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. 2019. hal-01379642v5

HAL Id: hal-01379642

https://hal.science/hal-01379642v5

Preprint submitted on 21 Oct 2019 (v5), 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 by 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 a Gaussian Process (GP) model 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 different scenarios of noise and batch size. We then apply it 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. The 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), 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, such as 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 $f : \mathbb{X} \to \mathbb{R}$, where $\mathbb{X}$ is a compact subset of $\mathbb{R}^d$. Section 5 shows an example of such systems. Given a few evaluations of f and a fixed closed set $T \subset \mathbb{R}$, we are interested in estimating

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

In the motivating test case in section 5, for example, $\Gamma(f)$ represents the *safe region*, i.e. all values of the physical parameters that lead the system of interest to a subcritical response, taking $T = (-\infty, t]$ with $t \in \mathbb{R}$.

There is much heterogeneity in the literature on how to name $\Gamma(f)$. Here we follow Adler and Taylor (2007) and we call $\Gamma(f)$ an *excursion set*. If $T = [t, +\infty)$, with $t \in \mathbb{R}$, $\Gamma(f)$ 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 primarily motivated by the case $T = [t, +\infty)$, however it may also applied to $T = (-\infty, t]$.

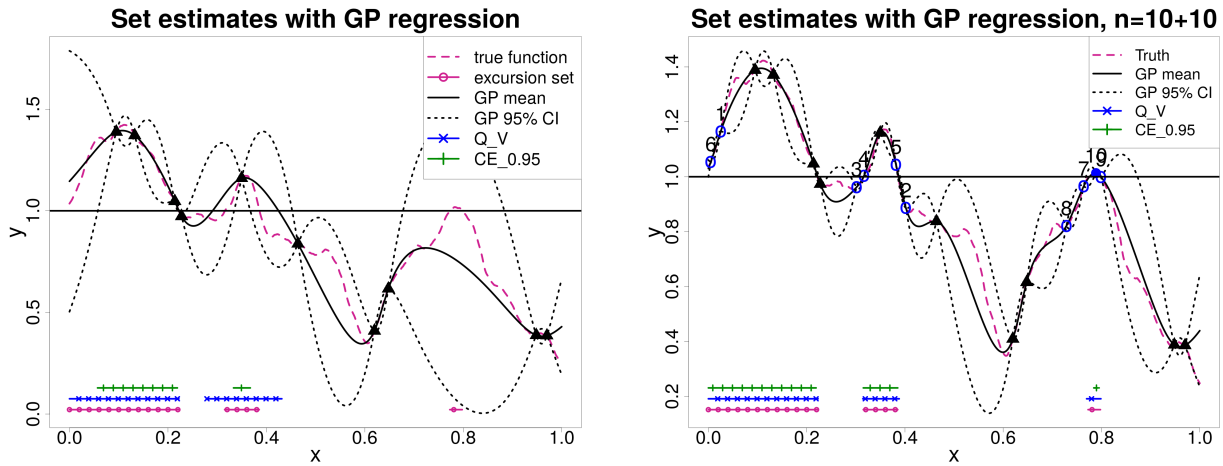
Throughout the article, we model f as a realization of a Gaussian process (GP) and, following Sacks et al. (1989) and Santner et al. (2018), we emulate f with the posterior GP distribution given the available function evaluations. The posterior GP distribution can be used as building block for different estimates of $\Gamma(f)$, see, e.g., Azzimonti (2016).

Consider now a generic estimate $\widetilde{\Gamma}$ for $\Gamma(f)$ and denote with $\mathrm{vol}(A)$ the volume of $A \subset \mathbb{X}$. Simple quality indicators for a set estimate are the volume of false positives $\mathrm{vol}(\widetilde{\Gamma} \setminus \Gamma(f))$, i.e. the volume of points estimated in the set while actually outside $\Gamma(f)$, and the volume of false negatives $\mathrm{vol}(\Gamma(f) \setminus \widetilde{\Gamma})$, i.e. the volume of points estimated not in the excursion set while actually inside. For example, in Chevalier (2013) and Chevalier et al. (2014a), $\Gamma(f)$ 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. 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) chosen with a maximin Latin hypercube sample (LHS) design and we estimate $\Gamma(f)$ (purple dotted horizontal line), where $T = [1, +\infty)$, with the Vorob'ev expectation (Q_V, middle horizontal blue line); a comparison of Q_V with the true excursion set shows that here Q_V has volumes of false positive (0.084) and negatives (0.025) of the same order of magnitude.

The Vorob'ev expectation, explained in more details in section 2.1, 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 be interested in set estimates which would *very likely* be included in an excursion set of the form $\Gamma(f) = \{x \in \mathbb{X} : f(x) \geq t\}$. Such a property naturally gives more importance to the minimization of (the volume of) false positives than of false negatives. French and Sain (2013) and Bolin and Lindgren (2015) introduced the concept of conservative estimates which select sets that are deliberately smaller – in volume – than $\Gamma(f)$ and are included in the excursion set with a *large probability* α. 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 at level $\alpha \approx 1$ 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 $\alpha = 0.95$ ($\mathrm{CE}_{0.95}$, green top horizontal line). In this example, $\mathrm{CE}_{0.95}$ 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,



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

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

Figure 1: Example of excursion set estimation. Comparison of Vorob'ev expectation (Q_V) and conservative estimate ($\mathrm{CE}_{0.95}$). The numbers near the new points indicate the order in which they were added to the DoE.

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 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 the problem of sequentially choosing evaluation points in order to reduce the uncertainty on conservative estimates. In order to illustrate this concept, consider the example introduced in figure 1a and notice how conservative set estimates do not intercept the excursion near $x = 0.8$. In this case we can increase the size of our design of experiments by adaptively choosing new evaluations of f that help us better localize the excursion. Figure 1b shows an example of such adaptive DoEs where, starting from the initial DoE in figure 1a, 10 additional points are selected with strategy C introduced in section 3.

Previous adaptive DoE strategies for excursion set estimation mainly focused on recovering the boundaries of the set. In particular, Picheny et al. (2010) introduced the targeted IMSE (integrated mean squared error), 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). 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 in the case of excursion sets above $t \in \mathbb{R}$, however our **R** implementation, available on CRAN, allows also for excursions below t.

1.1 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 uncertainty on such estimates. In section 3.2, we detail the proposed sequential strategies and, in appendix A, we derive fast-to-compute formulae for the associated criteria and illustrate their implementation. Section 4 presents a benchmark study where we analyze a trade-off between noisy evaluations and batch size in three scenarios. Section 5 shows the results obtained on a reliability engineering test case. In appendix B we provide more properties for conservative estimates that further justify the choices made in section 2.1. In supplementary material we also apply the proposed strategies on a coastal flood problem. All proofs are in appendices A and B.

2 Background

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

$$z_i = f(x_i) + \tau\epsilon_i \qquad x_i \in \mathbb{X}, \quad i = 1, \ldots, n \tag{2}$$

with ϵ_i independent realizations of standard Gaussian measurement noise and τ^2 a known homogeneous 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) $\xi \sim GP(\mathfrak{m}, \mathfrak{K})$, with mean function $\mathfrak{m}(x) := \mathbb{E}[\xi_x]$ and covariance function $\mathfrak{K}(x, x') := \mathrm{Cov}(\xi_x, \xi_{x'}),\ x, x' \in \mathbb{X}$. The mean function could potentially have a structure such as $\mathfrak{m}(x) = \sum_{i=1}^{p} \beta_j f_j(x)$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p) \in \mathbb{R}^p$ are parameters to be estimated and f_j are known basis functions. With this notation, z_i is a realization of $Z_i = \xi_{x_i} + \tau\varepsilon$ where $\varepsilon \sim N(0, 1)$. For $n > 0$, we denote by $\mathbf{z}_n = (z_1, \ldots, z_n) \in \mathbb{R}^n$ 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. (2018), Chapter 4, for closed-form formulae.

