



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-01379642v4

HAL Id: hal-01379642

<https://hal.science/hal-01379642v4>

Preprint submitted on 30 Nov 2018 (v4), 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 the problem of estimating the set of all inputs that leads a system to some particular behavior. The system is modeled with an expensive-to-evaluate function, such as a computer experiment, and we are interested in its excursion set, i.e. the set of points where the function takes values above or below some prescribed threshold. The objective function is emulated with Gaussian Process (GP) models based on an initial design of experiments enriched with evaluation results at (batch-) sequentially determined input points. The GP model provides conservative estimates for the excursion set, which control false positives while minimizing false negatives. We introduce adaptive strategies that sequentially select 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. Tractable formulae for the conservative criteria are derived which allow more convenient optimization. The method is benchmarked on random functions generated under the model assumptions in two and five dimensions and 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 model-free indicators.

Keywords: Batch sequential strategies; Conservative estimates; 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. All authors thank Michael McCourt for his question on model-free comparisons at the NIPS 2017 workshop on Bayesian Optimization.

[†]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), CentraleSupélec, 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). Here we consider a system modeled as a continuous, expensive-to-evaluate function $f : \mathbb{X} \rightarrow \mathbb{R}$, where \mathbb{X} is a compact subset of \mathbb{R}^d . Section 1.1 shows an example of such systems. Given few evaluations of f and a fixed closed set $T \subset \mathbb{R}$, we are interested in estimating

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

As an example, in the motivating test case in section 1.1, Γ^* represents the *safe region*, i.e. all values of the physical parameters that lead the system of interest to a subcritical response, in this case $T = (-\infty, t]$ with $t \in \mathbb{R}$.

There is much heterogeneity in the literature on how to name Γ^* . Here we follow Adler and Taylor (2007) and we call Γ^* an *excursion set*. If $T = [t, +\infty)$, with $t \in \mathbb{R}$, Γ^* is often also called excursion set above t (see, e.g., Azaïs and Wschebor, 2009; Bolin and Lindgren, 2015), but also level set (Berkenkamp et al., 2017). If $T = (-\infty, t]$ the set is sometimes referenced as sojourn set (Spodarev, 2014) or sublevel set (Gotovos et al., 2013). Our work is introduced for the case $T = [t, +\infty)$, however it is also applied to the case $T = (-\infty, t]$.

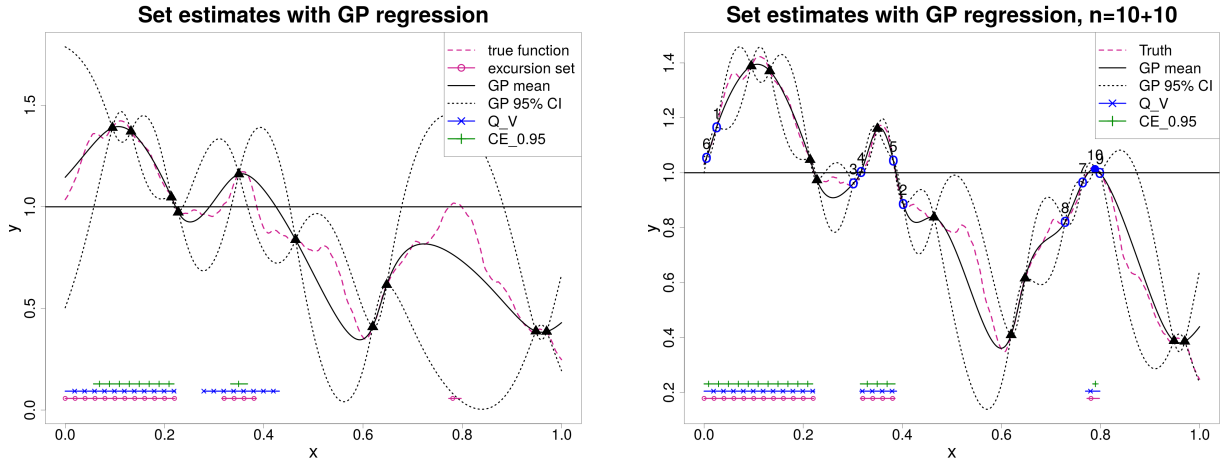
Throughout the article, we model f as a realization of a Gaussian process (GP) and, following Sacks et al. (1989); Santner et al. (2003), we emulate f with the posterior GP distribution given the available function evaluations. The posterior GP distribution can be used as building block for different estimates of Γ^* , see, e.g. Azzimonti (2016).

Consider now a generic estimate $\tilde{\Gamma}$ for Γ^* and denote with $\text{vol}(A)$ the volume of $A \subset \mathbb{X}$. Simple quality indicators for a set estimate are the volume of false positives $\text{vol}(\tilde{\Gamma} \setminus \Gamma^*)$, i.e. the volume of points estimated in the set while actually outside Γ^* , and the volume of false negatives $\text{vol}(\Gamma^* \setminus \tilde{\Gamma})$, i.e. the volume of points estimated not in the excursion set while actually inside. For example, in Chevalier (2013); Chevalier et al. (2014a), Γ^* is estimated with the Vorob'ev expectation, a notion borrowed from random set theory (Molchanov, 2005, Chapter 2), that aims to minimize the overall volume of misclassified points. Fig-

Figure 1a shows an analytical example where the input space is $\mathbb{X} = [0, 1]$ and the function f is generated as a realization of a GP (purple dashed line) with mean zero and Matérn covariance kernel with hyper-parameters $\nu = 3/2$, $l = 0.3$, $\sigma^2 = 0.3$, see, e.g., Rasmussen and Williams (2006), Chapter 4, for details on the parametrization. We build a GP model (black solid line) from $n = 10$ evaluations of f (black triangles) and we estimate Γ^* (purple dotted horizontal line) with $T = [1, +\infty)$. The figure shows the Vorob'ev expectation (Q_V , middle horizontal blue line), an estimate of the true excursion set; a comparison of Q_V with the true excursion set shows that Q_V has volumes of false positive (0.084) and negatives (0.025) of the same order of magnitude.

The Vorob'ev expectation gives a similar importance to false positives and false negatives. However, in a number of applications, the cost of misclassification is not symmetric with higher penalties for false positives, for instance, than for false negatives. Practitioners may hence be interested in set estimates which would *very likely* be included in an excursion set of the form $\Gamma^* = \{x \in \mathbb{X} : f(x) \geq t\}$. Such a 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 conservative estimates which select sets that are deliberately smaller – in volume – than Γ^* and are included in the excursion set with a *large probability*. The empty set trivially satisfies this probabilistic inclusion property, therefore conservative estimates are selected as sets with maximal volume in a family of possible estimates. A conservative estimate thus enforces a low probability of false positives. In a reliability engineering framework, the excursion set can be the set of safe configurations and a conservative estimate aims at selecting a region which is included in the safe set. Figure 1a shows a conservative estimate at level $\alpha = 0.95$ ($CE_{0.95}$, green top horizontal line). In this example, $CE_{0.95}$ has a false positive volume equal to zero, however a much higher volume of false negative (0.121) than the Vorob'ev expectation. For a fixed threshold t , the excursion set above t is trivially the complement of the sojourn set below t . Note, however, that this does not hold for their respective set estimates. In particular, the conservative estimate of an excursion set is not the complement of the conservative estimate of the corresponding sojourn set due to the probabilistic inclusion property.

French and Sain (2013) and Bolin and Lindgren (2015) proposed an approach to com-



(a) Initial DoE: maximin LHS, $n = 10$.

(b) Adaptive DoE: strategy $T2$, 10 new points.

Figure 1: Example of adaptive DoE. Comparison of Vorob'ev expectation (Q_V) and conservative estimate ($CE_{0.95}$).

pute 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 the problem of sequentially choosing numerical experiments in order to reduce the uncertainty on conservative estimates. Figure 1b shows an example of adaptive DoE where, starting from the initial DoE in figure 1a, 10 additional points are selected with strategy C introduced in section 3.

