Adaptive design criteria motivated by a plug-in percentile estimator

Rodrigo Cabral Farias, Luc Pronzato, Maria-João Rendas

Chapter 1
Adaptive design criteria motivated by a plug-in percentile estimator

Rodrigo Cabral-Farias, Luc Pronzato and Maria-João Rendas

1.1 Introduction

Increasingly complex numerical models are involved in a variety of modern engineering applications, ranging from evaluation of environmental risks to optimisation of sophisticated industrial processes. Study of climat change is an extremely well-known example, while its current use in other domains like pharmaceutics (the so-called *in vitro* experiments), aeronautics or even cosmetics are less well known of the general public. These models allow the prediction of a number of variables of interest for a given configuration of a number of factors that potentially affect them. Complex models depend in general on a large number of such factors, and their execution time may range from a couple of hours to several days.

In many cases, collectively falling in the domain of *risk analysis*, the interest is in identifying how often, under what conditions, or how strongly, a certain phenomenon may happen. In addition to the numerical model that predicts the variable of interest, it is then necessary to define a probabilistic structure in the set of its input factors, most often using a frequenciest approach. "How often" requires then the evaluation of the probability of occurence of the event of interest, while "how strongly" implies the determination of the set of the most extreme possible situations. In the former case we face a problem of estimation of an exceedance probability, while in latter is usually referred to as percentile estimation. For instance, in a study of the risk of flooding in a given coastal region, in the first case we want to estimate the probability α that a certain level of inundation η will not be exceeded, while in the second we are interest in the inundation level η that, with probability α, is not exceeded. In the context of the current planetary concern

R. Cabral-Farias, L. Pronzato, M.-J. Rendas
Laboratoire I3S, Université Côte d'Azur — CNRS, 2000 rte des Lucioles, 06903, Sophia Antipolis, France, e-mail: cabral@i3s.unice.fr,Luc.Pronzato@cnrs.fr,rendas@i3s.unice.fr

with the rise of the sea level, in the first case we may want to estimate the probability that it does not exceed one meter in a given region, while in the second the goal is to estimate the rise that, with 99% probability, will not be exceeded. We remark that both when estimating an exceedance probability or a percentile, most often the user is also interested in the delineation of the set of corresponding input configurations, a problem that often goes by the name of estimation of excursion sets. In the previous sea level example, where the parameters may be the evolution of tides, wave amplitudes, etc., one is interested in finding set of configurations of these factors that would entail a rise higher than a given η. Note that while the estimation of such a set is closely related to the other two problems, the three problems are, formally speaking, distinct problems.

A brute force approach to any of these problems would consider running the mathematical/numerical model for a representative set of configurations of input factors, and checking for each one whether the condition of interest is met or not. While this may be a viable alternative for very simple models, with a very small number of input factors, it cannot produce an answer in feasible time in any of the examples mentioned above: the size of any truly "representative set" allowing the observation of the (most often rare) situations of interest would be gigantic, each individual model run taking itself a long time.

State of the art methods resort, instead, to carefully chosen sets of configurations of the input factors, which are incrementally chosen by taking into consideration all previous model runs: they adaptively sample the model's input space, concentrating the computational effort in regions that are close, in a convenient sense, to the target excursion set. While efficient adaptive methods for exceedance probabilities have been proposed in the past [1, 2], mush less work has been devoted to the estimation of percentiles, which, as we will see later, is fundamentally more difficult.

This paper investigates wether the efficient solutions available for the easier problem of estimation of an excursion set can help finding solutions to the closely related percentile estimation problem, providing increased efficiency when compared to current methods. To formally relate the two problems, we introduce first a new percentile estimator (section 1.3), whose error analysis (under the assumption of small errors) enables the definition of a new family of criteria (section 1.4). We discuss their numerical implementation and complexity (section 1.5), and present a numerical study comparing their performance (section 1.6.3). The results obtained in the set of case-studies considered confirm the idea behind this study, showing that estimates of the percentile obtained on designs incrementally build to estimate the probability of exceedance of the current percentile estimate, converge to the correct value even when started with a poor initial design and for difficult situations.