2.1 Vorob'ev expectation and conservative estimates

The prior distribution on ξ induces a (random) set $\Gamma(\xi) = \{x \in \mathbb{X} : \xi_x \in T\}$. We will omit the dependency on ξ when obvious and refer to $\Gamma(\xi)$ as Γ when appropriate. By using the posterior distribution of ξ, we can provide estimates for $\Gamma(f)$. See, e.g. Chevalier et al. (2014a), Bolin and Lindgren (2015) and 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

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

In our case we consider the posterior coverage function p_n, defined with the posterior probability $P_n(\cdot) = P(\cdot \mid \mathbf{Z}_n = \mathbf{z}_n)$, where $\mathbf{Z}_n = (Z_1, \ldots, Z_n)$. If $T = (-\infty, t]$, then

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	$9.5 \times 10^{-5} \pm 4.7 \times 10^{-4}$	0.187 ± 0.063	0.95

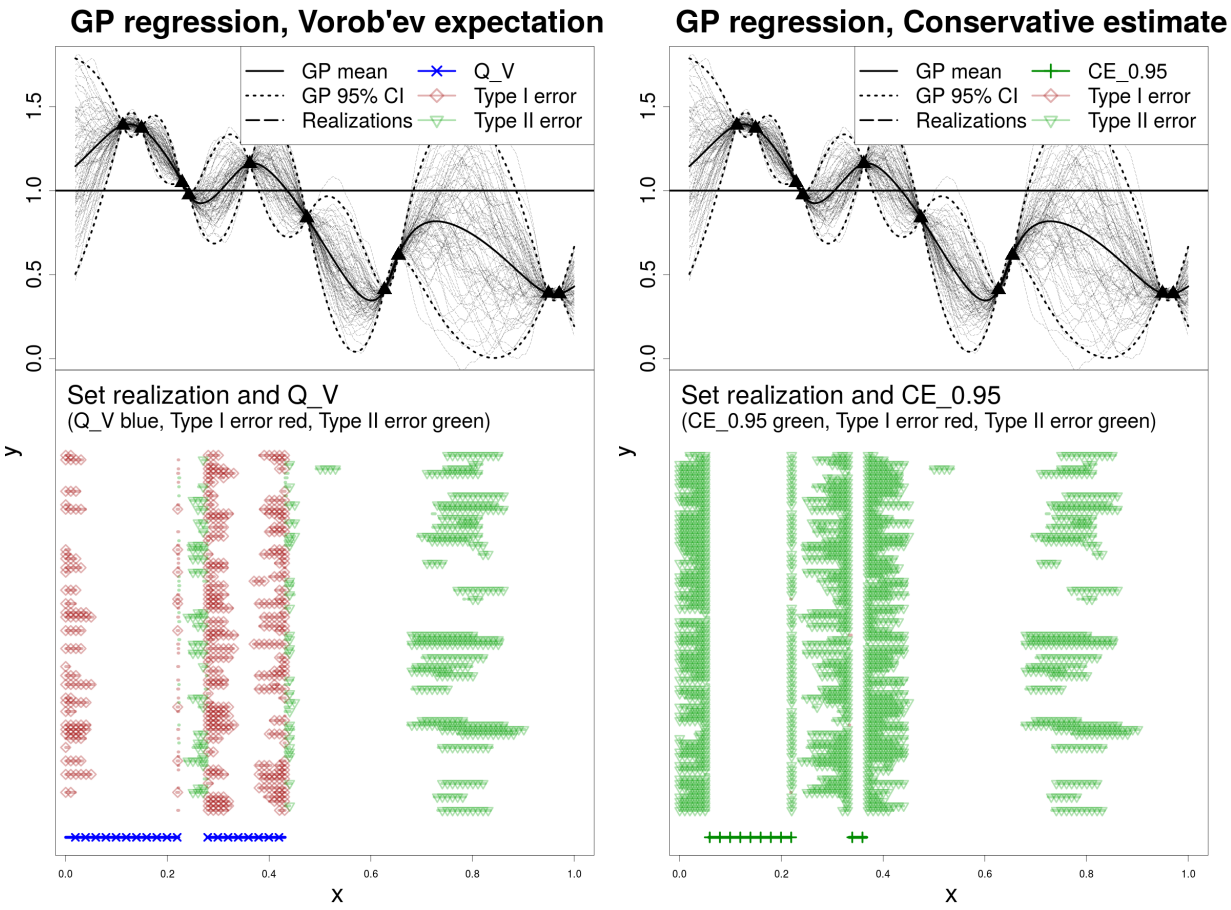
$p_n(x) = \Phi\left(\frac{t-\mathfrak{m}_n(x)}{\mathfrak{s}_n(x)}\right)$, where $\Phi(\cdot)$ is the CDF of a standard Normal random variable and $\mathfrak{s}_n(x) = \sqrt{\mathfrak{K}_n(x,x)}$. The coverage function defines the family of *Vorob'ev quantiles*

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

with $\rho \in [0,1]$. These sets are closed for each $\rho \in [0,1]$ (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. *Vorob'ev expectation* (Vorob'ev, 1984; Molchanov, 2005; Chevalier et al., 2013) relies on the notion of measure. In the example in figure 1b and in the applications presented here we use the standard volume, however here we introduce the concept in a slightly more general form by using a finite measure μ, for example, μ could be a probability distribution on $\mathbb{X}$. The Vorob'ev expectation is defined as the quantile Q_{n,ρ_V} such that $\mu(Q_{n,\rho}) \leq \mathbb{E}[\mu(\Gamma)] \leq \mu(Q_{n,\rho_V})$ for all $\rho > \rho_V$. The set Q_{n,ρ_V} is also the minimizer of $\mathbb{E}[\mu(\Gamma \Delta M)]$[1] among all measurable sets such that $\mu(M) = \mathbb{E}[\mu(\Gamma)]$, see, e.g., Molchanov (2005, Theorem 2.3, Chapter 2). Vorob'ev expectation minimizes a uniformly weighted combination of the expected measure of false positives ($\mathbb{E}[\mu(M \setminus \Gamma)]$, also called *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) used this distance to adaptively reduce the uncertainty on Vorob'ev expectations. In section 3.1 we adapt it for conservative estimates.

[1] For any set A, B, $A \Delta B := A \setminus B \cap B \setminus A$



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

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

Figure 2: 1-dimensional example, $n = 10$ evaluations. Type I (red, diamonds) and Type II (green, triangles) errors for Vorob'ev expectation and conservative estimate.

