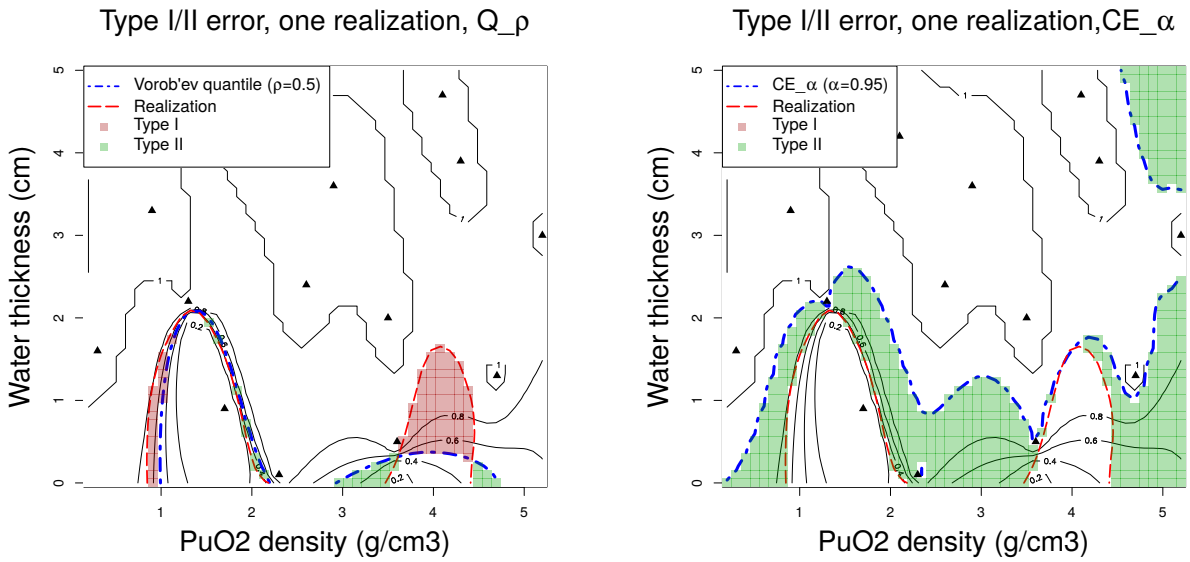
Proposition 1 is an extension of Theorem 2.3, Molchanov (2005) to a generic Vorob'ev quantile. As a consequence, a conservative estimate $\mathrm{CE}_{\alpha,n} = Q_{n,\rho_n^\alpha}$ computed with Vorob'ev quantiles minimizes the expected measure of false negatives ($\Gamma \setminus Q_{n,\rho_n^\alpha}$) for fixed probability of false positives ($Q_{n,\rho_n^\alpha} \setminus \Gamma$). In general, the Vorob'ev quantile chosen for $\mathrm{CE}_{\alpha,n}$ with this procedure is not the set S with the largest measure satisfying the property $P(S \subset \Gamma) \geq \alpha$. See supplementary material for a counterexample.

In the remainder of the paper we consider only conservative estimates obtained with Vorob'ev quantiles thus $\mathrm{CE}_{\alpha,n}$ is also denoted as Q_{n,ρ_n^α}.

3.2 Uncertainty quantification on conservative estimates

Our object of interest is Γ^*, therefore we require uncertainty functions that take into account the whole set. Chevalier et al. (2013); Chevalier (2013) evaluate the uncertainty on the Vorob'ev expectation with the *Vorob'ev uncertainty*, i.e. the expected distance in measure between the current estimate Q_{n,ρ_n} and the set Γ. In this section we introduce an uncertainty adaptive to conservative estimates. The idea is to describe the uncertainty by looking at the expected measure of false negatives. In the example of figure 2b, this quantity is the mean measure of the sets in green. Expected distance in measure and false negatives are related concepts and, in order to highlight this connection, let us first recall that the Vorob'ev uncertainty of a quantile Q_ρ is

$$H_{n,\rho} = \mathbb{E}_n[\mu(\Gamma \Delta Q_{n,\rho})] = \mathbb{E}_n[\mu(Q_{n,\rho} \setminus \Gamma)] + \mathbb{E}_n[\mu(\Gamma \setminus Q_{n,\rho})], \qquad \rho \in [0, 1]. \tag{7}$$



(a) $\rho_n = 0.5$. Expected type I/II $= 3.49 \times 10^{-2}$ / 1.76×10^{-2}.

(b) $\rho_n^\alpha = 0.997$. Expected type I/II $= 7.41 \times 10^{-5}$ / 2.18×10^{-1}.

Figure 3: Sets $Q_{n,\rho} \backslash \Gamma$ (red) and $\Gamma \backslash Q_{n,\rho}$ (green) for one realization $\Gamma(\omega)$ from the conditional distribution of Γ given $n = 15$ evaluations of k-eff.

In the following sections, this uncertainty measure is computed with $\rho = 0.5$, the Vorob'ev median and with $\rho = \rho_n^\alpha$, the conservative estimate at level α. Let us denote by $G_n^{(1)}(\rho) = \mu(Q_{n,\rho} \setminus \Gamma)$ and $G_n^{(2)}(\rho) = \mu(\Gamma \setminus Q_{n,\rho})$ the random variables associated with the measure of the first and the second set difference in equation (7).