1.2 Problem formulation and background

1.2.1 Problem formulation

Consider a real scalar function $f : \mathcal{A} \to \mathbb{R}$, where $\mathcal{A} \subset \mathbb{R}^d$ and let $p_{\mathcal{A}}$ be the probability distribution of the input factors of f. We denote by x a generic point of $\mathcal{A}$, and by $f(x)$ the value output by f to input configuration x.

Two characteristics of f are of interest in this paper. The η-exceedance probability α_η, defined by[1]

$$\alpha_\eta \stackrel{\triangle}{=} \mathbb{E}_{x\sim p_{\mathcal{A}}} \{x \ : \ f(x) \geq \eta\} \ ,$$

and the α-percentile of f, which is the "inverse" of the exceedance probability

$$\eta_\alpha \stackrel{\triangle}{=} \{\inf_y \ : \ \alpha_y = \alpha\} \ .$$

Obviously, $\eta_{\alpha_{\eta^\star}} = \eta^\star$ and $\alpha_{\eta^\star_\alpha} = \alpha^\star$. It is also easily verified that η_α is a non-increasing function of α and α_η a non-increasing function of η.

The following two problems are addressed in this paper.

Problem 1 *Percentile estimation.*
Given a set of observations $\mathcal{X}_n \stackrel{\triangle}{=} \{(x_i, f(x_i)), i = 1, \ldots, n\}$ *and an* $\alpha \in [0, 1]$, *estimate the* α*-percentile of a function* f. *We denote it generically by* $\hat{\eta}_\alpha(\mathcal{X}_n)$.

Problem 2 *Chose next observation (design problem).*
Given a set of observations $\mathcal{X}_n$ *of function* f *and an* $\alpha \in [0, 1]$ *chose the next observation point* $x_{n+1} \in \mathcal{A}$ *that leads to the best* $\hat{\eta}_\alpha(\mathcal{X}_{n+1})$.

The problems above are not well-posed, since the notion of "best performance" is not properly defined in Problem 2, and no criterion is given for choosing $\hat{\eta}_\alpha(\mathcal{X}_n)$ in Problem 1. They can be made precise in a Bayesian framework, by endowing f with a probability structure, i.e., assuming that the function f is randomly drawn according to a given distribution. For entities living in an infinite-dimensional space, like f, this distribution usually takes the form of a Gaussian process [7], i.e., for any finite collection $X_n = \{x_i\}_{i=1}^n$ of n points in $\mathcal{A}$, the vector $f(X_n)$ is normally distributed, with mean $\mu(X_n)$ and a covariance matrix R_{X_n} whose (i, j)-the element is $[R_{X_n}]_{(i,j)} = C(x_i, x_j)$ for some symmetric semi-positive definite operator C. Under this assumption, the function values at any collection of finite points conditioned on observations $\mathcal{X}_n$ are jointly normal random variables, with known distribution (with a covariance $R_{f|\mathcal{X}_n}$ that depends on the design X_n and a mean $\mu_{f|\mathcal{X}_n}$ that depends also on the observed values $\{f(x_i)\}_{i=1}^n$). This induces a conditional

[1] Symbol $\stackrel{\triangle}{=}$ indicates definitions and $\mathbb{E}_p\{\cdot\}$ denotes expectation under distribution p.

distribution on the value of η_α (for any fixed $\alpha \in [0,1]$). In the same manner, the conditional distribution of the errors of any percentile estimator $\hat{\eta}_\alpha$ is also well defined. A good performance criterion for the choice of $\hat{\eta}_\alpha$ in Problem 1 is thus the minimum expected posterior square error. The optimal estimator $\hat{\eta}^\star_\alpha(\mathcal{X}_n)$ under this setting is simply its expected value under the posterior distribution:

$$\hat{\eta}^\star_\alpha(\mathcal{X}_n) \triangleq \mathbb{E}_{f|\mathcal{X}_n}\left[\eta_\alpha\right] \ . \tag{1.1}$$

The same criterion can be used to give a precise meaning to the notion of best sampling point in Problem 2: using a myopic one-step-ahead approach the best point is the one that yields minimum expected squared error of $\hat{\eta}^\star_\alpha$ if that point is added to the design.