Conservative estimates (Bolin and Lindgren, 2015; French and Sain, 2013) embed 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 $\Gamma(f)$ 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 set $\mathrm{CE}_{\alpha,n}$ 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) and Azzimonti and Ginsbourger (2018), we choose $\mathfrak{C}$ as the family of Vorob'ev quantiles $\{Q_{n,\rho} : \rho \in [0,1]\}$ as introduced in equation (3). 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, $\mathbb{E}_n[\mu(\mathrm{CE}_{0.95} \setminus \Gamma)]$. Figure 2 shows 80 posterior realizations of the GP (light dashed black lines) and for each realization we computed the false positive (type I error, horizontal red lines) and false negatives (type II error, horizontal green lines). Notice how they are symmetrically minimized by the Vorob'ev expectation (figure 2a) while the conservative estimate with $\alpha = 0.95$ (figure 2b) 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 $\mathrm{CE}_{\alpha,n}$ gives control on the probability of false positives. Table 1 also reports the values for the Vorob'ev quantile $Q_{0.95}$, 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 joint probability of inclusion.

The computation of $\mathrm{CE}_{\alpha,n}$ in equation (4) requires finding 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 $P_n(Q_\rho \subset \Gamma) \approx P_n(\xi_{q_1} \leq t, \ldots, \xi_{q_\ell} \leq t)$, where $\{q_1, \ldots, q_\ell\} \subset Q_\rho$ is a set of

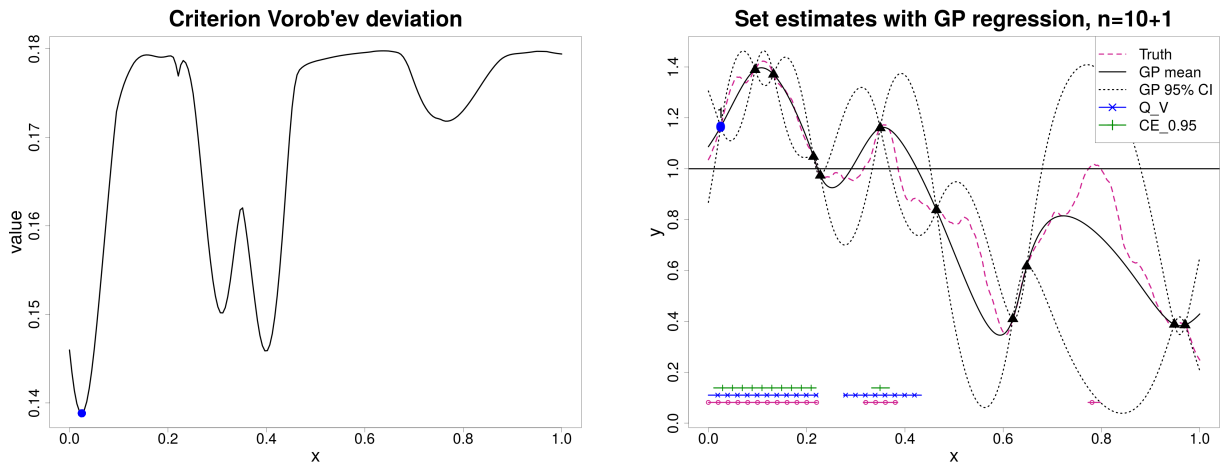
ℓ points in Q_ρ with ℓ large. The probability above is then computed efficiently with the integration technique proposed by Azzimonti and Ginsbourger (2018). The number ℓ is generally chosen as large as the computational budget allows. This technique can also be used for excursion sets above t. An alternative 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. The optimal Vorob'ev level chosen for conservative estimates at level α is denoted by ρ^α in what follows : $\mathrm{CE}_{\alpha,n} := Q_{n,\rho^\alpha}$.

2.2 SUR strategies

Sequential design of experiments adaptively chooses the next evaluation points according to a strategy with the aim of improving the estimation of a quantity (or set, here) of interest. As shown in the introduction we can improve the set estimates in figure 1a by carefully adding new function evaluations, see figure 1b. There are many ways to build a sequential DoE, see, e.g., Santner et al. (2018), chapter 6. Here 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., 2019) and select a sequence of points in order to reduce the uncertainty of a quantity of interest.

In the remainder of the paper we consider that the first n points $x_1, \ldots, x_n$ 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 $x_{n+1}, \ldots, x_{n+q}$. The advantage of batches with $q > 1$ lies in the fact that parallel function evaluations, when available, can save the user wall-clock time. In a sequential setting the response values at these points are unknown before evaluations, 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 $\mathbf{x}^{(q)} = (x_{n+1}, \ldots, x_{n+q}) \in \mathbb{X}^q$.

For a specific problem, we consider a quantity, denoted by H_n, which measures the residual uncertainty at step n. We define this quantity for the conservative estimation problem in section 3.1. If the first n locations and evaluations are known, then H_n is a (deterministic) real number quantifying the residual uncertainty on the estimate. As an example, consider the setup in figure 1 and the uncertainty $H_n := \mathbb{E}_n[\mu(\Gamma \Delta \, \mathrm{CE}_{0.95})]$; we



(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 3: Adaptive DoE with SUR strategy on the example introduced in figure 1.

can compute H_n, $n = 10$, with numerical integration and obtain $H_{10} = 0.23$. 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

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

a minimizer of the future uncertainty in expectation. For a more complete and theoretical perspective on SUR strategies see, e.g., Bect et al. (2019) 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 a *batch sequential one-step lookahead sampling criterion* and is denoted by $J_n : \mathbf{x}^{(q)} \in \mathbb{X}^q \mapsto \mathbb{E}_{n,\mathbf{x}^{(q)}}[H_{n+q}] \in \mathbb{R}$.

We can build a SUR strategy with the uncertainty $H_n = \mathbb{E}_n[\mu(\Gamma \Delta \, \mathrm{CE}_{0.95})]$ mentioned above. The criterion associated with this uncertainty has the remarkable property that it can be computed with fast-to-evaluate formulae, thus making its optimization more convenient. Figure 3a shows the function J_n for $n = 10$ and $q = 1$; the next evaluation x_{11} is chosen as the minimizer of this function restricted to a finite discretization of the domain. Figure 3b shows the updated GP model and $\mathrm{CE}_{0.95}$ which, compared to figure 1a,

is now larger while still included inside 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 unknown 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 use plug-in maximum likelihood estimates for the hyper-parameters. Model checking procedures, such as cross-validation, can be used to evaluate the robustness of hyper-parameters' estimates. If only few observations are available, a fully Bayesian approach might better capture the overall uncertainty. However such an approach is not straightforward for many SUR criteria (see, e.g., Stroh, 2018) and it involves 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 $\Gamma(f)$, therefore we require uncertainty functions that take into account the whole set. Chevalier et al. (2013) and 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 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 (Chevalier, 2013) that the Vorob'ev deviation of a quantile $Q_{n,\rho}$ is

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

In the following sections, ρ will be chosen either as the Vorob'ev median, $\rho = 0.5$, or as the threshold for a conservative estimate at level α after n evaluations, $\rho = \rho_n^\alpha$.