Proposition 2. *Consider the conservative estimate Q_{n,ρ_n^α}, then the ratio between the error $\mathbb{E}_n[G_n^{(1)}(\rho_n^\alpha)]$ and the measure $\mu(Q_{n,\rho_n^\alpha})$ is bounded by $1 - \alpha$, the chosen level for the conservative estimates.*

A conservative estimate Q_{n,ρ_n^α} aims at controlling the error $\mathbb{E}_n[G_n^{(1)}(\rho_n^\alpha)]$. We call *Type I* and *Type II* errors given n observations the quantities $\mathbb{E}_n[G_n^{(1)}(\rho_n^\alpha)]$ and $\mathbb{E}_n[G_n^{(2)}(\rho_n^\alpha)]$ respectively. Figure 3 shows the sets $\Gamma(\omega) \setminus Q_{n,\rho}$ (green region) and $Q_{n,\rho} \setminus \Gamma(\omega)$ (red region) for one realization $\Gamma(\omega)$, $\omega \in \Omega$, of the posterior distribution of Γ given $n = 15$ evaluations of f in the IRSN example presented in figure 1. Notice how the conservative estimate ($\alpha = 0.95$, figure 3b) results in a smaller red region and in a much larger green region than $Q_{0.5}$. Conservative estimates at high levels α tend to select regions inside Γ. In particular

if the posterior GP mean provides a good approximation of the function f, conservative estimates with high α tend to be inside the true excursion set Γ^*. In such situations the Type I error is usually very small, as shown in proposition 2, while Type II error could be rather large. For the IRSN example in figures 1 and 3, the type II error is equal to 2.18×10^{-1} for the conservative level, while it is equal to 1.76×10^{-2} for $\rho_n = 0.5$. Type II error provides a quantification of the residual uncertainty on the conservative estimate; we formalize this concept with the following definition.

Definition 1 (Type II uncertainty)**.** *Consider the Vorob'ev quantile Q_{n,ρ_n^α} corresponding to the conservative estimate at level α for Γ. The* Type II uncertainty *is defined as*

$$H^{\mathrm{T2}}_{n,\rho_n^\alpha} := \mathbb{E}_n[G_n^{(2)}(\rho_n^\alpha)] = \mathbb{E}_n[\mu(\Gamma \setminus Q_{n,\rho_n^\alpha})]. \tag{8}$$

4 SUR strategies for conservative estimates

The measures of residual uncertainty introduced in the previous section can be used to define SUR strategies for conservative estimates. We consider a set up where the $\mathbf{X}_n$ and the respective function evaluations are known and we introduce one-step lookahead SUR criteria for conservative estimates. In a sequential algorithm we minimize such criteria to select the next batch of $q > 0$ locations $X_{n+1}, \ldots, X_{n+q} \in \mathbb{X}$ where to evaluate the function.

4.1 SUR criteria

Since the locations $X_{n+1}, \ldots, X_{n+q}$ and the responses $Z_{X_{n+1}}, \ldots, Z_{X_{n+q}}$ are unknown, the uncertainty H_{n+q} and the conservative level ρ_{n+q}^α are $\mathcal{A}_{n+q}$-measurable random variables. The criteria introduced below (equations (9) and (12)) are properly defined for $\rho = \rho_{n+q}^\alpha$, however, there are no closed form formulae to solve the expectations in their definitions. For this reason, the criteria's implementations use the last known level ρ_n^α. We consider two sampling criterion based on the uncertainty functions in equations (7) and (8).

The *conservative* J_n criterion is defined as

$$J_n(\mathbf{x}^{(q)};\rho_n^\alpha) = \mathbb{E}_{n,\mathbf{x}^{(q)}}\left[H_{n+q,\rho_n^\alpha}\right] = \mathbb{E}_{n,\mathbf{x}^{(q)}}\left[\mu(\Gamma\Delta Q_{n+q,\rho_n^\alpha})\right] \tag{9}$$

for $\mathbf{x}^{(q)} = (x_{n+1},\dots,x_{n+q}) \in \mathbb{X}^q$, where Q_{n+q,ρ_n^α} is the Vorob'ev quantile obtained with $n+q$ evaluations of the function at level ρ_n^α, the conservative level obtained with n evaluations. This is an adaptation of the Vorob'ev criterion introduced by Chevalier (2013) based on the Vorob'ev deviation (Vorob'ev, 1984; Molchanov, 2005; Chevalier et al., 2013).

Chevalier (2013), Chapter 4.2, derives the formula for this criterion for the Vorob'ev expectation, i.e. the quantile at level $\rho = \rho_{n,V}$. In the following proposition we extend this result to any $\mathcal{A}_n$ measurable quantile ρ_n.

Proposition 3 (Criterion J_n)**.** *Consider $\Gamma^* = \{x \in \mathbb{X} : f(x) \in T\}$ with $T = [t,+\infty)$, where $t \in \mathbb{R}$ is a fixed threshold, then the criterion J_n can be expanded in closed-form as*

$$\begin{aligned} J_n(\mathbf{x}^{(q)};\rho_n) &= \mathbb{E}_{n,\mathbf{x}^{(q)}}\left[\mu\left(\Gamma\Delta Q_{n+q,\rho_n}\right)\right] \\ &= \int_{\mathbb{X}} \left(2\Phi_2\left(\begin{pmatrix} a_{n+q}(u) \\ \Phi^{-1}(\rho_n) - a_{n+q}(u) \end{pmatrix}; \begin{pmatrix} 1+\gamma_{n+q}(u) & -\gamma_{n+q}(u) \\ -\gamma_{n+q}(u) & \gamma_{n+q}(u) \end{pmatrix} \right) \right. \\ &\quad \left. - p_n(u) + \Phi\left(\frac{a_{n+q}(u) - \Phi^{-1}(\rho_n)}{\sqrt{\gamma_{n+q}(u)}} \right) \right) d\mu(u), \end{aligned} \tag{10}$$

where

$$\begin{aligned} a_{n+q}(u) &= \frac{\mathfrak{m}_n(u) - t}{\mathfrak{s}_{n+q}(u)}, & \mathbf{b}_{n+q}(u) &= \frac{K_q^{-1}\mathfrak{K}_n(\mathbf{x}^{(q)},u)}{\mathfrak{s}_{n+q}(u)}, \\ \gamma_{n+q}(u) &= \mathbf{b}_{n+q}^T(u) K_q \mathbf{b}_{n+q}(u) & p_n(u) &= \Phi\left(\frac{\mathfrak{m}_n(u) - t}{\mathfrak{s}_n(u)}\right), \ u \in \mathbb{X}, \end{aligned} \tag{11}$$

with $\mathfrak{K}_n(\mathbf{x}^{(q)},u) = (\mathfrak{K}_n(x_{n+1},u),\dots,\mathfrak{K}_n(x_{n+q},u))^T$, $K_q = \mathfrak{K}_n(\mathbf{x}^{(q)},\mathbf{x}^{(q)}) + \tau^2(\mathbf{x}^{(q)})I_q$, where $\mathfrak{K}_n(\mathbf{x}^{(q)},\mathbf{x}^{(q)}) = [\mathfrak{K}_n(x_{n+i},x_{n+j})]_{i,j=1,\dots,q}$ and $\Phi_2(\cdot;\Sigma)$ is the bivariate centered Normal distribution with covariance matrix Σ.