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 around a certain level of the response variable. Bect et al. (2012) investigated the concept of Stepwise Uncertainty Reduction (SUR) strategies for GP (see also Vazquez and Bect, 2009; Chevalier et al., 2014a; Bect et al., 2017). Existing instances of such strategies, however, do not provide any control on false positives and as such 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. For example, notice how in figure 1b, some points (e.g. numbers 1, 2, 8) are chosen far from the boundary, in order to improve

the confidence on the classification of those regions. Here 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. The adaptive strategies are introduced here in the case of excursion sets above $t \in \mathbb{R}$, however our **R** implementation, available on line, allows also for excursions below t .

1.1 Motivating test case

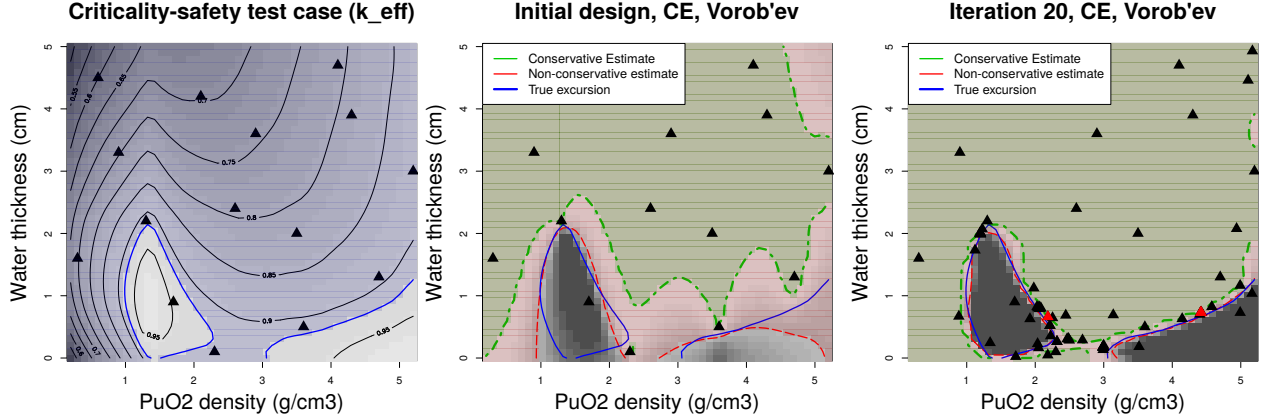
In reliability engineering applications, the set Γ^* in equation (1) often represents safe inputs for a system. In such settings, it is vital to avoid flagging unsafe regions as safe. With a broad use of hypothesis testing terminology we refer to this as type I error or false positive.

Figure 2 shows an example of such reliability engineering applications: a test case from the French Institute for Radiological Protection and Nuclear Safety (IRSN). We briefly outline the test case in this subsection, see section 4 for detailed results. The problem 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 : $\mathbb{X} \rightarrow [0, +\infty)$. In our application $\mathbb{X} = [0.2, 5.2] \times [0, 5]$ 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

$$\Gamma^* = \{(\text{PuO}_2, \text{H}_2\text{O}) \in \mathbb{X} : k\text{-eff}(\text{PuO}_2, \text{H}_2\text{O}) \leq 0.92\}, \quad (2)$$

where the threshold $t = 0.92$ was chosen, for safety reasons, lower than the true critical case ($k\text{-eff} = 1.0$) where an uncontrolled chain reaction occurs. Figure 2a 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 2b 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 $k\text{-eff}$, 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 2c shows that, as more evaluations are



(a) Function k -eff, set of interest Γ^* (shaded blue, $\text{vol}(\Gamma^*)=0.8816 \text{ vol}(\mathbb{X})$) 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 2: 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.

available, conservative and non-conservative estimates both get closer to the true safe 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 3).

In this paper we study conservative estimates well suited to examples such as figure 2 and how to quantify uncertainties on such estimates. The definition of a measure of residual uncertainty allows us to define natural SUR strategies, where new evaluations are chosen by optimizing a specific criterion. The evaluation of such criterion is potentially challenging, however, here we provide tractable formulae 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. In particular, section 2.1 reviews set estimates preliminary to this work and section 2.2 recalls the concept of SUR strategies. In section 3.1 we introduce the metrics used to quantify the uncertainty on such estimates. In section 3.2, we detail the

proposed sequential strategies, we derive closed-form formulae for the associated criteria and we illustrate their implementation. Section 4 presents the results obtained on the IRSN test case and section 5 shows a benchmark study on Gaussian process realizations. In appendix B we provide more properties for conservative estimates that further justify the choices made in section 2.1. All proofs are in appendices A and B.

2 Background

Let us consider n observations of the function f , possibly tampered by measurement noise

$$z_i = f(x_i) + \tau(x_i)\epsilon_i \quad x_i \in \mathbb{X}, \quad i = 1, \dots, n$$

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 almost surely continuous Gaussian process (GP) $\xi \sim GP(\mathbf{m}, \mathfrak{K})$, with mean function $\mathbf{m}(x) := \mathbb{E}[\xi_x]$ and covariance function $\mathfrak{K}(x, x') := \text{Cov}(\xi_x, \xi_{x'})$, $x, x' \in \mathbb{X}$. With this notation, z_i is a realization of $Z_i = \xi_{x_i} + \tau(x_i)\epsilon$ where $\epsilon \sim N(0, 1)$. For $n > 0$, we denote by $\mathbf{z}_n = (z_1, \dots, z_n) \in \mathbb{R}^n$ the observations at an initial design of experiments (DoE) $\mathbf{X}_n = (x_1, \dots, x_n) \in \mathbb{X}^n$. The posterior distribution of the process is Gaussian with mean and covariance computed as the conditional mean \mathbf{m}_n and conditional covariance \mathfrak{K}_n given the observations, see, e.g., Santner et al. (2003) for closed form formulae.

2.1 Vorob'ev expectation and conservative estimates

The prior distribution on ξ induces a (random) set $\Gamma = \{x \in \mathbb{X} : \xi_x \in T\}$. By using the posterior distribution of ξ , we can provide estimates for Γ^* , see, e.g. Chevalier et al. (2014a); Bolin and Lindgren (2015); Azzimonti (2016) for summaries of different approaches. A central tool for the approach presented here is the *coverage probability function* of a random closed set Γ , defined as

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

In our case we consider the posterior coverage function $p_{\Gamma, n} = p_n$, defined with the posterior probability $P_n(\cdot) = P(\cdot \mid \mathbf{Z}_n = \mathbf{z}_n)$, where $\mathbf{Z}_n = (Z_1, \dots, Z_n)$ and we drop the subscript Γ

as the set is clear from the context. If $T = (-\infty, t]$, then $p_n(x) = \Phi\left(\frac{t - m_n(x)}{s_n(x)}\right)$, where $\Phi(\cdot)$ is the CDF of a standard Normal random variable and $s_n(x) = \sqrt{\widehat{\mathfrak{R}}_n(x, x)}$. The coverage function defines the family of *Vorob'ev quantiles*

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

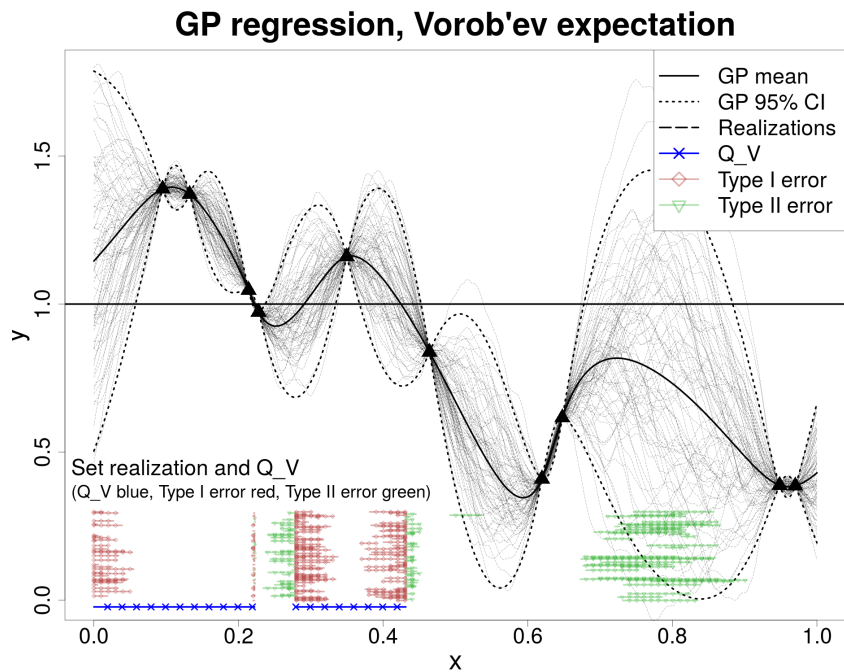
with $\rho \in [0, 1]$. These sets are closed for each $\rho \in [0, 1]$ (see Molchanov, 2005, Proposition 1.34) and form a family of possible estimates parametrized by ρ .