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) and recall that their expectations, i.e. $\mathbb{E}_n[G_n^{(1)}(\rho_n^\alpha)]$ and $\mathbb{E}_n[G_n^{(2)}(\rho_n^\alpha)]$, are called *Type I* and *Type II* errors respectively. 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*

$$H_{n,\rho_n^\alpha}^{\mathrm{T2}} := \mathbb{E}_n[G_n^{(2)}(\rho_n^\alpha)] = \mathbb{E}_n[\mu(\Gamma \setminus Q_{n,\rho_n^\alpha})]. \tag{7}$$

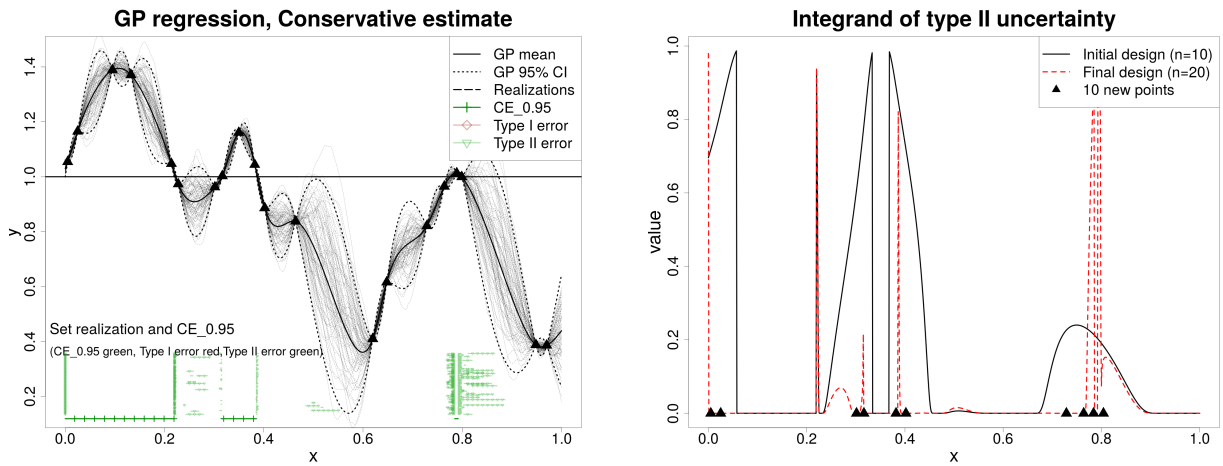
This definition of residual uncertainty is reasonable for conservative estimates because they aim at controlling the error $\mathbb{E}_n[G_n^{(1)}(\rho_n^\alpha)]$. In particular it is possible to show that the ratio between the Type I error and the measure of a conservative estimate is bounded by a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

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 $\Gamma(f)$. 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.

Figure 2b provides a visualization of type II uncertainty: the green horizontal lines are realizations of $\Gamma \setminus Q_{n,\rho_n^\alpha}$ obtained from posterior GP draws. Type II uncertainty is the expected value of the measure of such sets. In the example shown, this amounts to 0.187. Consider now the updated GP estimate in figure 1b where 10 new points were added to the initial DoE by following an adaptive strategy that will be described later. Figure 4a



(a) Type I/II errors final model.

(b) Integrand of type II uncertainty.

Figure 4: 1d example, DoE with initial points of figure 1b plus 10 adaptive new points.

shows how the type II uncertainty is reduced in the updated model: the green horizontal lines are much shorter, resulting in a smaller expected measure of 0.035 ± 0.022.

The expectation and integration operators can be exchanged in equation (7), therefore type II uncertainty can be further written as

$$H^{\mathrm{T2}}_{n,\rho^{\alpha}_{n}} = \mathbb{E}_n[\mu(\Gamma \setminus Q_{n,\rho^{\alpha}_{n}})] = \int_{\mathbb{X}} p_n(x) \mathbb{1}_{(Q_{n,\rho^{\alpha}_{n}})^C}(x) d\mu(x) \tag{8}$$

where $\mathbb{1}_{A^C}$ denotes the indicator function of the complement of set A. Figure 4b plots the integrand in equation (8) for the example in figure 1b with $n = 10$, after the initial LHS design, (black solid line), and for $n = 20$ (red dashed line), after 10 points were added with the same adaptive strategy as in figure 4a. The large bumps shown for $n = 10$ are reduced to small spikes after the new points are added in appropriate locations.

3.2 SUR criteria

Suppose that the first n locations and their respective function evaluations are known. Here we introduce one-step lookahead SUR criteria for conservative estimates based on the measures of residual uncertainty previously introduced. In a sequential algorithm we minimize such criteria to select the next batch of $q > 0$ locations $x_{n+1}, \ldots, x_{n+q} \in \mathbb{X}$.

Since the locations $x_{n+1}, \ldots, x_{n+q}$ and the responses $Z_{n+1}, \ldots, Z_{n+q}$ are unknown, the

uncertainty H_{n+q} and the conservative level ρ^{α}_{n+q} are random variables. The criteria introduced below (equations (9) and (10)) are properly defined for $\rho = \rho^{\alpha}_{n+q}$, the conservative level after $n+q$ evaluations, however, in that case the expectations involved can only be computed with an expensive Monte Carlo procedure. The criteria's implementations use the last known level $\rho = \rho^{\alpha}_{n}$ which allows to expand the criteria in fast-to-evaluate formulae. We consider two sampling criterion based on the uncertainty functions in equations (6) and (7).

The *conservative* J_n criterion is defined as

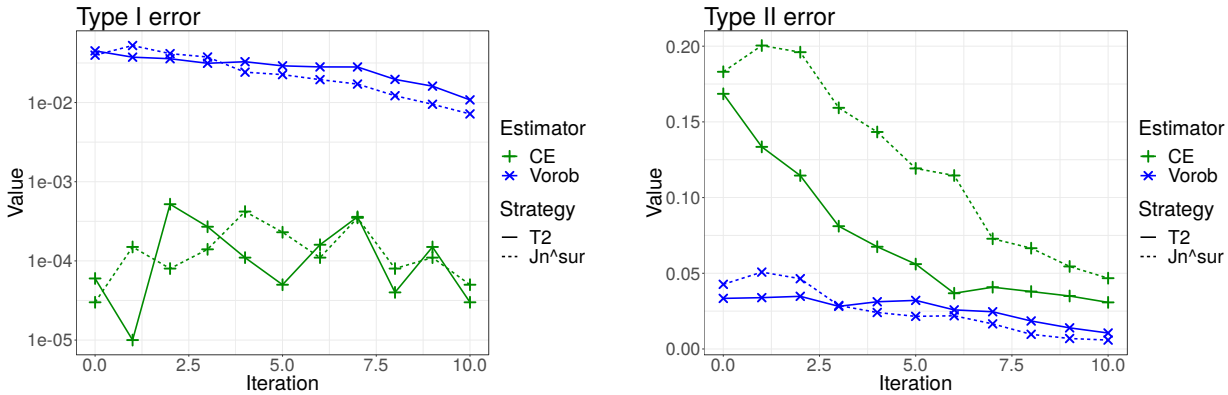
$$J_n(\mathbf{x}^{(q)}; \rho^{\alpha}_n) = \mathbb{E}_{n,\mathbf{x}^{(q)}}\left[H_{n+q,\rho^{\alpha}_n}\right] = \mathbb{E}_{n,\mathbf{x}^{(q)}}\left[\mu(\Gamma\Delta Q_{n+q,\rho^{\alpha}_n})\right] \tag{9}$$

for $\mathbf{x}^{(q)} = (x_{n+1}, \ldots, x_{n+q}) \in \mathbb{X}^q$, where Q_{n+q,ρ^{α}_n} is the Vorob'ev quantile obtained with $n+q$ evaluations of the function at level ρ^{α}_n, 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). In Chapter 4.2, Chevalier (2013), derives the formula for this criterion for the Vorob'ev expectation, i.e. the quantile at level $\rho = \rho_{n,V}$.

Note that each evaluation of $J_n(\mathbf{x}^{(q)})$ requires calculating the expectation $\mathbb{E}_{n,\mathbf{x}^{(q)}}[\cdot]$. This could, in principle, be achieved with a Monte Carlo procedure that draws samples from the posterior distribution of Z, generate posterior samples for Γ and uses such samples to empirically evaluate the expectation. This procedure however could potentially be very costly and, since many evaluations of J_n are required in order to find its minimum, we provide in appendix A, proposition 2 a fast-to-evaluate formula to compute this criterion for any ρ.