In the case of conservative estimates with high level α, each term of equation (7) does not contribute equally to the expected distance in measure, as observed in proposition 2. It is thus reasonable to consider the *Type II criterion*, based on the Type II uncertainty

(definition 1), defined as follows.

$$
\begin{aligned}
J_n^{\mathrm{T2}}(\mathbf{x}^{(q)};\rho_n^\alpha) &= \mathbb{E}_{n,\mathbf{x}^{(q)}}\left[H_{n+q}^{\mathrm{T2}}(\rho_n^\alpha)\right] \\
&= \mathbb{E}_{n,\mathbf{x}^{(q)}}\left[G_n^{(2)}(Q_{n+q,\rho_n^\alpha})\right], \qquad \text{for } \mathbf{x}^{(q)} \in \mathbb{X}^q.
\end{aligned} \tag{12}
$$

Proposition 4 (Type II criterion)**.** *In the case* $\Gamma^* = \{x \in \mathbb{X} : f(x) \in T\}$ *with* $T = [t, +\infty)$*, where* $t \in \mathbb{R}$ *is a fixed threshold, the criterion* $J_n^{\mathrm{T2}}(\cdot;\rho_n^\alpha)$ *can be expanded in closed-form as*

$$
\begin{aligned}
J_n^{\mathrm{T2}}(\mathbf{x}^{(q)};\rho_n^\alpha) &= \mathbb{E}_{n,\mathbf{x}^{(q)}}\left[G_n^{(2)}(Q_{n+q,\rho_n^\alpha})\right] \\
&= \int_{\mathbb{X}} \Phi_2\left(\begin{pmatrix} a_{n+q}(u) \\ \Phi^{-1}(\rho_n^\alpha) - a_{n+q}(u) \end{pmatrix}; \begin{pmatrix} 1+\gamma_{n+q}(u) & -\gamma_{n+q}(u) \\ -\gamma_{n+q}(u) & \gamma_{n+q}(u) \end{pmatrix}\right) d\mu(u).
\end{aligned} \tag{13}
$$

4.2 Implementation details

Propositions 3 and 4 provide closed-form expressions for the criteria, however their computation requires numerical approximations. In particular, the evaluation of J_n and J_n^{T2} require the computation of an integral over $\mathbb{X}$ with respect to μ. We compute this integral with an importance sampling Monte Carlo procedure, where the evaluation points for the integrand are chosen with space filling designs, such as Sobol' sequence or uniform sampling. The Monte Carlo weights were chosen uniform in the applications presented in the next sections. We exploit the kriging update formulas (Chevalier et al., 2014b; Emery, 2009) for faster updates of the posterior mean and covariance when new evaluations are added. A sequential strategy then adds new evaluations by minimizing the criteria introduced above. We use the genetic algorithm using derivatives of Mebane and Sekhon (2011) to solve the optimization problem.

The strategies are implemented in the **R** programming language (R Core Team, 2016), with the packages `DiceKriging` (Roustant et al., 2012) for Gaussian modelling, `KrigInv` (Chevalier et al., 2014c) for already existing sampling criterion, `rgenoud` for the optimization routine and `anMC` (Azzimonti and Ginsbourger, 2017) to compute the conservative estimates. The current implementation only allows a homogeneous noise variance $\tau(x) \equiv \tau \in \mathbb{R}$. See **??**, supplementary materials, for a full description.

Table 2: Strategies implemented in the test cases.

Strategy	Criterion	Parameters
Benchmark 1	IMSE	
Benchmark 2	tIMSE	target=t
A	$J_n(\cdot;\rho_n)$	$\rho_n = 0.5$
B	$J_n(\cdot;\rho_n^\alpha)$	$\alpha = 0.95$
C	$J_n^{\mathrm{T2}}(\cdot;\rho_n^\alpha)$	$\alpha = 0.95$

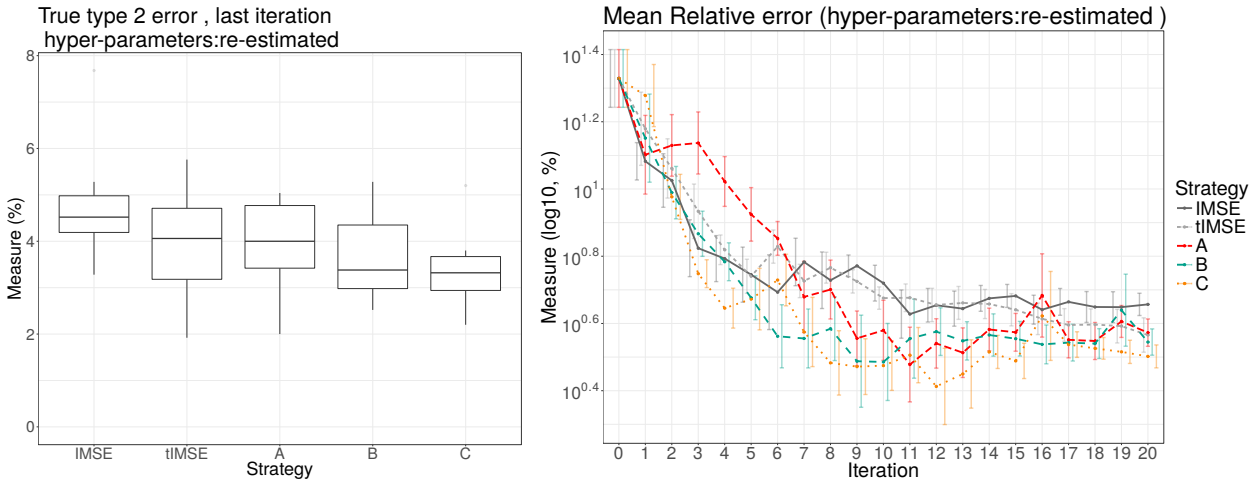
5 Results on reliability engineering test case

In this section we review the test case introduced in figure 1b and apply the sequential strategies from the previous section to reduce the uncertainty on the conservative estimate.