The level ρ can be selected in different ways. The choice $\rho = 0.5$ leads to the *Vorob'ev median*, which is not conservative. The *Vorob'ev expectation* (Vorob'ev, 1984; Molchanov, 2005; Chevalier et al., 2013), as introduced in section 1, relies on the notion of volume. Here we consider a slightly more general form by using a finite measure μ in place of the volume. In the applications presented here, μ is the Lebesgue measure on \mathbb{X} , however the theory holds for more general measures on \mathbb{X} , e.g. a probability distribution on inputs. The Vorob'ev expectation is defined 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 of Γ . The set Q_{n,ρ_V} is also the minimizer of $\mathbb{E}[\mu(\Gamma \Delta M)]^1$ 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 a uniformly weighted combination of the expected measure of false positives ($\mathbb{E}[\mu(M \setminus \Gamma)]$, type I error) and false negatives ($\mathbb{E}[\mu(\Gamma \setminus M)]$, type II error) among sets with measure equal to $\mathbb{E}[\mu(\Gamma)]$. In appendix B.2 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 often called *expected distance in measure*. Chevalier (2013) suggested this distance to adaptively reduce the uncertainty on Vorob'ev expectations. In section 3.1 we adapt it for conservative estimates.

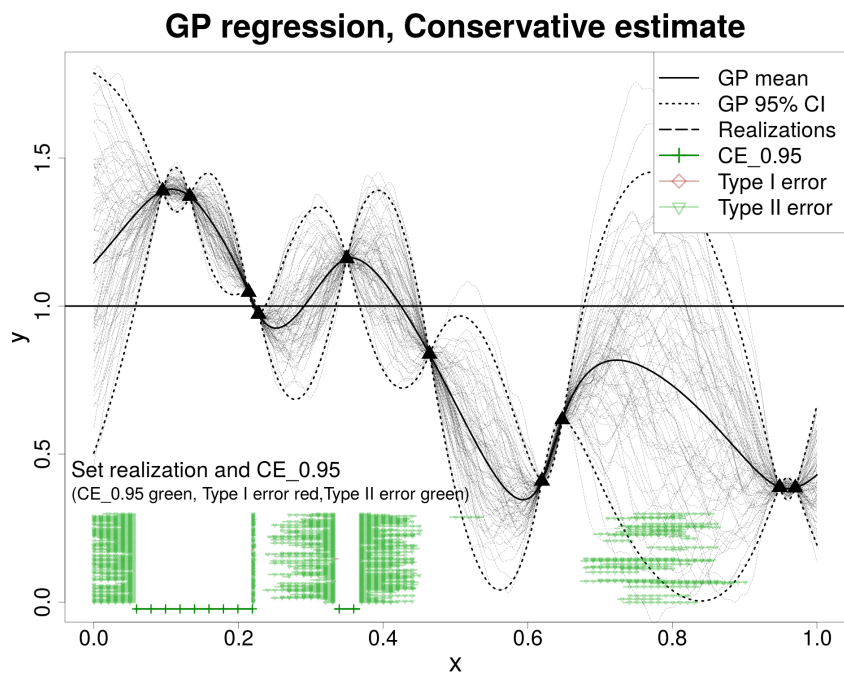
Conservative estimate (Bolin and Lindgren, 2015; French and Sain, 2013) constitutes a different type of set estimate that embed a probabilistic control on false positives in the estimator. Denote with \mathfrak{C} a family of closed subsets in \mathbb{X} . A *conservative estimate at level α* for Γ^*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

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

¹For any A, B , $A \Delta B := A \setminus B \cup B \setminus A$



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



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

Figure 3: 1-dimensional example, $n = 10$ evaluations. Type I (red, diamonds) and Type II (green, triangles) errors for Vorob'ev expectation and conservative estimate.

Table 1: Summary values for example in figure 3, 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
$CE_{0.95}$	0.987	$9.5 \times 10^{-5} \pm 4.7 \times 10^{-4}$	0.187 ± 0.063	0.95

The set $CE_{\alpha,n}$ is therefore a maximal set (according to μ) in the family \mathfrak{C} such that the posterior probability of inclusion is at least α . Here, by following French and Sain (2013); Bolin and Lindgren (2015); Azzimonti and Ginsbourger (2018), we choose \mathfrak{C} as the family of Vorob'ev quantiles $\{Q_\rho : \rho \in [0, 1]\}$. While the concept of probabilistic inclusion might seem unusual at first, conservative estimates are actually linked with the well known concept of confidence regions, as we briefly show in appendix B.1. Note further that the condition $P_n(C \subset \Gamma) = P_n(C \setminus \Gamma = \emptyset) \geq \alpha$ controls the probability of false positives. We can visualize this property on the one dimensional example introduced in figure 1 by empirically estimating the expected measure of false positives, $\mathbb{E}_n[\mu(CE_{0.95} \setminus \Gamma)]$. Figure 3 shows 80 posterior realizations of the GP (light dashed black lines) and for each realization we computed the false positive (type I error, red) and false negatives (type II error, green). Notice how they are symmetrically minimized by the Vorob'ev expectation (figure 3a) while the conservative estimate with $\alpha = 0.95$ (figure 3b) has small false positives and 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, $\hat{P}(Q_\rho \subset \Gamma)$. The Vorob'ev expectation may be closer to the truth than conservative estimates, especially for small DoEs, however $CE_{\alpha,n}$ gives control on the probability of false positives. Table 1 also reports the values for the Vorob'ev quantile $Q_{0.95}$, i.e. a non adaptive high quantile choice for ρ . Note that $P(Q_{0.95} \subset \Gamma) < 0.95$, in fact, the quantile's definition based on the marginal probability $p_n(x) \geq 0.95, x \in \mathbb{X}$, does not imply any statement on the probability of inclusion.

The computation of $CE_{\alpha,n}$ in equation (4) requires to find a set C of maximum measure among sets included in the random set Γ with probability at least α . When \mathfrak{C} is the

family of Vorob'ev quantiles Q_ρ , $\rho \in [0, 1]$, this optimization can be solved with a simple dichotomic search on ρ . See appendix B.2 for more details. If $T = (-\infty, t]$, for $\rho \in [0, 1]$, we approximate $P_n(Q_\rho \subset \Gamma) \approx P_n(\xi_{q_1} \leq t, \dots, \xi_{q_\ell} \leq t)$, where $\{q_1, \dots, q_\ell\} \subset Q_\rho$ is a set of ℓ points in Q_ρ with ℓ large. The probability above is then computed with the computationally efficient integration proposed by Azzimonti and Ginsbourger (2018). The number ℓ is generally chosen as large as the computational budget allows. This technique can also be used for excursion sets above t . An alternative method, not used here, is Monte Carlo with conditional realizations of the field, see, e.g. Azzimonti et al. (2016) for fast approximations of conditional realizations.