Note that although simply defined, the optimal estimator above can only be computed by resorting to heavy numerical simulations from the posterior distribution of f, as discussed in [1, 4, 5]where both Problems 1 and 2 are addressed. We will come back to its implementation in a subsequent section of this chapter.

Problems analogue to those defined above can be formulated for the estimation of α_η, the exceedance probability. As above, the assumption of a Gaussian process prior allows the definition of criteria appropriate for both the estimation and design problems. However, as we present below, while the optimal Bayesian estimate above is hard to compute, the derivation of the optimal (minimum mean-square error) estimate $\hat{\alpha}^\star_\eta(\mathcal{X}_n)$ of α_η is much simpler. The expression for the estimate is the following [2]:

$$\hat{\alpha}^\star_\eta(\mathcal{X}_n) \triangleq \mathbb{E}_{f|\mathcal{X}_n}\left[\mathbb{E}_{x\sim p_\mathcal{A}}\left(I[f(x) \geq \eta]\right)\right] \ , \tag{1.2}$$

where $I[\cdot]$ is the indicator function of a set. Under the Gaussian assumption, analytical expressions for the posterior distribution of the indicator function at each domain point are known. Therefore, by exchanging the order of expectations in (1.2), we have

$$\hat{\alpha}^\star_\eta(\mathcal{X}_n) \triangleq \mathbb{E}_{x\sim p_\mathcal{A}}\left[\mathbb{E}_{f|\mathcal{X}_n}\left(I[f(x) \geq \eta]\right)\right] = \mathbb{E}_{x\sim p_\mathcal{A}}\left[\Phi\left(\frac{\mu_{f|\mathcal{X}_n}(x) - \eta}{\sqrt{r_{f|\mathcal{X}_n}(x)}}\right)\right] \ , \tag{1.3}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal variable and $\mu_{f|\mathcal{X}_n}(x)$ and $r_{f|\mathcal{X}_n}(x)$ are the mean and variance of the function at point x conditioned on the observations (see Appendix 1 for details on the evaluation of the conditional mean and variance). Nevertheless, the presence of the expected value over the input factors of f in the above expression ($\mathbb{E}_{x\sim p_\mathcal{A}}$) indicates that its precise computation may be problematic.

To summarise, while the assumption of a Gaussian process model for f enables a precise mathematical formulation of both Problems 1 and 2 for which optimal solutions are known, these solutions are computationally infeasible unless for simple functions defined over low-dimensional input spaces. We will

define later an alternative percentile estimator which we will subsequently use to derive sub-optimal design criteria, that may be efficient alternatives to the computationally demanding optimal Bayesian criterion. First, we present literature results that will prove useful in the rest of the paper.

1.2.2 Background

We start by presenting the more computational version of eq. (1.1) used in [1, 4] and the corresponding sampling criteria for point x_{n+1}. To approximate the expectation operator $\mathbb{E}_{f|\mathcal{X}_n}[\cdot]$ in (1.1), M_x independent and identically distributed (i.i.d.) samples $\{x^i\}_{i=1}^{M_x}$ are drawn from $p_{\mathcal{A}}$. Then, M_f realizations on the set of M_x points of the function, $f_i^j = \{f^j(x^i)\}_{j=1}^{M_f}$ are drawn from the posterior distribution. From the j-th set of M_x samples, one sample from the posterior distribution of the percentile η_α^j can be approximately obtained as follows:

$$\begin{aligned} \eta_\alpha^j &\triangleq \eta \ : \ \mathbb{E}_{x\sim p_{\mathcal{A}}}\left(I\left[\mathbb{E}_{f|\mathcal{X}_n}[f(x)] \geq \eta\right]\right) = \alpha \\ &\simeq f^j_{(\lfloor M_x(1-\alpha)\rfloor)} \ . \end{aligned} \tag{1.4}$$

where $f^j_{(r)}$ denotes the r-th rank order statistic of f_i^j. Using these posterior samples of $\eta|_{\mathcal{X}_n}$ the optimal estimator can then be approximated by

$$\hat{\eta}_\alpha^\star(\mathcal{X}_n) \simeq \frac{1}{M_f}\sum_{j=1}^{M_f} \eta_\alpha^j \triangleq \hat{\eta}_\alpha^{\mathrm{MC}} \ . \tag{1.5}$$