Recall that the object of interest is the excursion set of k-eff below $t = 0.92$, equation (2), which represents the set of safe configurations for a nuclear storage facility. One evaluation of the function k-eff is an expensive computer experiment and we aim to provide an estimate for Γ^* from few evaluations of k-eff and to quantify its uncertainty. The true data result from a MCMC simulation and have a heterogeneous noise variance. Here we consider the k-eff function in figure 1 obtained from 50×50 evaluations of k-eff smoothed with a GP model that accounts for the true noise variance. In what follows k-eff refers to this posterior mean given 2500 evaluations.

We consider a GP model with covariance function Matérn ($\nu = 5/2$, see, e.g., Rasmussen and Williams, 2006, Chapter 4, for details on the parametrization) and homogeneous noise variance estimated from the data. The initial DoE is a Latin hypercube sample design with $n_0 = 15$ function evaluations at the points plotted as triangles in figure 1a. We compare the strategies in table 2 on $m_{\mathrm{doe}} = 10$ different initial DoEs of size $n_0 = 15$, obtained with the function `optimumLHS` from the package `lhs` in **R**. The design of experiments in figure 1a is one of those 10. The covariance hyper-parameters and the noise variance are estimated with maximum likelihood from each initial DoE.

We now test how to adaptively reduce the uncertainty on the estimate with the strategies in table 2. We run $n = 20$ iteration of each strategy and at each step we select a batch



(a) True type II error, last iteration.

(b) Relative volume error versus iteration number.

Figure 4: Nuclear criticality safety test case, randomized initial DoEs.

of $q = 3$ new points where k-eff is evaluated. The covariance hyper-parameters are re-estimated at each iteration. The conservative estimates are computed with the Lebesgue measure μ on $\mathbb{X}$, figure 1b shows $\mathrm{CE}_{\alpha,15}$, $\alpha = 0.95$. Figure 1c shows the coverage function of Γ obtained after 75 function evaluations at locations selected with Strategy C and $\mathrm{CE}_{\alpha,75}$.

Figure 4a shows a comparison of the type II error at the last iteration, i.e. after 75 evaluations of the function, for each initial DoE and each strategy. Strategy C achieves a median type II error 27% lower than IMSE. Strategy B median type II error is 25% lower than IMSE and strategy A's 12% lower than IMSE.

Figure 4b shows the relative volume error as a function of the iteration number for strategies IMSE, tIMSE, A, B, C. The relative volume error is computed by comparing the conservative estimate with a ground truth for Γ^* obtained from evaluations of k-eff on a grid 50×50. The volume of Γ^* computed with Monte Carlo integration from this grid of evaluations is 88.16%. All strategies show a strong decrease in relative volume error in the first 10 iteration, i.e. until 30 evaluations of k-eff are added. In particular strategies B, C show the strongest decline in error in the first 5 iterations. Overall, strategy C, the minimization of the expected type II error, seem to provide the best uncertainty reductions both in terms of relative volume error and in terms of type II error.

Table 3: Test cases parameter choices.

Test case	d	covariance parameters	m_{doe}	n_{init}
GP	2	$\nu = 3/2$, $\theta = [0.2, 0.2]^T$, $\sigma^2 = 1$	10	3
GP	5	$\nu = 3/2$, $\theta = [0.2, 0.2, 0.2, 0.2, 0.2]^T$, $\sigma^2 = 1$	10	6

6 Numerical benchmarks

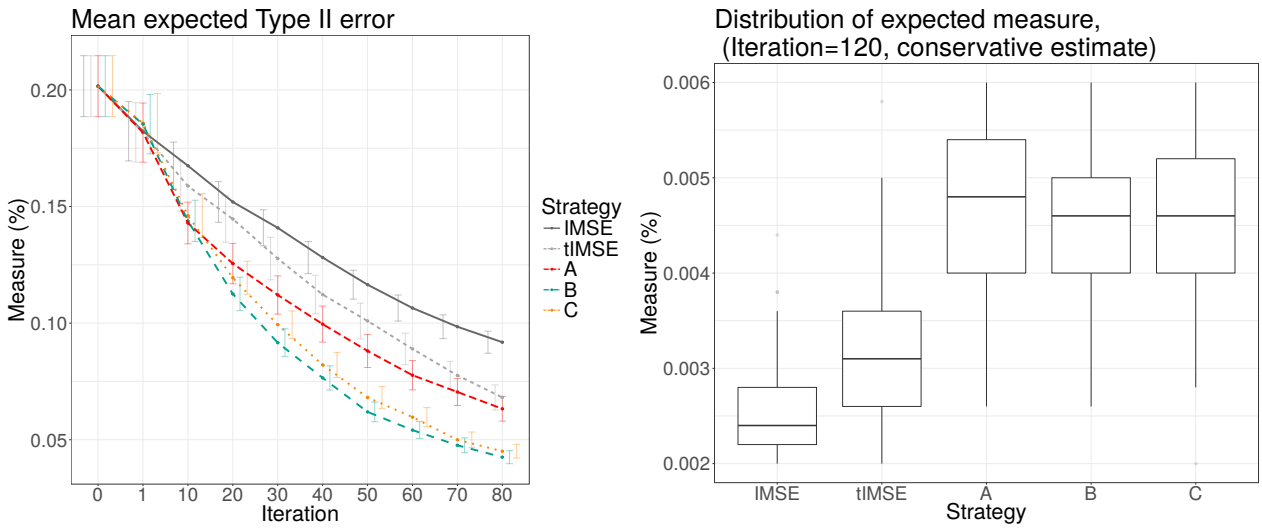
In this section we develop a benchmark study with Gaussian process realizations to study the different behavior of the proposed strategies. We consider two cases with the following shared setup. The input space is the unit hypercube $\mathbb{X} = [0, 1]^d$, for $d = 2, 5$ and $(Z_x)_{x \in \mathbb{X}} \sim GP(\mathfrak{m}, \mathfrak{K})$ with constant prior mean $\mathfrak{m} \equiv 0$ and tensor product Matérn covariance function with known hyper-parameters fixed as in table 3. The noise variance here is constant and equal to zero. The objective is a conservative estimate at level $\alpha = 0.95$ for $\Gamma = \{x \in \mathbb{X} : Z_x \geq 1\}$ and μ is the Lebesgue measure on $\mathbb{X}$. We test the strategies in table 2.