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. It is thus reasonable to consider the following criterion.

Definition 2 (Type II criterion)**.** *Consider Q_{n+q,ρ^{α}_n}, the Vorob'ev quantile from $n+q$ evaluations with ρ^{α}_n, the conservative level obtained with n evaluations. The* Type II criterion



(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.

is defined as

$$\begin{aligned} J_n^{\mathrm{T2}}(\mathbf{x}^{(q)};\rho_n^\alpha) &= \mathbb{E}_{n,\mathbf{x}^{(q)}}\left[H_{n+q}^{\mathrm{T2}}(\rho_n^\alpha)\right] \\ &= \mathbb{E}_{n,\mathbf{x}^{(q)}}\left[G_n^{(2)}(Q_{n+q,\rho_n^\alpha})\right], \qquad \textit{for } \mathbf{x}^{(q)} \in \mathbb{X}^q. \end{aligned} \tag{10}$$

Proposition 3, appendix A, provides a fast-to-evaluate formula for (10).

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. A fully Bayesian approach would lead to a higher degree of conservativeness for the final estimate, as the hyper-parameter uncertainty would be accounted for. The formulae introduced in propositions 2 and 3 could be adapted to a fully Bayesian approach, however their evaluation requires advanced Monte Carlo techniques (Stroh, 2018) 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^{T2} in the experimental setting of figure 1. Note how for the Vorob'ev expectation (blue $\times$ lines) the two strategies (dashed or solid lines) produce very similar results, however for conservative estimates (green + lines), J_n^{T2} (solid lines) reduces type II error faster than J_n^{SUR} (dashed lines).

Table 2: MC function evaluation scenarios, total cost $O(n_{MC}kq)$ fixed.

q	τ^2	n_{MC}	k	n
1	$6.25 \cdot 10^{-5}$	$1.6 \cdot 10^4$	50	50
1+7	$5 \cdot 10^{-4}$	$2 \cdot 10^3$	50	400
8	$5 \cdot 10^{-4}$	$2 \cdot 10^3$	50	400
16	10^{-3}	10^3	50	800

3.3 Implementation details

Propositions 2 and 3, in appendix A, provide fast-to-evaluate expressions for the criteria, however their computation requires numerical approximations. See appendix A for more details. New points are obtained by minimizing numerically the selected criterion, 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, 2018) in the package `KrigInv` (Chevalier et al., 2014c). The function `EGIparallel` in `KrigInv` produces adaptive designs such as the one in figure 1b by automatically optimizing the criterion $J_n^{\mathrm{T}2}$. `KrigInv` interfaces with `DiceKriging` (Roustant et al., 2012) for GP modeling, `rgenoud` (Mebane and Sekhon, 2011) for the optimization routine and `anMC` (Azzimonti and Ginsbourger, 2018) for conservative estimates. See **??**, in supplementary materials.

4 Numerical benchmark for batch-sequential criteria

4.1 Noisy function evaluation scenarios

In this section we consider a synthetic numerical study that shows a practical use for batch-sequential criteria. The implementation of uncertainty quantification for expensive-to-evaluate computer experiments is often run on cloud computing platforms. When using such platforms, practitioners often have a fixed total computational budget which is deter-

mined, for example, by the money/time available to run experiments. Resources can be deployed in parallel by creating new computational nodes or sequentially by employing one node for longer time. Nodes are often virtual on such platforms, so there is no restriction on the number of parallel units available.

Here we consider how to allocate resources in order to provide a conservative estimate and reduce the uncertainties for the set $\Gamma(f)$ in (1). In our setting the evaluations of the function f are approximated with Monte Carlo sampling, i.e. for $i = 1, \ldots, n$, we obtain a value $z_i = \frac{1}{n_{MC}} \sum_{j=1}^{n_{MC}} (f(x_i) + \epsilon_{i,j})$, where $\epsilon_{i,j}$ are realizations of i.i.d. Gaussian random variables with zero mean and variance ν^2. The number of Monte Carlo samples, n_{MC}, is fixed before the experiment starts and kept constant throughout. The observation model above can be written as $z_i = f(x_i) + \tau\epsilon_i$ with overall measurement noise $\tau^2 = \nu^2/n_{MC}$. The cost of one observation z_i is proportional to n_{MC} and for larger costs we achieve smaller variance τ^2. Note that noise variance is assumed here homogeneous, i.e. ν^2 does not depend on the location x_i. See Picheny et al. (2013) for an example of online allocation applied to the problem of minimizing a noisy function with tunable precision. We consider an adaptive strategy with k iterations of a batch sequential strategy that selects $q \geq 1$ new points for each iteration, i.e. $n = kq$ function evaluations overall. The total cost of the procedure is therefore $c_{tot} = O(kqn_{MC}) = O(nn_{MC})$, where we assume that the costs of training the GP and of optimizing the criterion are negligible. Since c_{tot} is fixed, then the choices of n_{MC}, k and q are linked. A larger n_{MC} leads to more precise observations, however to a smaller overall number of evaluation n.

We study four possible allocations of resources described in table 2 which range from a "purely sequential" strategy, where at each iteration all resources are used to reduce evaluation noise at a single input location, to a batch-sequential strategy where at each iteration q=16 different locations are explored with a high noise level. Note that the second strategy is a hybrid strategy where we add 1 new evaluation with the criterion selected and 7 others with a randomized LHS space-filling criterion.

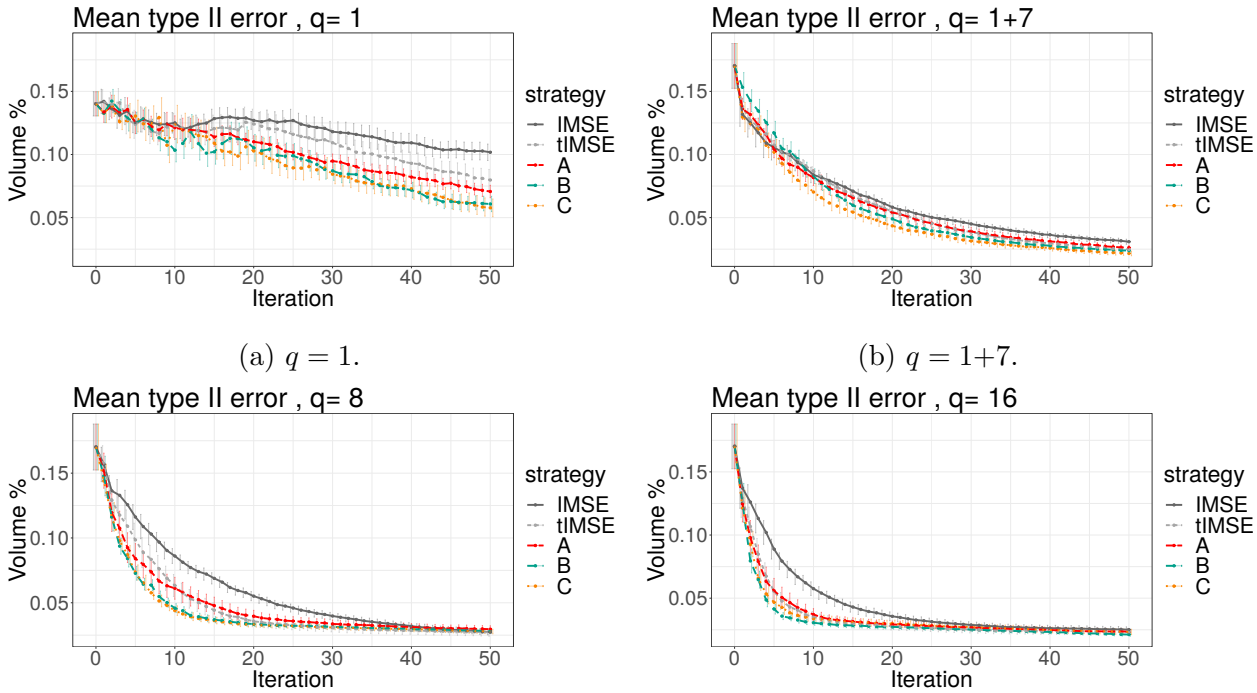
We analyze the trade-off between batch size and noise level on a synthetic test case where we assume that the function f is a realization of $(\xi_x)_{x\in\mathbb{X}} \sim GP(\mathfrak{m}, \mathfrak{K})$ with constant mean function $\mathfrak{m}$ and Matérn covariance kernel $\mathfrak{K}$ with smoothness parameter $\nu = 3/2$,