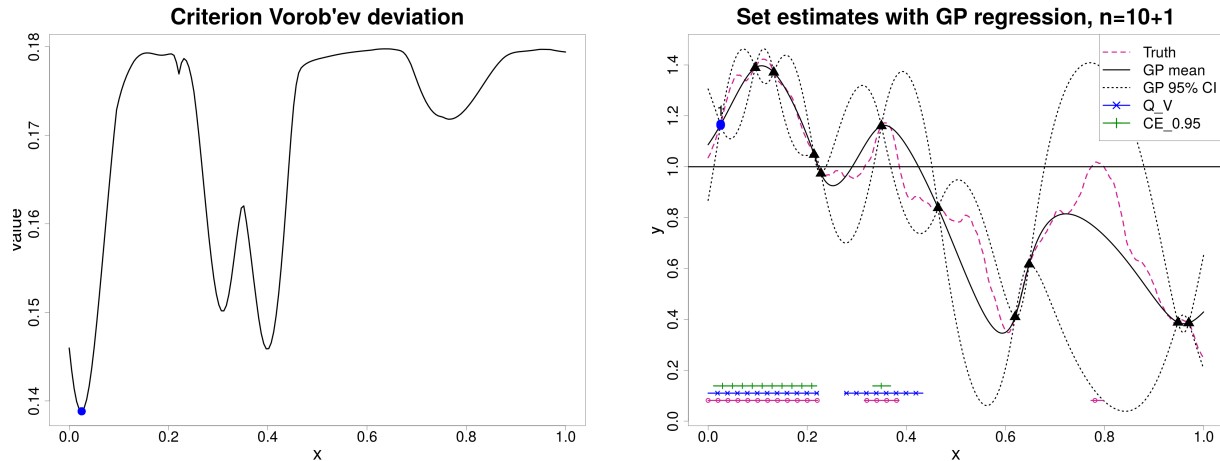
2.2 SUR strategies

Sequential design of experiments adaptively chooses the next evaluation points according to a strategy with the aim of improving the final estimate. We follow the Stepwise Uncertainty Reduction approach (SUR, see, e.g., Fleuret and Geman, 1999; Bect et al., 2012; Chevalier et al., 2014a; Bect et al., 2018) and we select a sequence of points in order to reduce the uncertainty on selected quantities of interest. In the remainder of the paper we consider that the first n points x_1, \dots, x_n and the respective evaluations \mathbf{z}_n are known and we denote by $\mathbb{E}_n[\cdot] = \mathbb{E}[\cdot \mid \mathbf{Z}_n = \mathbf{z}_n]$ the expectation conditional on $\mathbf{Z}_n = \mathbf{z}_n$. We are interested in selecting the next batch of q locations x_{n+1}, \dots, x_{n+q} . In a sequential setting the evaluations at these points are random, therefore we denote by $\mathbb{E}_{n, \mathbf{x}^{(q)}}[\cdot]$ the conditional expectation given the first n evaluations and with the next locations fixed at $\mathbf{x}^{(q)} = (x_{n+1}, \dots, x_{n+q}) \in \mathbb{X}^q$.

For a specific problem, we define a measure of residual uncertainty at step n , denoted by H_n . If the first n locations and evaluations are known, then H_n is a (deterministic) real number quantifying the residual uncertainty on the estimate. Consider the example in figure 1, we can compute, with $n = 10$, the uncertainty defined as $H_n = \mathbb{E}_n[\mu(\Gamma \Delta \text{CE}_{0.95})]$. With numerical integration we find the value $H_{10} = 0.23$. On the other hand, the quantity H_{n+1} , seen from step n , is random because Z_{n+1} is random. The next batch of q locations can then be selected following the principles of a SUR strategy, i.e. by setting

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

a minimizer of the future uncertainty in expectation. For a more complete and theoretical



(a) Criterion J_n . The next point is the minimizer of this function, blue dot.

(b) Adaptive DoE with $n = 11$, the last point (blue dot) is chosen with J_n .

Figure 4: Adaptive DoE with SUR strategy on the example introduced in figure 1.

perspective on SUR strategies see, e.g., Bect et al. (2018) 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. 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. We can build a SUR strategy with the uncertainty $H_n = \mathbb{E}_n[\mu(\Gamma \Delta \text{CE}_{0.95})]$ introduced above. Moreover, the criterion associated with this uncertainty has the remarkable property that it can be computed with semi-analytical formulae, thus making its optimization more convenient. Figure 4a shows the function J_n for $n = 10$ and $q = 1$; the next evaluation x_{11} is chosen as the minimizer of this function. Figure 4b shows the updated GP model, note how $\text{CE}_{0.95}$ now better covers the true set.

The expectation \mathbb{E}_n can only be computed if we know \mathfrak{K} which is often chosen from a parametric family depending on few hyper-parameters, l and σ in the analytical example. In practice, the hyper-parameters are not known and can be estimated with a plug-in or with a fully Bayesian approach. In this work we follow the previous literature on boundary

estimation with GP models (see, e.g. Ranjan et al., 2008; Picheny et al., 2013; Chevalier et al., 2014a; Azzimonti and Ginsbourger, 2018) and we plug-in maximum likelihood estimates for the hyper-parameters. If only few observations are available, a fully Bayesian approach might better capture the overall uncertainty at an additional computational cost.

In the next section we detail two uncertainty functions tailored for conservative estimates and we show how their respective SUR criteria can be computed.

3 SUR strategies for conservative estimates

3.1 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 deviation*, 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 suited for conservative estimates. The idea is to describe the uncertainty by looking at the expected measure of false negatives. In the example of figure 3b, 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 deviation of a quantile $Q_{n,\rho}$ is

$$H_{n,\rho} = \mathbb{E}_n[\mu(\Gamma \Delta Q_{n,\rho})] = \mathbb{E}_n[\mu(Q_{n,\rho} \setminus \Gamma)] + \mathbb{E}_n[\mu(\Gamma \setminus Q_{n,\rho})], \quad \rho \in [0, 1]. \quad (6)$$

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 after n evaluations.

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 (6). We call *Type I* and *Type II* errors given n observations the quantities $\mathbb{E}_n[G_n^{(1)}(\rho_n^\alpha)]$ and $\mathbb{E}_n[G_n^{(2)}(\rho_n^\alpha)]$. 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

$$H_{n,\rho_n^\alpha}^{\Gamma^2} := \mathbb{E}_n[G_n^{(2)}(\rho_n^\alpha)] = \mathbb{E}_n[\mu(\Gamma \setminus Q_{n,\rho_n^\alpha})]. \quad (7)$$

This definition of residual uncertainty is reasonable for conservative estimates because they aim at controlling the error $\mathbb{E}_n[G_n^{(1)}(\rho_n^\alpha)]$. In particular it is possible to show that the ratio between the Type I error and the measure of a conservative estimate is bounded.

Proposition 1. *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 $1 - \alpha$.*

Proof. See appendix B. □

If the posterior GP mean provides a good approximation of the function f , conservative estimates with high α tend to be inside the true set Γ^* . In such situations the Type I error is usually very small while Type II error could be rather large. Note the differences in Type I/II errors reported in table 1 for the analytical example. Type II uncertainty is thus a relevant quantity when evaluating conservative estimates. In the test case studies we also compute the expected type I error to check that it is consistently small.

3.2 SUR criteria

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 first n locations and their 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 $x_{n+1}, \dots, x_{n+q} \in \mathbb{X}$.

Since the locations x_{n+1}, \dots, x_{n+q} and the responses Z_{n+1}, \dots, Z_{n+q} are unknown, the uncertainty H_{n+q} and the conservative level ρ_{n+q}^α are random variables. The criteria introduced below (equations (8) and (11)) are properly defined for $\rho = \rho_{n+q}^\alpha$, however, there are no closed form formulae to compute 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 (6) and (7).

The *conservative* J_n criterion is defined as

$$J_n(\mathbf{x}^{(q)}; \rho_n^\alpha) = \mathbb{E}_{n, \mathbf{x}^{(q)}} [H_{n+q, \rho_n^\alpha}] = \mathbb{E}_{n, \mathbf{x}^{(q)}} [\mu(\Gamma \Delta Q_{n+q, \rho_n^\alpha})] \quad (8)$$

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 quantile ρ_n which is a function of past n observations.