We consider an initial design of experiments $\mathbf{X}_{n_{\text{init}}}$, obtained with the function `optimumLHS` from the package `lhs` and we simulate the field at $\mathbf{X}_{n_{\text{init}}}$. The size n_{init} (see table 3) is chosen small to highlight the differences between the sequential strategies. We select the next evaluations by minimizing each sampling criteria detailed in table 2. Each strategy is run for $n = 80$ ($n = 120$ if $d = 5$) iterations, updating the model with $q = 1$ new evaluations at each step. We consider m_{doe} different initial design of experiments and, for each design, we replicate the procedure 10 times with different initial values $Z_{\mathbf{X}_{n_{\text{init}}}}$.

We evaluate the strategies by looking at the type I and type II errors for Q_{n,ρ_n^α}, defined in section 3.2, and by computing the measure $\mu(Q_{n,\rho_n^\alpha})$. We report mean and median result for each initial design. Expected type I error does not vary much among the different strategies as it is controlled by the condition defining conservative estimate, as shown in section 3. Type I error and total computing time are reported in supplementary material.

6.1 Dimension 2 GP realizations

Figure 5a shows the expected type II error at selected iteration numbers averaged across different initial DoE. This quantity decreases for all strategies, however strategy B and



(a) Mean type II error for Q_{n,ρ_n^α} across different initial DoE, $n = 80$ iterations.

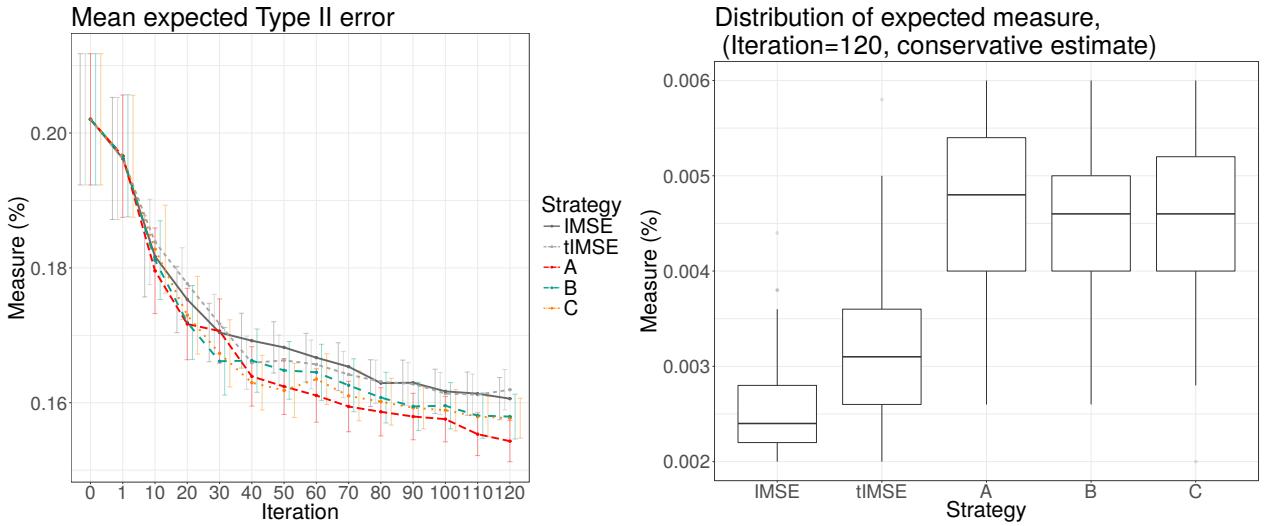
(b) Measure $\mu(Q_{n,\rho_n^\alpha})$ across different initial DoE, after $n = 80$ iterations.

Figure 5: Gaussian process realizations test case in dimension 2.

C outperform the others. The distribution of the expected volume $\mathbb{E}_n[\mu(Q_{n,\rho_n^\alpha})]$ after $n = 80$ new evaluations, is shown in figure 5b. The strategies A, B, C all provide better uncertainty reduction for conservative estimates than a standard IMSE strategy or than a tIMSE strategy. In particular strategy C has the lowest mean type 2 error while at the same time providing an estimate with the largest measure, thus yielding a conservative set likely to be included in Γ^* and, at the same time, not trivial. All estimates, however, are very conservative: the final median ratio between the expected type I error and the estimate's volume is 0.016%, much smaller than the upper bound $1 - \alpha = 5\%$ computed in proposition 2. On the other hand, the median ratio between the expected type II error and the volume at the last iteration is between 31% (C) and 143% (IMSE).

6.2 Dimension 5 GP realizations

Figures 6a and 6b shows the mean expected type II error over selected iterations and the expected measure $\mathbb{E}_n[\mu(Q_{n,\rho_n^\alpha})]$ after 120 iterations of each strategy. Strategies A, B, C provide better uncertainty reduction for conservative estimates than IMSE or tIMSE. Strategies A and C provide a faster reduction of the type II error and a smaller final mean value



(a) Mean type II error for Q_{n,ρ_n^α} across different initial DoE, $n = 120$ iterations.

(b) Measure $\mu(Q_{n,\rho_n^\alpha})$ across different initial DoE, after $n = 120$ iterations.