Table 3: Strategies implemented in the test cases.

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

variance $\sigma^2 = 1$ and lengthscales $\theta_i = 0.2$, $i = 1, 2$. The noise variance is described by the column τ^2 in table 2 and depends on the specific scenario. The set to estimate is $\Gamma(f) = \{x \in [0,1]^2 : f(x) \geq 1\}$, an excursion above $t = 1$. We use the volume on $[0,1]^2$ as the measure μ. For each scenario we consider an initial DoE of size $n_{\text{init}} = 3$ and a GP model with zero prior mean and Matern 3/2 covariance kernel with hyperparameters σ^2 and θ_i known and set to the values specified above. We select the next function evaluation with the strategies listed in table 3, 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) and tIMSE is the targeted IMSE strategy described in Picheny et al. (2010). We run each strategy for the number of iteration specified in table 2. We consider $m_{\text{doe}} = 10$ different initial DoE and, for each design, we replicate the procedure 10 times with different values for $\xi_{\mathbf{X}_{\text{init}}}$.

Figure 6 shows a comparison of expected type II errors for the strategies listed in table 3 averaged over the 10 replications and the 10 initial DoEs. Scenario $q = 1$ shows a faster decrease in type II error for strategies B and C, however there are not enough function evaluations to reach convergence. For the other scenarios, the change in the median value of type II error is smaller than 1% between successive iterations before the final iteration. Note that in figure 6 (a), (c), (d) the differences between strategies are clear: strategies C and B are the fastest in reducing the error, followed by A and tIMSE; IMSE instead achieves the slowest error reduction. This reflects the fact that adaptive strategies tailored to the problem require fewer iterations to reduce type II error. In figure 6 (b), the ranking



(a) $q = 1$.

(b) $q = 1{+}7$.

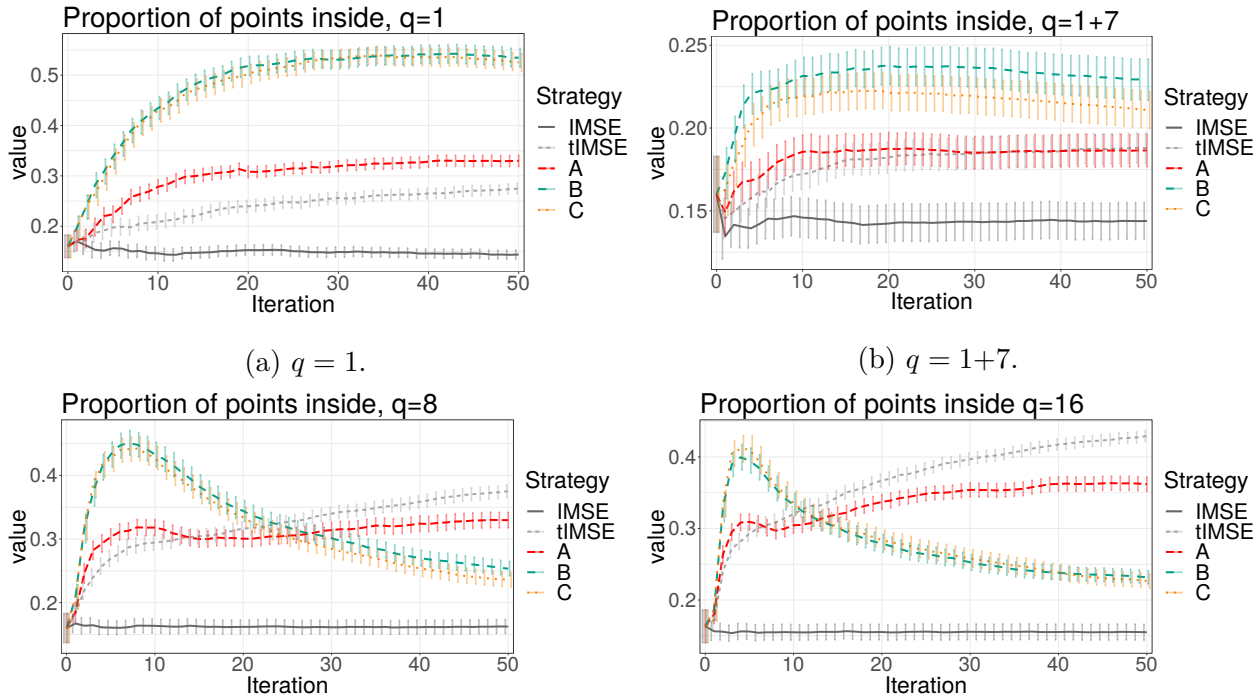
(c) $q = 8$.

(d) $q = 16$.

Figure 6: Expected Type II error in different batch sequential scenarios.

between strategies is similar however the differences between strategies are less important. This is due to the effect of adding 7 new input points with the same space-filling strategy independently of the criterion used to select the first point.

Figure 6 shows that, in this example, the parallel scenarios provide a much faster convergence for all strategies considered. Another aspect to consider is wall-clock time: under the assumption that Monte Carlo samples can be evaluated in parallel, then all scenarios require the same wall-clock time. In some cases, however, the MC samples required for evaluating the function at one new input can be computed only on one computational node. Then the procedure with $q = 1$ would be sequential so the wall-clock time would be $O(nn_{MC})$, however, for $q > 1$, each new input could be evaluated in parallel and the wall-clock time would become $O(kn_{MC})$, where $k = \frac{n}{q}$. Therefore, in case of non-parallelizable MC computations, a greatly reduced wall-clock time would also be an additional benefit of batch sequential scenarios with $q > 1$. Note that here the noise is set relatively low in all scenarios. In high noise scenarios, where a strong trade-off between noise and number of evaluations



(a) $q = 1$.

(b) $q = 1{+}7$.

(c) $q = 8$.

(d) $q = 16$.

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

is required, the situation is less clear. In supplementary material we present an example where the observations have higher noise variance. In such example, batch-sequential strategies still outperform sequential ones however the difference between scenarios is less pronounced.