In the same manner, the mean square error of its error is also directly approximated using a number of realisations f^j simulated from the posterior distribution of f:

$$\mathbb{E}_{f|\mathcal{X}_n}\left[\left(\eta_\alpha(f) - \hat{\eta}_\alpha(f)\right)^2\right] \simeq \frac{1}{M_f}\sum_{j=1}^{M_f}\left(\eta_\alpha^j - \hat{\eta}_\alpha^{\mathrm{MC}}\right)^2 \ , \tag{1.6}$$

Clearly, the approximation (1.6) above could be used as an ideal sampling criterion to choose x_{n+1}, since it directly approximates the mean square error of $\hat{\eta}^\star(\mathcal{X}_n)$. However, since it depends on $f(x_{n+1})$, it does not define a proper design criterion.

A solution to the issue above, presented in [1] and [4], is to consider that the estimate depends on the new design point x_{n+1} but the expectation in the mean squared error criterion is conditioned only on the previous observations $\mathcal{X}_n$. Using nested conditional expectations, we can write this criterion as a function of $x = x_{n+1}$ as follows:

$$J_n^\star(x;\alpha) = \mathbb{E}_{f|\mathcal{X}_n}\left[\left(\eta_\alpha - \hat{\eta}_\alpha(\mathcal{X}_{n+1})\right)^2\right]. \tag{1.7}$$

This criterion can be approximated at a point x by the following Monte-Carlo method: first, N_f realisations $\{f^k\}_{k=1}^{N_f}$ are drawn according to the posterior distribution of $f|\mathcal{X}_n$. Then, a posterior sample of the percentile η_α^k is calculated as the $(\lfloor M_x(1-\alpha)\rfloor)$-th order statistics, and the estimate $\hat{\eta}_\alpha^{\mathrm{MC}}(\mathcal{X}_{n+1}^k)$ is computed relying itself on M_f independent realizations of $f|\mathcal{X}_{n+1}^k$. Finally, $J_n^\star(x;\alpha)$ can be approximated as follows:

$$J_n^\star(x;\alpha) \simeq \frac{1}{N_f}\sum_{k=1}^{N_f}\left(\eta_\alpha^k - \hat{\eta}_\alpha^{\mathrm{MC}}(\mathcal{X}_{n+1}^k)\right)^2 . \tag{1.8}$$

Note that the computational cost of this criterion is very high, since it requires drawing N_f realisations of $f|\mathcal{X}_n(x_{n+1})$ and $N_f M_f$ trajectories of $f|\mathcal{X}_n$. Moreover, evaluation of each $\hat{\eta}_\alpha^{\mathrm{MC}}(\mathcal{X}_{n+1}^k)$ requires one sorting operation on M_x points.

An alternative criterion for adaptive estimation of a percentile has been presented in [5]. It is based on the (sub-optimal) empirical estimate of η_α implicitly defined by:

$$\hat{\eta}_\alpha^{\mathrm{emp}}(\mathcal{X}_n) \stackrel{\triangle}{=} \eta \ : \ \alpha = \mathrm{E}_{x\sim p_{\mathcal{A}}}\left(I\left[\mathbb{E}_{f|\mathcal{X}_n}[f(x)] \geq \eta\right]\right) . \tag{1.9}$$

This estimator can be approximated by Monte-Carlo using the $(\lfloor M_x(1-\alpha)\rfloor)$-th order statistic of the predicted function evaluated at randomly drawn M_x input values. Observe that the empirical estimate $\hat{\eta}_\alpha^{\mathrm{emp}}$ directly uses the Bayes estimate of f as if it was the true function, completely neglecting the uncertainty about its unobserved values.