Proposition 2 (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*

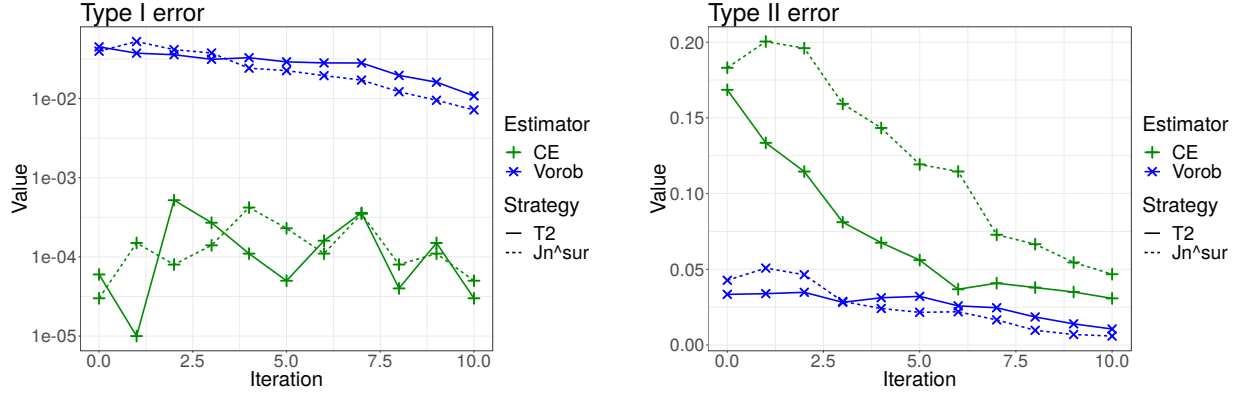
$$\begin{aligned} J_n(\mathbf{x}^{(q)}; \rho_n) &= \mathbb{E}_{n, \mathbf{x}^{(q)}} [\mu(\Gamma \Delta Q_{n+q, \rho_n})] \\ &= \int_{\mathbb{X}} \left(2\Phi_2 \left(\begin{pmatrix} a_{n+q}(u) \\ \Phi^{-1}(\rho_n) - a_{n+q}(u) \end{pmatrix}; \begin{pmatrix} 1 + \gamma_{n+q}(u) & -\gamma_{n+q}(u) \\ -\gamma_{n+q}(u) & \gamma_{n+q}(u) \end{pmatrix} \right) \right. \\ &\quad \left. - p_n(u) + \Phi \left(\frac{a_{n+q}(u) - \Phi^{-1}(\rho_n)}{\sqrt{\gamma_{n+q}(u)}} \right) \right) d\mu(u), \end{aligned} \quad (9)$$

where

$$\begin{aligned} a_{n+q}(u) &= \frac{\mathbf{m}_n(u) - t}{\mathfrak{s}_{n+q}(u)}, & \mathbf{b}_{n+q}(u) &= \frac{K_q^{-1} \mathfrak{K}_n(\mathbf{x}^{(q)}, u)}{\mathfrak{s}_{n+q}(u)}, \\ \gamma_{n+q}(u) &= \mathbf{b}_{n+q}^T(u) K_q \mathbf{b}_{n+q}(u) & p_n(u) &= \Phi \left(\frac{\mathbf{m}_n(u) - t}{\mathfrak{s}_n(u)} \right), \quad u \in \mathbb{X}, \end{aligned} \quad (10)$$

with $\mathfrak{K}_n(\mathbf{x}^{(q)}, u) = (\mathfrak{K}_n(x_{n+1}, u), \dots, \mathfrak{K}_n(x_{n+q}, u))^T$, $K_q = \mathfrak{K}_n(\mathbf{x}^{(q)}, \mathbf{x}^{(q)}) + \text{diag}(\tau^2(\mathbf{x}^{(q)}))$ is assumed invertible, $\mathfrak{K}_n(\mathbf{x}^{(q)}, \mathbf{x}^{(q)}) = [\mathfrak{K}_n(x_{n+i}, x_{n+j})]_{i,j=1,\dots,q}$ 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 (6) does not contribute equally to the expected distance in measure, as observed in proposition 1.



(a) Empirical expected type I error computed on 100 posterior realizations.

(b) Empirical expected type II error computed on 100 posterior realizations.

Figure 5: Comparison of J_n^{SUR} (Bect et al., 2012) and J_n^{T2} on the example in figure 1.

It is thus reasonable to consider the *Type II criterion*, based on the Type II uncertainty (definition 1), defined as follows.

$$\begin{aligned}
 J_n^{T2}(\mathbf{x}^{(q)}; \rho_n^\alpha) &= \mathbb{E}_{n, \mathbf{x}^{(q)}} [H_{n+q}^{T2}(\rho_n^\alpha)] \\
 &= \mathbb{E}_{n, \mathbf{x}^{(q)}} [G_n^{(2)}(Q_{n+q, \rho_n^\alpha})], \quad \text{for } \mathbf{x}^{(q)} \in \mathbb{X}^q.
 \end{aligned} \tag{11}$$

The criterion J_n^{T2} can also be expressed in a semi-analytical form that lends itself better to optimization.

Proposition 3 (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 $J_n^{T2}(\cdot; \rho_n^\alpha)$ can be expanded in closed-form as*

$$\begin{aligned}
 J_n^{T2}(\mathbf{x}^{(q)}; \rho_n^\alpha) &= \mathbb{E}_{n, \mathbf{x}^{(q)}} [G_n^{(2)}(Q_{n+q, \rho_n^\alpha})] \\
 &= \int_{\mathbb{X}} \Phi_2 \left(\begin{pmatrix} a_{n+q}(u) \\ \Phi^{-1}(\rho_n^\alpha) - a_{n+q}(u) \end{pmatrix}; \begin{pmatrix} 1 + \gamma_{n+q}(u) & -\gamma_{n+q}(u) \\ -\gamma_{n+q}(u) & \gamma_{n+q}(u) \end{pmatrix} \right) d\mu(u).
 \end{aligned} \tag{12}$$

The criteria J_n, J_n^{T2} are implemented in this work with a plug-in approach for covariance hyper-parameters, i.e. at each step the hyper-parameters $\theta \in \Theta$ are estimated with maximum likelihood and plugged-in the covariance kernel. The formulae in propositions 2 and 3 could be adapted to a fully Bayesian approach, however their evaluation requires advanced Monte Carlo techniques and it will be a future topic of research.

Figure 5 shows a comparison of estimated type I and type II errors obtained with strategy J_n^{SUR} (Bect et al., 2012, equation (23)) and with strategy $J_n^{T^2}$ in the experimental setting of figure 1. For the Vorob’ev expectation the two strategies produce very similar results, however for conservative estimates, $J_n^{T^2}$ reduces type II error faster than J_n^{SUR} .

3.3 Implementation details

Propositions 2 and 3 provide closed-form expressions for the criteria, however their computation requires numerical approximations. In particular, the evaluation of J_n and $J_n^{T^2}$ require the computation of an integral over \mathbb{X} with respect to μ . The integral can be computed with an importance sampling Monte Carlo method. Here, however, we fix the integration points with space filling designs, such as Sobol’ sequence or uniform sampling. If the dimension of \mathbb{X} is high, the region of interest for sampling could become very small with respect to \mathbb{X} and this would make simple Monte Carlo or importance sampling methods very inefficient. We did not observe this behavior in our experiments, however, in such cases sequential Monte Carlo (SMC) methods could provide better results, see, e.g., Bect et al. (2017) and references therein. 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. We minimize the criteria with the genetic algorithm using derivatives of Mebane and Sekhon (2011).