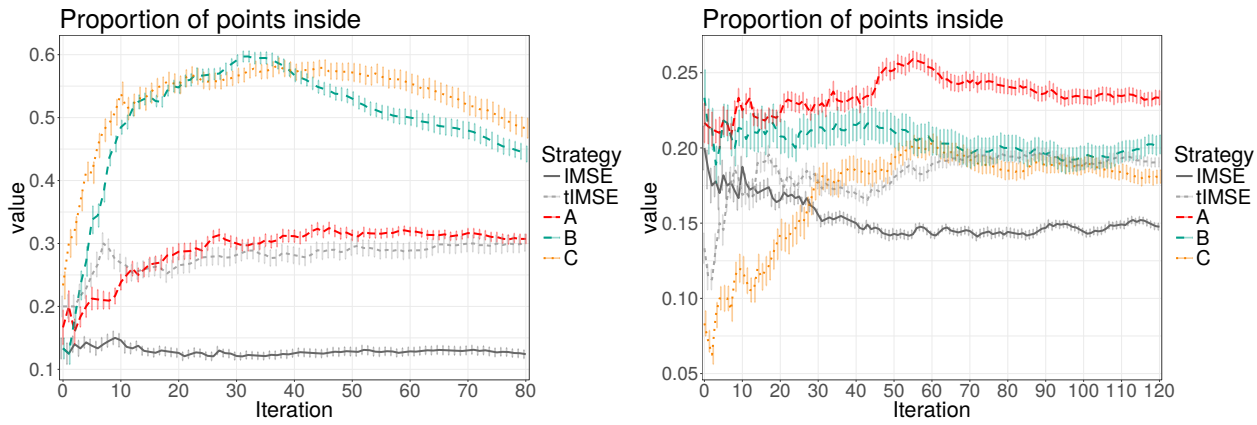
Figure 6: Gaussian process realizations test case in dimension 5.

than the others with strategy A obtaining a slightly higher median value for the expected measure at iteration 120. Also in this case, even if the iteration number is higher, the final estimates provided by all methods are very conservative. The median ratio over all DoEs and all replications between the expected type I error and volume is 0.02%, much smaller than the upper bound 5%. The expected type II error is instead 3 orders of magnitude larger than the estimate's volume. This indicates that we have only recovered a small portion of the true set Γ^* and this estimate is very conservative.

6.3 Model-free comparison of strategies

The metrics presented in the previous sections are based on the GP model. In this section we compare the strategies with a simpler metric independent from the underlying model.

We consider the number of evaluation points that are selected inside and outside the excursion set. At each iteration i, this quantity is computed as $\frac{\#\{j:y_j \geq t,\ j=1,\dots,n_i\}}{n_i}$, where n_i is the total number of points at iteration i and $y_1, \dots, y_{n_i}$ are the evaluations. Figure 7 shows the proportion of points inside the excursion set at each iteration for the two GP test cases. Strategy IMSE is a space filling strategy therefore the proportion of points



(a) Test case $d = 2$.

(b) Test case $d = 5$.

Figure 7: GP realizations. Average number of design points inside the excursion region.

inside the excursion reflects the volume of excursion. Strategies A, tIMSE are adaptive to the problem of estimating an excursion set, however they are not adaptive for conservative estimation, as such they tend to select points around the boundary of Γ and not inside. Strategies B, C instead select more points inside the excursion leading to a good trade-off between a good global approximation of the set and a good approximation of the boundary. These observations are reflected in two dimensions, figure 7a, by the proportion of points inside. This metric in the five dimensional test case, figure 7b, differentiates IMSE from the other strategies, however indicates little differences between adaptive strategies.

7 Discussion

In this paper we introduced sequential uncertainty reduction strategies for conservative estimates. This type of set estimates proved to be useful in reliability engineering, however they could be of interest in all situations where practitioners aim at controlling the overestimation of the set. The estimator CE, however, is based on a global quantity and depends on the quality of the underlying GP model. Under the model, conservative estimates control, by definition, the false positive or type I error. If the GP model is not reliable then such estimates are not necessarily conservative. For a fixed model, increasing the level of confidence might mitigate this problem. We presented test cases with fixed

$\alpha = 0.95$, however testing different levels, e.g. $\alpha = 0.99, 0.995$, and comparing the results is a good practice. The computation of the estimator CE requires the approximation of the exceedance probability of a Gaussian process. This is currently achieved with a discrete approximation, however continuous approximations might be more effective.

The sequential strategies proposed here provide a way to reduce the uncertainty on conservative estimates by adding new function evaluations. The numerical studies presented showed that adapted strategies provide a better uncertainty reduction that generic strategies. In particular, strategy C, i.e. the criterion $J_n^{\mathrm{T}2}(\cdot; \rho_n^\alpha)$, resulted among the best criteria in terms of Type 2 uncertainty and relative volume error in all test cases. In this work we mainly focused on showing the differences between the strategy with a-posteriori measures of uncertainty. Nonetheless the expected type I and II errors could be used to provide stopping criteria for the sequential strategies. Further studies on those quantities could lead to a better understanding of their the limit behavior as n increases.

The strategies proposed in this work focus on reducing the uncertainty on conservative estimates. This objective does not necessarily lead to better overall models for the function or to good covariance hyper-parameters estimation. The sequential behavior of hyper-parameters maximum likelihood estimators under SUR strategies needs to be studied in more details. See supplementary material for a preliminary study on this aspect.

References

Arnaud, A., Bect, J., Couplet, M., Pasanisi, A., and Vazquez, E. (2010). Évaluation d'un risque d'inondation fluviale par planification séquentielle d'expériences. In *42èmes Journées de Statistique de la SFdS*.

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

Azzimonti, D. and Ginsbourger, D. (2017). Estimating orthant probabilities of high dimensional gaussian vectors with an application to set estimation. *Journal of Computational and Graphical Statistics*.

Bayarri, M. J., Berger, J. O., Calder, E. S., Dalbey, K., Lunagomez, S., Patra, A. K., Pitman, E. B., Spiller, E. T., and Wolpert, R. L. (2009). Using statistical and computer models to quantify volcanic hazards. *Technometrics*, 51(4):402–413.

Bect, J., Bachoc, F., and Ginsbourger, D. (2017). A supermartingale approach to Gaussian process based sequential design of experiments. *arXiv:1608.01118 [stat.ML]*.