4.2 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 $\frac{\#\{j:z_j \geq t,\ j=1,\ldots,n_i\}}{n_i}$, where n_i is the total number of points at iteration i and $z_1, \ldots, z_{n_i}$ are the evaluations. Figure 7 shows the proportion of points inside the excursion set at each iteration for the three scenarios outlined in the previous section. Strategy IMSE is a space filling strategy therefore the proportion of points inside the excursion approximates the volume of excursion. In the

scenario $q = 1$, the strategies have not yet reached convergence, therefore both B and C tend to select more points inside the set than A and tIMSE to consolidate the conservative estimation. The effect of the 7 points chosen with a space-filling strategy in scenario $q = 1 + 7$ is clear in figure 7b where all strategies show proportions very close to IMSE which remain stable as the iteration number grows. On the other hand, figure 7c and figure 7d again show a fast increase in the proportion of points inside for strategies B and C, however this is a transitory behavior and this proportion starts to decrease already after iterations 8 and 10 respectively. This highlights a tendency to explore the space by those strategies which was also verified visually by looking at the sequence of DoEs. Strategies A and tIMSE instead tend to choose points around the boundary of the set therefore they initially choose fewer points inside the set and even after convergence they do not show an exploratory behavior.

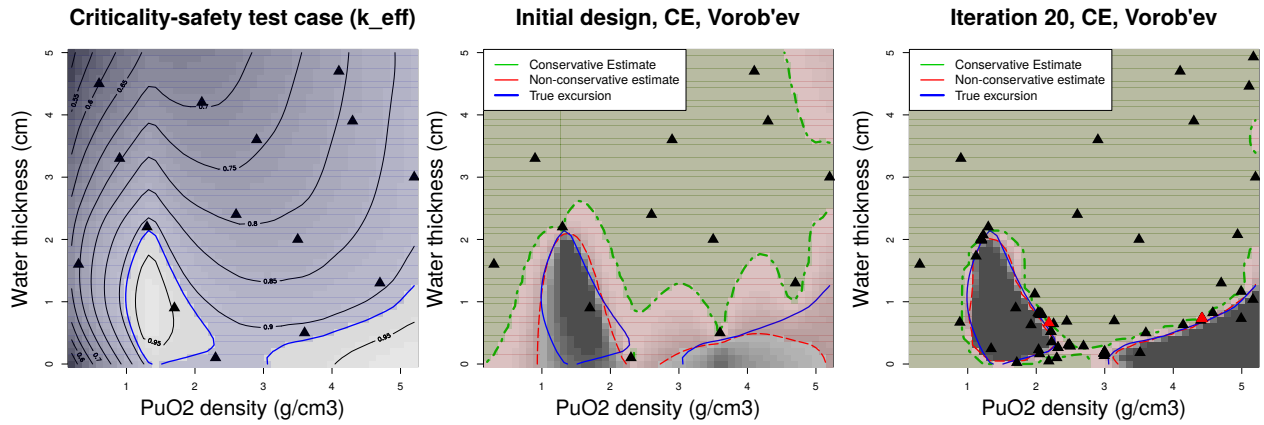
5 Reliability engineering test case

In reliability engineering applications, the set $\Gamma(f)$ in equation (1) often represents safe inputs for a system. In such settings, it is vital to avoid flagging unsafe regions as safe.

Figure 8 shows an example of such reliability engineering applications: a test case from the French Institute for Radiological Protection and Nuclear Safety (IRSN). The problem concerns a nuclear storage facility and we are interested in estimating the set of parameters that lead to 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 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

$$\Gamma(\text{k-eff}) = \{(\text{PuO}_2, \text{H}_2\text{O}) \in \mathbb{X} : \text{k-eff}(\text{PuO}_2, \text{H}_2\text{O}) \leq 0.92\}, \tag{11}$$

where the threshold $t = 0.92$ was chosen, for safety reasons, lower than the true critical case (k-eff $= 1.0$) where an uncontrolled chain reaction occurs. Figure 8a shows the set Γ(k-eff) shaded in blue and the contour levels for the true function computed from evaluations over a 50×50 grid, used as ground truth. The true data result from a MCMC simulation



(a) Function k-eff, set of interest (shaded blue, vol(Γ(k-eff))=0.8816 vol(𝕏)) and initial DoE ($n = 15$).

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

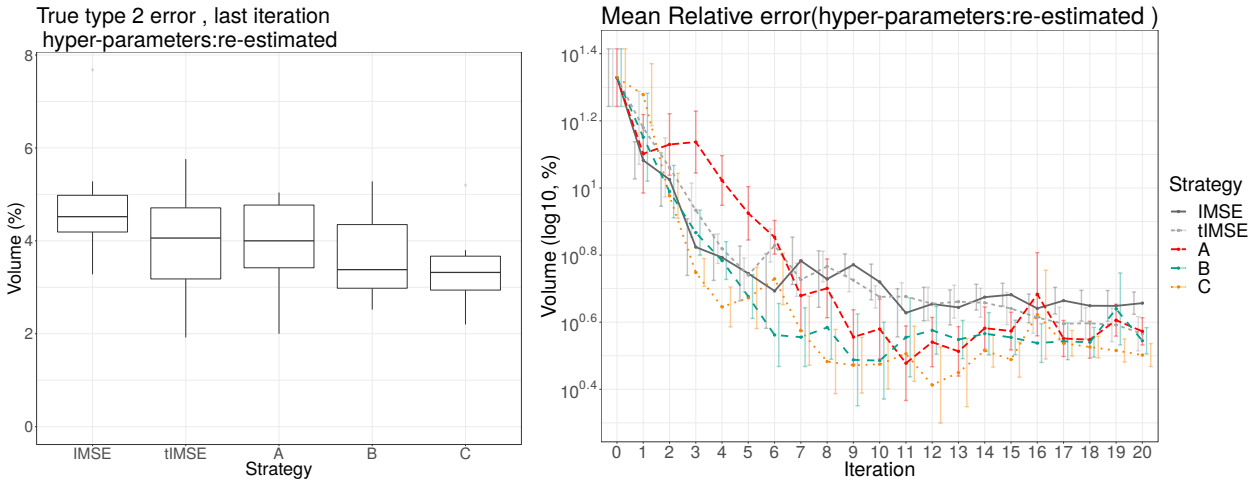
(c) Conservative ($\alpha = 0.95$, green) and non-conservative estimate (Vorob'ev expectation, red) after 75 evaluations.

Figure 8: 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.

and have a heterogeneous noise variance. Here we consider the k-eff function in figure 8 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 8a. We consider the five strategies listed in table 3 and we compare them on $m_{\text{doe}} = 10$ different initial DoEs of size $n_0 = 15$ obtained with the function `optimumLHS` from the package `lhs` (Carnell, 2019) in **R**.

Figure 8b shows a conservative estimate at level $\alpha = 0.95$ (shaded green) and a non conservative one (Vorob'ev expectation, shaded red) obtained from one of the 10 DoEs, the true set Γ(k-eff) is delimited in blue. Figure 8c shows that, as more evaluations are available, conservative and non-conservative estimates both get closer to the true safe set.



(a) True type II error, last iteration.

(b) Relative volume error versus iteration number.

Figure 9: Nuclear criticality safety test case, randomized initial DoEs.

The estimates in this example are computed from 75 function evaluations, where the last 60 points were selected sequentially with strategy C.

We now test how to adaptively reduce the uncertainty on the estimate with the strategies in table 3. 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 9a 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. 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 9b 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 $\Gamma(f)$ obtained from evaluations of k-eff on a 50×50 grid. The volume of $\Gamma(f)$ 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, and 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 yield the best uncertainty reductions both in terms of relative volume error and type II error.