The strategies are implemented in the **R** programming language (R Core Team, 2018) in the package `KrigInv` (Chevalier et al., 2014c). The function `EGI` in `KrigInv` produces adaptive designs such as the one in figure 1b by automatically optimizing the criterion $J_n^{T^2}$. `KrigInv` interfaces with `DiceKriging` (Roustant et al., 2012) for Gaussian modelling, `rgenoud` for the optimization routine and `anMC` (Azzimonti and Ginsbourger, 2018) for conservative estimates. The current implementation only allows a homogeneous noise standard deviation $\tau(x) \equiv \tau \geq 0$.

4 Results on reliability engineering test case

In this section we review the test case introduced in section 1.1 and apply the sequential strategies from the previous section to reduce the uncertainty on the conservative estimate.

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. Each evaluation of k-eff is done with an expensive computer experiment. We thus aim to compute an estimate for Γ^* from as few evaluations of k-eff as possible and to quantify its uncertainty. The true data result from a MCMC simulation and have a heterogeneous noise variance. Here we consider our k-eff function in figure 2 obtained from 50×50 evaluations of k-eff smoothed with a GP model that accounts for a prescribed value of noise variance provided by the simulator and considered as the true variance.

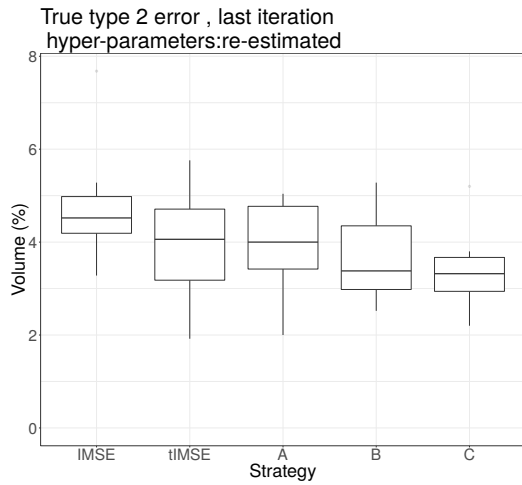
We consider a GP model with covariance function from the Matérn family and homogeneous noise variance estimated from the data. We choose the regularity parameter $\nu = 5/2$ in order to represent the regularity of the underlying phenomenon. The initial DoE is a Latin hypercube sample design with $n_0 = 15$ function evaluations at the points plotted as triangles in figure 2a. We consider the five strategies listed in table 2, where we recall that the strategy IMSE chooses the next evaluation by minimizing the integrated mean squared error of the prediction, see, e.g. Sacks et al. (1989). The five strategies are compared on $m_{\text{doe}} = 10$ 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 2a 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 $n = 20$ iterations of each strategy and at each step we select a batch 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 2c shows the coverage function of Γ obtained after 75 function evaluations at locations selected with Strategy C and $\text{CE}_{\alpha,75}$.

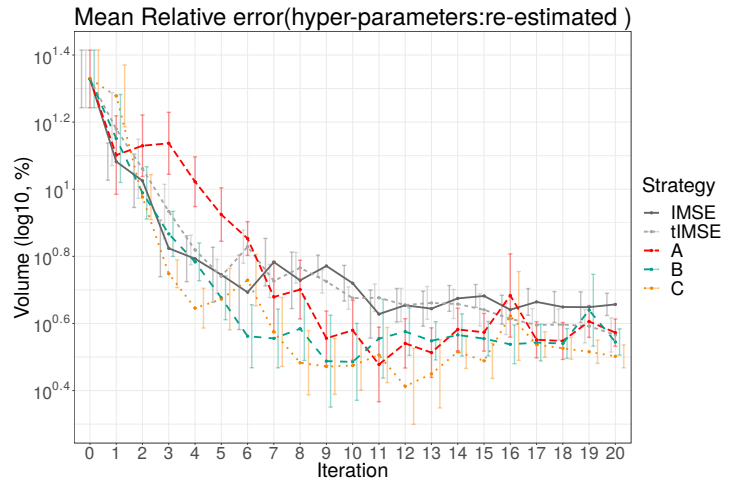
Figure 6a shows the type II error (as percentage of the total measure of \mathbb{X}) at the last iteration, i.e. after 75 evaluations of the function, for each initial DoE and each strategy.

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	$J_n^{\text{T}2}(\cdot; \rho_n^\alpha)$	$\alpha = 0.95$



(a) True type II error, last iteration.



(b) Relative volume error versus iteration number.

Figure 6: Nuclear criticality safety test case, randomized initial DoEs.

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 6b 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 50×50 grid. The volume of Γ^* computed with numerical integration from this grid of evaluations is 88.16% of the total volume of the input space. All strategies show a strong decrease in relative volume error in the first 10 iterations, 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, seems to provide the best uncertainty reductions both in terms of relative volume error and in terms of type II error.

5 Numerical benchmarks

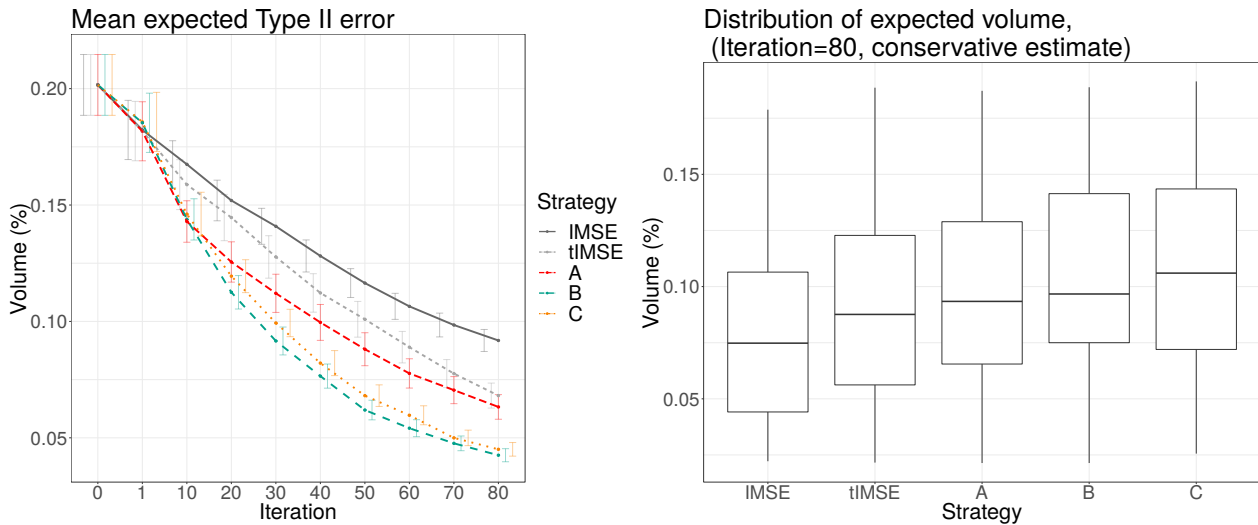
In this section we develop a benchmark study with Gaussian process realizations to study the different behavior of the proposed strategies. We consider two cases with the following shared setup. The input space is the unit hypercube $\mathbb{X} = [0, 1]^d$, for $d = 2, 5$ and $(\xi_x)_{x \in \mathbb{X}} \sim GP(\mathbf{m}, \mathfrak{K})$ with constant prior mean $\mathbf{m} \equiv 0$ 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 $\Gamma = \{x \in \mathbb{X} : \xi_x \geq 1\}$ and μ is the usual volume. We test the strategies listed 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 criterion 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 $\xi_{\mathbf{X}_{n_{\text{init}}}}$.