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.

Bolin, D. and Lindgren, F. (2015). Excursion and contour uncertainty regions for latent Gaussian models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 77(1):85–106.

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

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

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. In Uciński, D., Atkinson, A., and Patan, C., editors, *mODa 10 – Advances in Model-Oriented Design and Analysis*. Physica-Verlag HD.

Chevalier, C., Ginsbourger, D., and Emery, X. (2014b). Corrected kriging update formulae for batch-sequential data assimilation. In *Mathematics of Planet Earth*, Lecture Notes in Earth System Sciences, pages 119–122. Springer Berlin Heidelberg.

Chevalier, C., Picheny, V., and Ginsbourger, D. (2014c). The KrigInv package: An efficient and user-friendly R implementation of kriging-based inversion algorithms. *Comput. Statist. Data Anal.*, 71:1021–1034.

Chilès, J.-P. and Delfiner, P. (2012). *Geostatistics: Modeling Spatial Uncertainty, Second Edition*. Wiley, New York.

Emery, X. (2009). The kriging update equations and their application to the selection of neighboring data. *Comput. Geosci.*, 13(3):269–280.

Fleuret, F. and Geman, D. (1999). Graded learning for object detection. In *Proceedings of the workshop on Statistical and Computational Theories of Vision of the IEEE international conference on Computer Vision and Pattern Recognition (CVPR/SCTV)*, volume 2. Citeseer.

French, J. P. and Sain, S. R. (2013). Spatio-temporal exceedance locations and confidence regions. *Ann. Appl. Stat.*, 7(3):1421–1449.

Ginsbourger, D. and Le Riche, R. (2010). Towards gaussian process-based optimization with finite time horizon. In *mODa 9 – Advances in Model-Oriented Design and Analysis*, pages 89–96. Springer.

González, J., Osborne, M., and Lawrence, N. D. (2016). GLASSES: Relieving The Myopia Of Bayesian Optimisation. In *19th International Conference on Artificial Intelligence and Statistics*, pages 790–799.

Mebane, W. R. J. and Sekhon, J. S. (2011). Genetic optimization using derivatives: The rgenoud package for R. *Journal of Statistical Software*, 42(11):1–26.

Molchanov, I. (2005). *Theory of Random Sets*. Springer, London.

Osborne, M. A., Garnett, R., and Roberts, S. J. (2009). Gaussian processes for global optimization. In *3rd international conference on learning and intelligent optimization (LION3)*, pages 1–15.

Picheny, V., Ginsbourger, D., O., R., Haftka, R., and Kim, N. (2010). Adaptive designs of experiments for accurate approximation of a target region. *ASME. J. Mech. Des.*, 132(7):071008–071008–9.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian processes for machine learning.* MIT Press.

Roustant, O., Ginsbourger, D., and Deville, Y. (2012). DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodelling and optimization. *Journal of Statistical Software*, 51(1):1–55.

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

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

Vazquez, E. and Bect, J. (2009). A sequential bayesian algorithm to estimate a probability of failure. *IFAC Proceedings Volumes*, 42(10):546–550.

Vorob'ev, O. Y. (1984). Srednemernoje modelirovanie (mean-measure modelling). *Nauka, Moscow*, In Russian.

Wheeler, M. W., Dunson, D. B., Pandalai, S. P., Baker, B. A., and Herring, A. H. (2014). Mechanistic hierarchical gaussian processes. *J. Amer. Statist. Assoc.*, 109(507):894–904.

A Sequential strategies

Proof of proposition 3. Recall that

$$\mathbb{E}_{n,\mathbf{x}^{(q)}}\left[\mu(\Gamma\Delta Q_{n+q,\rho_n})\right] = \mathbb{E}_{n,\mathbf{x}^{(q)}}[\underbrace{\mu(Q_{n+q,\rho_n}\setminus\Gamma)}_{=G^{(1)}_{n+q}(\rho_n)}] + \mathbb{E}_{n,\mathbf{x}^{(q)}}[\underbrace{\mu(\Gamma\setminus Q_{n+q,\rho_n})}_{=G^{(2)}_{n+q}(\rho_n)}]. \tag{14}$$

From the definitions of $G^{(1)}_{n+q}, G^{(2)}_{n+q}$ and the law of total expectation we have

$$\mathbb{E}_{n,\mathbf{x}^{(q)}}\left[G^{(2)}_{n+q}(\rho_n)\right] = \int_{\mathbb{X}} \mathbb{E}_n[p_{n+q}(u)\mathbb{1}_{\{p_{n+q}(u)<\rho_n\}}]d\mu(u) \tag{15}$$

$$\begin{aligned}\mathbb{E}_{n,\mathbf{x}^{(q)}}\left[G^{(1)}_{n+q}(\rho_n)\right] &= \int_{\mathbb{X}} \mathbb{E}_n[\mathbb{1}_{\{p_{n+q}(u)\geq\rho_n\}}(1-p_{n+q}(u))]d\mu(u) \\ &= \int_{\mathbb{X}} \left(\mathbb{E}_n[\mathbb{1}_{\{p_{n+q}(u)\geq\rho_n\}}] - \mathbb{E}_n[\mathbb{1}_{\{p_{n+q}(u)\geq\rho_n\}}p_{n+q}(u))]\right) d\mu(u) \\ &= \int_{\mathbb{X}} \left(\mathbb{E}_n[\mathbb{1}_{\{p_{n+q}(u)\geq\rho_n\}}] - p_n(u)\right) d\mu(u) + \mathbb{E}_{n,\mathbf{x}^{(q)}}\left[G^{(2)}_{n+q}(\rho_n)\right]\end{aligned} \tag{16}$$