6 Discussion

In this paper we introduced sequential uncertainty reduction strategies for conservative estimates. Such set estimates proved to be useful in a reliability engineering example, however they could be of interest in any situation 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. They were introduced with a homogeneous noise variance observation model, however as shown in appendix A, the criteria implementations are available also in the heterogeneous noise variance case. Under such observation model, estimating the heteroskedastic noise variance structure can be challenging, see Binois et al. (2018) for more details. The numerical studies presented in the homogeneous and noise-free cases showed that adaptive strategies provide a better uncertainty reduction than generic strategies. In particular, strategy C, i.e. the criterion $J_n^{\mathrm{T2}}(\cdot; \rho_n^\alpha)$, is 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 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 and, in supplementary material, we report a small preliminary study on this aspect. On the other hand, a fully Bayesian approach, accounting for hyper-parameter uncertainty, could be used to strengthen the procedure's overall conservativeness.

SUPPLEMENTARY MATERIAL

Supplementary Materials: additional test case, more details on the numerical benchmarks and theoretical complements to section 3.1 and appendix B.2. (pdf)

On-line code: a git repository that allows (partial) reproducibility for the experiments in sections 4 and 5 and supplementary material sections 4 and 5 is available at supplemental_adoece. (git repository)

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. (2019). A supermartingale approach to Gaussian process based sequential design of experiments. *Bernoulli*, 25(4A):2883–2919.

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.

Binois, M., Gramacy, R. B., and Ludkovski, M. (2018). Practical Heteroscedastic Gaussian Process Modeling for Large Simulation Experiments. *Journal of Computational and Graphical Statistics*, 27(4):808–821.

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.

Carnell, R. (2019). *lhs: Latin Hypercube Samples*. R package version 1.0.1.

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.

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. (2018). *The Design and Analysis of Computer Experiments*. Springer New York, New York, NY.

Spodarev, E. (2014). *Limit Theorems for Excursion Sets of Stationary Random Fields*, pages 221–241. Springer International Publishing, Cham.

Stroh, R. (2018). *Sequential design of numerical experiments in multi-fidelity : Application to a fire simulator*. Theses, Université Paris-Saclay.

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 Fast-to-evaluate formulae for sequential strategies

In this section we prove two propositions that allow for the computation of the criteria in equations (9) and (10). We consider a more generic observation model than equation (2), where the noise variance τ is heterogeneous, i.e. $\tau(\cdot)$ is a function of $x \in \mathbb{X}$.

First we extend the result in Chevalier (2013) to any level ρ_n which is a function of past n observations.

Proposition 2 (Criterion J_n)**.** *Consider* $\Gamma(f) = \{x \in \mathbb{X} : f(x) \in T\}$ *with* $T = [t, +\infty)$*, where* $t \in \mathbb{R}$ *is a fixed threshold, then the criterion* J_n *defined by equation* (9) *can be expanded as*

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

where

$$a_{n+q}(u) = \frac{\mathfrak{m}_n(u) - t}{\mathfrak{s}_{n+q}(u)}, \qquad \mathbf{b}_{n+q}(u) = \frac{K_q^{-1}\mathfrak{K}_n(\mathbf{x}^{(q)}, u)}{\mathfrak{s}_{n+q}(u)}, \tag{13}$$

$$\gamma_{n+q}(u) = \mathbf{b}_{n+q}^T(u) K_q \mathbf{b}_{n+q}(u) \qquad p_n(u) = \Phi\left(\frac{\mathfrak{m}_n(u) - t}{\mathfrak{s}_n(u)}\right), \ u \in \mathbb{X},$$

with $\mathfrak{K}_n(\mathbf{x}^{(q)}, u) = (\mathfrak{K}_n(x_{n+1}, u), \ldots, \mathfrak{K}_n(x_{n+q}, u))^T$, $K_q = \mathfrak{K}_n(\mathbf{x}^{(q)}, \mathbf{x}^{(q)}) + \mathrm{diag}(\tau^2(\mathbf{x}^{(q)}))$ *is assumed invertible,* $\mathfrak{K}_n(\mathbf{x}^{(q)}, \mathbf{x}^{(q)}) = [\mathfrak{K}_n(x_{n+i}, x_{n+j})]_{i,j=1,\ldots,q}$, $\Phi_2(\cdot; \Sigma)$ *is the cumulative distribution of the bivariate centered Normal with covariance matrix* Σ *and* Φ *is the standard Normal cumulative distribution.*

Proof. Recall that

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

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

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

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

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 (13) 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, $N_1 \sim N(0,1)$. By equations (14) to (16), (18) and (19) we obtain equation (12).

□

We provide below a formulation for the SUR criterion J_n^{T2} in equation (10) which is fast-to-evaluate and allows for faster optimization.

Proposition 3 (Type II criterion)**.** *In the case* $\Gamma(f) = \{x \in \mathbb{X} : f(x) \in T\}$ *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 as*

$$\begin{aligned}
J_n^{\mathrm{T2}}(\mathbf{x}^{(q)}; \rho_n^\alpha) &= \mathbb{E}_{n,\mathbf{x}^{(q)}}\left[G_n^{(2)}(Q_{n+q,\rho_n^\alpha})\right] \\
&= \int_{\mathbb{X}} \Phi_2\left(\begin{pmatrix} a_{n+q}(u) \\ \Phi^{-1}(\rho_n^\alpha) - a_{n+q}(u) \end{pmatrix}; \begin{pmatrix} 1 + \gamma_{n+q}(u) & -\gamma_{n+q}(u) \\ -\gamma_{n+q}(u) & \gamma_{n+q}(u) \end{pmatrix}\right) d\mu(u).
\end{aligned} \tag{20}$$

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

The evaluation of J_n and J_n^{T2} 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 as in Chevalier et al. (2014c) or by fixing the integration points with space filling designs, such as a 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.

B Properties of conservative estimates

B.1 Conservative estimates and confidence regions

Consider an excursion set $\Gamma = \{x \in \mathbb{X} : Z_x \geq t\}$, $t \in \mathbb{R}$ 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 $\mu(Q_\rho^C) = \mu(\mathbb{X}) - \mu(Q_\rho)$, then we have that ρ^* is also the minimizer of $\rho \to \mu(Q_\rho^C)$ under the constraint $P(\Gamma^C \subset Q_\rho^C) \geq \alpha$. 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 $\mathrm{CE}_{\alpha,n}$. 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 $\rho_1 > \rho_2$,

$$Q_{\rho_1} \subset Q_{\rho_2}, \qquad Q_{\rho_1}, Q_{\rho_2} \in \mathfrak{C}. \tag{21}$$

We now detail how to compute $\mathrm{CE}_{\alpha,n}$ based on $\mathfrak{C}$, for a fixed $\alpha \in [0,1]$ from n observations. For each $\rho \in [0,1]$, we define the function $\psi_\Gamma : [0,1] \to [0,1]$ 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 (21). 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 $\rho = \rho_n^\alpha$ 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 $\mathrm{CE}_{\alpha,n} = Q_{n,\rho_n^\alpha}$ 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.

B.3 Proofs

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

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

$$\mathbb{E}\left[\mu(Q_\rho \Delta \Gamma)\right] \leq \mathbb{E}\left[\mu(M \Delta \Gamma)\right], \tag{22}$$

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

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

By applying the expectation on both sides, since $\mu(Q_\rho^C) = \mu(M^C)$, 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 (22). □

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

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

□