We evaluate the strategies by looking at the type I and type II errors for Q_{n, ρ_n^α} , defined in section 3.1, and by computing the measure $\mu(Q_{n, \rho_n^\alpha})$. Since the estimate Q_{n, ρ_n^α} has a

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



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

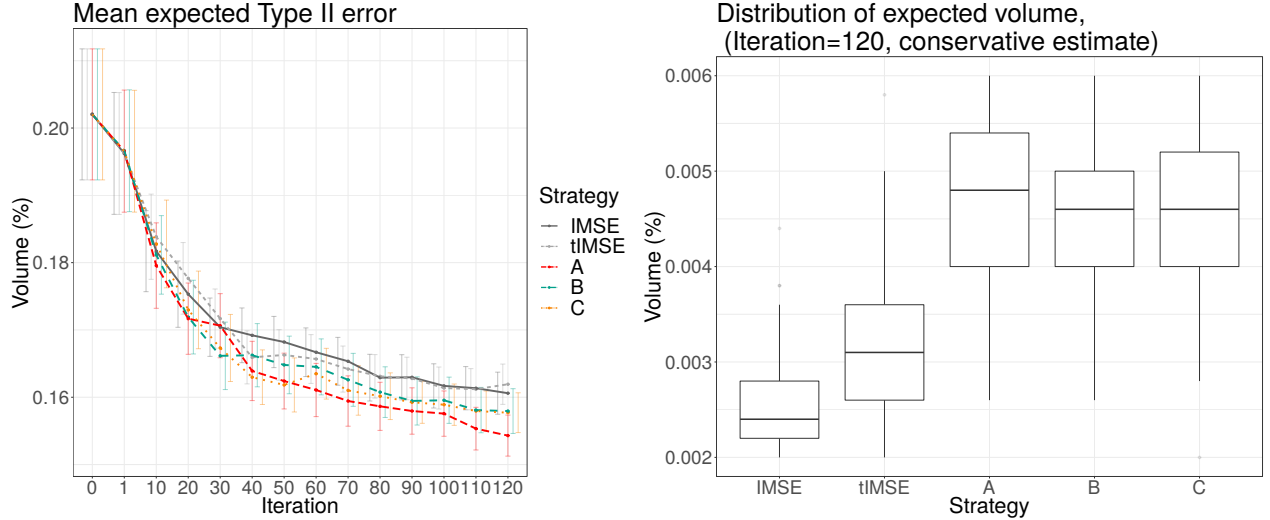
(b) Volume $\mu(Q_{n,\rho_n^\alpha})$ across different initial DoE, after $n = 80$ iterations.

Figure 7: Gaussian process realizations test case in dimension 2.

guaranteed low type I error, a large measure is an indicator of a non trivial conservative estimate. 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 3.1.

5.1 Dimension 2 GP realizations

Figure 7a 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 C outperform the others. Figure 7b shows the values of expected volume $\mathbb{E}_n[\mu(Q_{n,\rho_n^\alpha})]$ obtained after $n = 80$ new evaluations, across different initial DoEs. The strategies A, B, C all provide better uncertainty reduction for conservative estimates than a standard IMSE



(a) Mean type II error for Q_{n,ρ_n^α} across different initial DoE, $n = 120$ iterations. (b) Volume $\mu(Q_{n,\rho_n^\alpha})$ across different initial DoE, after $n = 120$ iterations.

Figure 8: Gaussian process realizations test case in dimension 5.

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 1. 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).

5.2 Dimension 5 GP realizations

Figures 8a and 8b show 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 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. Over all DoEs and replications,

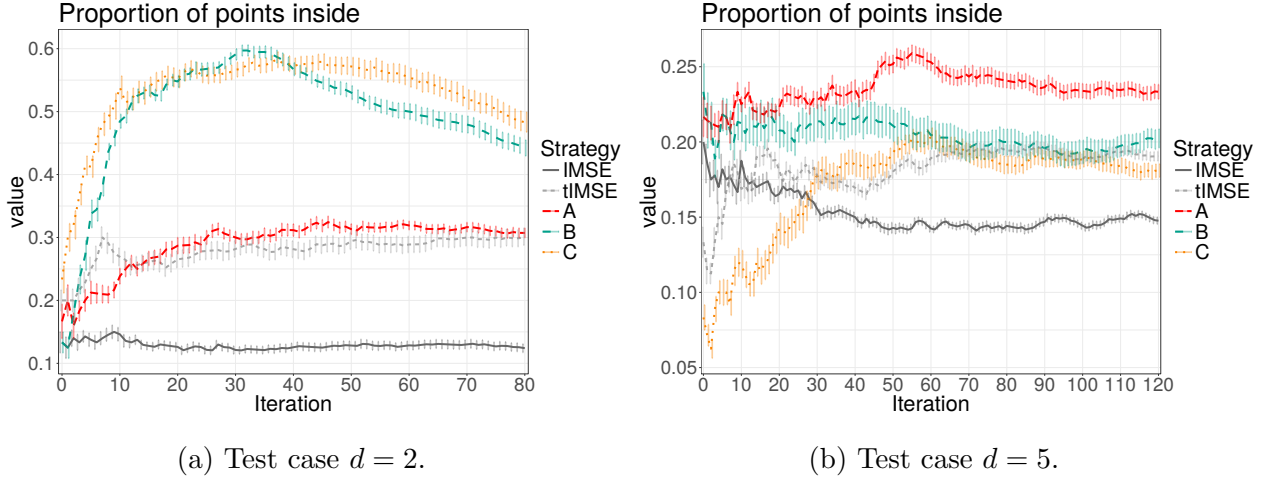


Figure 9: GP realizations. Average proportion of points inside the excursion region.

the median ratio between the expected type I error and the volume of Q_{n,ρ_n^α} 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.

5.3 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 $\frac{\#\{j:z_j \geq t, j=1,\dots,n_i\}}{n_i}$, where n_i is the total number of points at iteration i and z_1, \dots, z_{n_i} are the evaluations. Figure 9 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 inside the excursion reflects the volume of excursion. Strategies A and tIMSE are adapted to the problem of estimating an excursion set, however they are not adapted for conservative estimation, as such they tend to select points around the boundary of Γ and not inside. Strategies B and 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 9a, by the proportion

of points inside the excursion set. In the five dimensional test case, figure 9b, the proportion of points inside the excursion set is similar across all strategies, except for the IMSE strategy which tends to have a smaller proportion.

6 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, 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 than generic strategies. In particular, strategy *C*, i.e. the criterion $J_n^{\text{T}2}(\cdot; \rho_n^\alpha)$, 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 strategies with a-posteriori measures of uncertainty. Expected type I and II errors could also be used to provide stopping criteria for the sequential strategies. Further studies on those quantities could lead to a better understanding of their 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. A fully Bayesian procedure accounting for hyper-parameter uncertainty could mitigate this issue, however, as already suggested in section 3, further studies on the subject are required.

References

- Adler, R. J. and Taylor, J. E. (2007). *Random Fields and Geometry*. Springer, Boston.
- 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*.
- Azaïs, J.-M. and Wschebor, M. (2009). *Level sets and extrema of random processes and fields*. Wiley Online Library.
- Azzimonti, D. (2016). *Contributions to Bayesian set estimation relying on random field priors*. PhD thesis, University of Bern.
- Azzimonti, D., Bect, J., Chevalier, C., and Ginsbourger, D. (2016). Quantifying uncertainties on excursion sets under a Gaussian random field prior. *SIAM/ASA J. Uncertain. Quantif.*, 4(1):850–874.
- Azzimonti, D. and Ginsbourger, D. (2018). Estimating orthant probabilities of high dimensional gaussian vectors with an application to set estimation. *J. Comput. Graph. Statist.*, 27(2):255–267.
- 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. (2018). 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.
- Bect, J., Li, L., and Vazquez, E. (2017). Bayesian subset simulation. *SIAM/ASA J. Uncertain. Quantif.*, 5(1):762–786.
- Berkenkamp, F., Turchetta, M., Schoellig, A., and Krause, A. (2017). Safe model-based reinforcement learning with stability guarantees. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 908–918. Curran Associates, Inc.
- 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.
- Gotovos, A., Casati, N., Hitz, G., and Krause, A. (2013). Active learning for level set estimation. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, IJCAI '13*, pages 1344–1350. AAAI Press.
- 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.
- Picheny, V., Ginsbourger, D., Richet, Y., and Caplin, G. (2013). Quantile-based optimization of noisy computer experiments with tunable precision. *Technometrics*, 55(1):2–13.
- R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Ranjan, P., Bingham, D., and Michailidis, G. (2008). Sequential experiment design for contour estimation from complex computer codes. *Technometrics*, 50(4):527–541.
- 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.
- Spodarev, E. (2014). *Limit Theorems for Excursion Sets of Stationary Random Fields*, pages 221–241. Springer International Publishing, Cham.
- 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.