Notice that, for each $x \in \mathbb{X}$, the coverage function $p_{n+q,\mathbf{x}^{(q)}}$ can be written as

$$p_{n+q,\mathbf{x}^{(q)}}(x) = \Phi\left(a_{n+q}(x) + \mathbf{b}_{n+q}^T Y_q\right), \tag{17}$$

where $a_{n+q}, \mathbf{b}_{n+q}$ are defined in equation equation (11) and $Y_q \sim N_q(0, K_q)$ is a q dimensional normal random vector. The first part of equation (15) is

$$\begin{aligned}\mathbb{E}_n[\mathbb{1}_{p_{n+q}(u)\geq\rho_n}] &= P_n(p_{n+q}(u) \geq \rho_n) = P_n(\mathbf{b}_{n+q}^T(u)Y_q \geq \Phi^{-1}(\rho_n) - a_{n+q}(u)) \\ &= \Phi\left(\frac{a_{n+q}(u) - \Phi^{-1}(\rho_n)}{\sqrt{\mathbf{b}_{n+q}^T(u)K_q\mathbf{b}_{n+q}(u)}}\right)\end{aligned} \tag{18}$$

where the second equality follows from equation (17) and the third from $Y_q \sim N(0, K_q)$. Moreover

$$\begin{aligned}\mathbb{E}_n[\mathbb{1}_{\{p_{n+q}(u)<\rho_n\}}p_{n+q}(u)] &= \int \Phi\left(a_{n+q}(u) + \mathbf{b}_{n+q}^T(u)y\right)\mathbb{1}_{\{b_{n+q}^T(u)y<\Phi^{-1}(\rho_n)-a_{n+q}(u)\}}\Psi(y) \\ &= \int P(N_1 \leq a_{n+q}(u) + \mathbf{b}_{n+q}^T(u)y)\mathbb{1}_{\{b_{n+q}^T(u)y<\Phi^{-1}(\rho_n)-a_{n+q}(u)\}}\Psi(y) \\ &= \mathbb{E}\left[P(N_1 \leq a_{n+q}(u) + \mathbf{b}_{n+q}^T y,\ \mathbf{b}_{n+q}^T(u)y < \Phi^{-1}(\rho_n) - a_{n+q}(u))\right] \\ &= \Phi_2\left(\begin{pmatrix} a_{n+q}(u) \\ \Phi^{-1}(\rho_n) - a_{n+q}(u)\end{pmatrix}; \begin{pmatrix} 1+\gamma_{n+q}(u) & -\gamma_{n+q}(u) \\ -\gamma_{n+q}(u) & \gamma_{n+q}(u)\end{pmatrix}\right).\end{aligned} \tag{19}$$

where Ψ is the p.d.f. of Y_q, $N_1 \sim N(0,1)$ and $\Phi_2(a;\Sigma)$ is the c.d.f. of a centered bivariate Gaussian random variable with covariance Σ evaluated at a. By equations (14) to (16), (18) and (19) we obtain equation (10).

□

Proof of proposition 4. The proof follows from equations (15) and (19). □

B Properties of conservative estimates

In the following, let us denote by $(\Omega, \mathcal{F}, P)$ a probability space.

Proof of proposition 1. We want to show that the set Q_ρ satisfies

$$\mathbb{E}\left[\mu(Q_\rho \Delta \Gamma)\right] \leq \mathbb{E}\left[\mu(M \Delta \Gamma)\right], \tag{20}$$

for each measurable set M such that $\mu(M) = \mu(Q_\rho)$. Let us consider a measurable set M such that $\mu(M) = \mu(Q_\rho)$. For each $\omega \in \Omega$, we have

$$\mu(M\Delta\Gamma(\omega)) - \mu(Q_\rho\Delta\Gamma(\omega)) = 2\Big(\mu(\Gamma(\omega) \cap (Q_\rho \setminus M)) - \mu(\Gamma(\omega) \cap (M \setminus Q_\rho))\Big) + \mu(Q_\rho^C) - \mu(M^C).$$

By applying the expectation on both sides and by remembering that $\mu(Q_\rho^C) = \mu(M^C)$ we obtain

$$\begin{aligned}\mathbb{E}\left[\mu(M\Delta\Gamma) - \mu(Q_\rho\Delta\Gamma)\right] &= \mathbb{E}\left[2\Big(\mu(\Gamma \cap (Q_\rho \setminus M)) - \mu(\Gamma \cap (M \setminus Q_\rho))\Big)\right] \\ &= 2\int_{Q_\rho\setminus M} p_\Gamma(u)d\mu(u) - 2\int_{M\setminus Q_\rho} p_\Gamma(u)d\mu(u),\end{aligned}$$

where the second equality comes from the definition of Q_ρ. Moreover, since $p_\Gamma(x) \geq \rho$ for $x \in Q_\rho \setminus M$ and $p_\Gamma(x) \leq \rho$ for $x \in M \setminus Q_\rho$ we have

$$\begin{aligned}2\left[\int_{Q_\rho\setminus M} p_\Gamma(u)d\mu(u) - \int_{M\setminus Q_\rho} p_\Gamma(u)d\mu(u)\right] &\geq 2\rho[\mu(Q_\rho \setminus M) - \mu(M \setminus Q_\rho)] \\ &= 2\rho[\mu(Q_\rho) - \mu(M)] = 0,\end{aligned}$$

which shows that Q_ρ verifies equation equation (20). □

Proof of proposition 2. Notice that for all $\omega \in \Omega$ such that $Q_{n,\rho_n^\alpha} \subset \Gamma(\omega)$, we have $G_n^{(1)}(\omega) = 0$. By applying the law of total expectation we obtain

$$\begin{aligned}\mathbb{E}_n[G_n^{(1)}] &= \mathbb{E}_n[G_n^{(1)} \mid Q_{n,\rho_n^\alpha} \subset \Gamma]P(Q_{n,\rho_n^\alpha} \subset \Gamma) \\ &\quad + \mathbb{E}_n[G_n^{(1)} \mid Q_{n,\rho_n^\alpha} \setminus \Gamma \neq \emptyset](1 - P(Q_{n,\rho_n^\alpha} \subset \Gamma)) \\ &\leq 0 + \mathbb{E}_n[G_n^{(1)} \mid Q_{n,\rho_n^\alpha} \setminus \Gamma \neq \emptyset](1-\alpha) \leq \mu(Q_{n,\rho_n^\alpha})(1-\alpha).\end{aligned}$$

□