In [5], the authors present two criteria for sampling x_{n+1} based on dispersion measures of the estimator $\hat{\eta}_\alpha^{\mathrm{emp}}(\mathcal{X}_{n+1})$ conditioned on $\mathcal{X}_n$. One criterion is based on the minimization of

$$J_n^{\mathrm{prob}}(x;\alpha) = \left|\mathbb{E}_{f|\mathcal{X}_n}\left[\alpha_{\hat{\eta}_\alpha^{\mathrm{emp}}(\mathcal{X}_n)}(f)\right] - \alpha\right| , \tag{1.10}$$

where $\alpha_{\hat{\eta}_\alpha^{\mathrm{emp}}(\mathcal{X}_{n+1})}(f)$ denotes the exceedance probability for a percentile fixed at $\hat{\eta}_\alpha^{\mathrm{emp}}(\mathcal{X}_{n+1})$ as a function of the random function f. This criterion corresponds to the absolute deviation from the target probability of exceedance α that we obtain with respect to the optimal estimate of α using $\mathcal{X}_n$ and setting the percentile to its estimate $\hat{\eta}_\alpha^{\mathrm{emp}}(\mathcal{X}_{n+1})$.

The second sampling criterion is based on the conditional variance of the empirical estimator:

$$J_n^{\mathbb{V}\mathrm{ar}}(x;\alpha) = \mathbb{V}\mathrm{ar}_{f|\mathcal{X}_n}\left[\hat{\eta}_\alpha^{\mathrm{emp}}(\mathcal{X}_{n+1})\right] . \tag{1.11}$$

The authors of [5] argue that this criterion should be seen as an information criterion and, as such, the new design point x_{n+1} should maximize it.

A detailed analysis of which order relations on the predicted function values may be inverted when a new observation is added to the dataset is put forward in [5]. Using results of this analysis, the authors show that analytical expressions for $J_n^{\mathrm{prob}}(x;\alpha)$ and $J_n^{\mathbb{V}\mathrm{ar}}(x;\alpha)$ can be obtained. Even if the evaluation of $J_n^{\mathrm{prob}}(x;\alpha)$ and $J_n^{\mathbb{V}\mathrm{ar}}(x;\alpha)$ does not require computationally expensive conditional sampling of the trajectories of f as in the approximation (1.8) of $J_n^{\star}(x;\alpha)$, the computational complexity involved in the evaluation of J_n^{prob} and $J_n^{\mathbb{V}\mathrm{ar}}$ is nonetheless still high, requiring $O(M_x^2)$ operations.

Unfortunately, no MATLAB code implementing these two criteria is publicly available. For this reason, the numerical study presented in a later section of this chapter will not consider them.

The estimators presented above are deeply rooted in the inverse relation between percentiles and exceedance probabilities. Several authors have addressed the definition of adaptive design algorithms for the estimation of the η-exceedance probability α_η. Most notably in [2], several sampling criteria adapted for this setting have been introduced. They are all based on the expectation of quantities related to the mean squared error of the optimal estimate $\hat{\alpha}_\eta^{\star}(\mathcal{X}_{n+1})$ conditioned on $\mathcal{X}_n$ and collectively presented under the name of stepwise uncertainty reduction (SUR) criteria.

While in the initial publication [2] the conditional expectations involved in the SUR criteria are approximated by Monte-Carlo, the subsequent reference [3] presents analytical expressions for some of the SUR criteria which require integration only over the input factors. Most remarkably, it is shown [3] that, while the exact determination of the conditional expected mean-square value of $\hat{\alpha}_\eta^{*}(\mathcal{X}_{n+1})$ involves a double integration over $\mathcal{A}$, a simple upper bound on the criterion – which we will designate here simply by $J_n^{\mathrm{SUR}}(x;\alpha)$ – is much easier to compute, requiring only a simple integration over $\mathcal{A}$ (see Appendix 1). Since the upper bound $J_n^{\mathrm{SUR}}(x;\alpha)$ is much easier to compute than the criterion directly targeting the mean square error and often leads to designs with similar performance, in this paper we use it as the state-of-the-art design criterion for the estimation of exceedance probabilities.