A Sequential strategies

Proof of proposition 2. Recall that

$$\mathbb{E}_{n,\mathbf{x}^{(q)}} [\mu(\Gamma \Delta Q_{n+q,\rho_n})] = \mathbb{E}_{n,\mathbf{x}^{(q)}} [\underbrace{\mu(Q_{n+q,\rho_n} \setminus \Gamma)}_{=G_{n+q}^{(1)}(\rho_n)}] + \mathbb{E}_{n,\mathbf{x}^{(q)}} [\underbrace{\mu(\Gamma \setminus Q_{n+q,\rho_n})}_{=G_{n+q}^{(2)}(\rho_n)}]. \quad (13)$$

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

$$\mathbb{E}_{n,\mathbf{x}^{(q)}} [G_{n+q}^{(2)}(\rho_n)] = \int_{\mathbb{X}} \mathbb{E}_n [p_{n+q}(u) \mathbb{1}_{\{p_{n+q}(u) < \rho_n\}}] d\mu(u) \quad (14)$$

$$\begin{aligned} \mathbb{E}_{n,\mathbf{x}^{(q)}} [G_{n+q}^{(1)}(\rho_n)] &= \int_{\mathbb{X}} \mathbb{E}_n [\mathbb{1}_{\{p_{n+q}(u) \geq \rho_n\}} (1 - p_{n+q}(u))] d\mu(u) \quad (15) \\ &= \int_{\mathbb{X}} (\mathbb{E}_n [\mathbb{1}_{\{p_{n+q}(u) \geq \rho_n\}}] - \mathbb{E}_n [\mathbb{1}_{\{p_{n+q}(u) \geq \rho_n\}} p_{n+q}(u)]) d\mu(u) \\ &= \int_{\mathbb{X}} (\mathbb{E}_n [\mathbb{1}_{\{p_{n+q}(u) \geq \rho_n\}}] - p_n(u)) d\mu(u) + \mathbb{E}_{n,\mathbf{x}^{(q)}} [G_{n+q}^{(2)}(\rho_n)] \end{aligned}$$

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

$$p_{n+q,\mathbf{x}^{(q)}}(x) = \Phi(a_{n+q}(x) + \mathbf{b}_{n+q}^T Y_q), \quad (16)$$

where a_{n+q} , \mathbf{b}_{n+q} are defined in equation (10) and $Y_q \sim N_q(0, K_q)$ is a q dimensional normal random vector. The first part of equation (14) is

$$\begin{aligned} \mathbb{E}_n [\mathbb{1}_{\{p_{n+q}(u) < \rho_n\}}] &= P_n(p_{n+q}(u) < \rho_n) = P_n(\mathbf{b}_{n+q}^T Y_q \geq \Phi^{-1}(\rho_n) - a_{n+q}(u)) \\ &= \Phi\left(\frac{a_{n+q}(u) - \Phi^{-1}(\rho_n)}{\sqrt{\mathbf{b}_{n+q}^T(u) K_q \mathbf{b}_{n+q}(u)}}\right) \end{aligned} \quad (17)$$

where the second equality follows from equation (16) and the third from $Y_q \sim N(0, K_q)$.

Moreover

$$\begin{aligned} \mathbb{E}_n [\mathbb{1}_{\{p_{n+q}(u) < \rho_n\}} p_{n+q}(u)] &= \int \Phi(a_{n+q}(u) + \mathbf{b}_{n+q}^T(u)y) \mathbb{1}_{\{\mathbf{b}_{n+q}^T(u)y < \Phi^{-1}(\rho_n) - a_{n+q}(u)\}} \Psi(y) \\ &= \int P(N_1 \leq a_{n+q}(u) + \mathbf{b}_{n+q}^T(u)y) \mathbb{1}_{\{\mathbf{b}_{n+q}^T(u)y < \Phi^{-1}(\rho_n) - a_{n+q}(u)\}} \Psi(y) \\ &= \mathbb{E} [P(N_1 \leq a_{n+q}(u) + \mathbf{b}_{n+q}^T y, \mathbf{b}_{n+q}^T(u)y < \Phi^{-1}(\rho_n) - a_{n+q}(u))] \\ &= \Phi_2\left(\left(\begin{array}{c} a_{n+q}(u) \\ \Phi^{-1}(\rho_n) - a_{n+q}(u) \end{array}\right); \left(\begin{array}{cc} 1 + \gamma_{n+q}(u) & -\gamma_{n+q}(u) \\ -\gamma_{n+q}(u) & \gamma_{n+q}(u) \end{array}\right)\right). \end{aligned} \quad (18)$$

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 (13) to (15), (17) and (18) we obtain equation (9). □

Proof of proposition 3. The proof follows from equations (14) and (18). □

B Properties of conservative estimates

B.1 Conservative estimates and confidence regions

Consider an excursion set $\Gamma = \{x \in D : Z_x \geq t\}$, $t \in \mathbb{R}$ and recall that a conservative estimate Q_{ρ^*} for Γ is chosen as the Vorob'ev quantile with $\rho^* \in \arg \max_{\rho \in [0,1]} \{\mu(Q_\rho) : P(Q_\rho \subset \Gamma) \geq \alpha\}$. Since $Q_\rho \subset \Gamma \Leftrightarrow \Gamma^C \subset Q_\rho^C$ and $\mu(Q_\rho^C) = \mu(D) - \mu(Q_\rho)$, then we have that ρ^* is also the minimizer of $\rho \rightarrow \mu(Q_\rho^C)$ under the constraint $P(\Gamma \subset Q_\rho^C) \geq \alpha$. We can look at $Q_{\rho^*}^C$ as a confidence region for Γ^C , in the sense that it is the smallest set that contains Γ^C with a given probability.

In a reliability framework, if Γ is the set of safe configurations, then by selecting a conservative estimate for the safe set, we are actually selecting a confidence region for the dangerous configurations.

B.2 Conservative estimates with Vorob'ev quantiles

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

$$Q_{\rho_1} \subset Q_{\rho_2}, \quad Q_{\rho_1}, Q_{\rho_2} \in \mathfrak{C}. \tag{19}$$

We now detail how to compute $\text{CE}_{\alpha,n}$ based on \mathfrak{C} , for a fixed $\alpha \in [0, 1]$ from n observations. For each $\rho \in [0, 1]$, we define the function $\psi_\Gamma : [0, 1] \rightarrow [0, 1]$ 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 (19). Moreover, $\mu(Q_{\rho_1}) \leq \mu(Q_{\rho_2})$ for $\rho_1 \geq \rho_2$. The computation of $\text{CE}_{\alpha,n}$ amounts to finding the smallest $\rho = \rho_n^\alpha$ such that $\psi_\Gamma(\rho_n^\alpha) \geq \alpha$, which is achievable, for example, with a simple dichotomic search. 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 4. *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 4 is an extension of Theorem 2.3, Molchanov (2005) to a generic Vorob'ev quantile. As a consequence, a conservative estimate $\text{CE}_{\alpha,n} = Q_{n,\rho_n^\alpha}$ 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 $\text{CE}_{\alpha,n}$ with this procedure is not the set S with the largest measure satisfying the property $P(S \subset \Gamma) \geq \alpha$.

B.3 Proofs

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

Proof of proposition 4. We want to show that the set Q_ρ satisfies

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

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

$$\begin{aligned} \mu(M \Delta \Gamma(\omega)) - \mu(Q_\rho \Delta \Gamma(\omega)) &= 2 \left(\mu(\Gamma(\omega) \cap (Q_\rho \setminus M)) - \mu(\Gamma(\omega) \cap (M \setminus Q_\rho)) \right) \\ &\quad + \mu(Q_\rho^C) - \mu(M^C). \end{aligned}$$

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

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

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 (20). □

Proof of proposition 1. 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

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

□