1.3 The plug-in estimator

We have seen above that the empirical estimate $\hat{\eta}_\alpha^{\mathrm{emp}}$ completely neglects the uncertainty affecting the predicted values of f. We propose in this paper a new estimator, which is implicitly defined by

$$\hat{\eta}_\alpha^{\mathrm{plg}}(\mathcal{X}_n) \stackrel{\triangle}{=} \eta \;:\; \alpha = \mathbb{E}_{x\sim p_{\mathcal{A}}}\left(I\left[\mathbb{E}_{f|\mathcal{X}_n}[f(x)\geq\eta]\right)\right) \;. \tag{1.12}$$

This new estimator is thus defined as the percentile that would lead to an optimal estimate of α, see eq. (1.2), equal to the target α:

$$\alpha = \hat{\alpha}^{\star}_{\hat{\eta}^{\rm plg}_{\alpha}}(\mathcal{X}_n) \ . \tag{1.13}$$

Note the subtle but important distinction between (1.9) and (1.12) regarding the expression to which conditional expectation is applied, $f(x)$ in the empirical estimate, which leads to the predicted value at point x, and the event $f(x) \geq \eta$ in the expression above, which is the conditional (given $\mathcal{X}_n$) probability that the function at x is larger than η. The plug-in estimator defined above can be considered as a compromise between the simple empirical estimate, which totally neglects uncertainty

$$p_{f|\mathcal{X}_n}(u) \leftarrow \delta\left(u - \mathbb{E}_{f|\mathcal{X}_n}[f(x)]\right) \ ,$$

where $\delta(\cdot)$ is the Dirac's delta measure, and the consideration, as done by the optimal estimate, of the full posterior distribution for f, which incorporates the statistical dependency between its values at different points x.

Numerical computation of $\hat{\eta}^{\rm plg}_{\alpha}(\mathcal{X}_n)$ is efficiently done by numerical search of the (scalar) root of the monotone equation $\alpha - \hat{\alpha}^{\star}_{\eta}(\mathcal{X}_n) = 0$. The search can be initialised at any simple estimate, e.g. at $\eta = \hat{\eta}^{\rm emp}(\mathcal{X}_n)$.

1.4 Adaptive "plug-in" criteria

We exploit now the dual relation (1.12) between $\hat{\eta}^{\rm plg}_{\alpha}$ and $\hat{\alpha}^{\star}_{\eta}$. Under the assumption of small errors, defining $\hat{\eta}^{\rm plg}_{\alpha}$ as the solution of (1.13) allows us to establish an approximate relation between the error in the estimation of the exceedance probability for a given percentile and the error of estimation of the percentile corresponding to that exceedance probability.

Rewrite (1.13) as

$$\alpha = \mathbb{E}_{x\sim p_{\mathcal{A}}}\left[\int_{\hat{\eta}^{\rm plg}_{\alpha}(\mathcal{X}_n)}^{\infty} p_{f|\mathcal{X}_n}(u)\,{\rm d}u\right]$$

and define

$$F_n(\alpha,\eta) \stackrel{\triangle}{=} \alpha - \mathbb{E}_{x\sim p_{\mathcal{A}}}\left[\int_{\eta}^{\infty} p_{f|\mathcal{X}_n}(u)\,{\rm d}u\right] \ . \tag{1.14}$$

For both $\hat{\alpha}^{\star}_{\eta}(\mathcal{X}_n)$ and $\hat{\eta}^{\rm plg}_{\alpha}(\mathcal{X}_n)$, F_n is zero:

$$F_n(\hat{\alpha}^{\star}_{\eta}(\mathcal{X}_n),\eta) = 0, \qquad F_n(\alpha,\hat{\eta}^{\rm plg}_{\alpha}(\mathcal{X}_n)) = 0 \ .$$

Denote the ideal (mean square error) Bayesian sampling criteria for x_{n+1} targeting the mean square error of $\alpha^{\star}_{\eta}(\mathcal{X}_{n+1})$ and $\hat{\eta}^{\rm plg}_{\alpha}(\mathcal{X}_{n+1})$ by $J^{\star}_n(x;\eta)$ and

$J_n^{\mathrm{plg}}(x;\alpha)$, respectively:

$$J_n^\star(x;\eta) \triangleq \mathbb{E}_{f|\mathcal{X}_n}\left[(\alpha_\eta - \hat{\alpha}_\eta^\star(\mathcal{X}_{n+1}))^2\right] ,$$
$$J_n^{\mathrm{plg}}(x;\alpha) \triangleq \mathbb{E}_{f|\mathcal{X}_n}\left[(\eta_\alpha - \hat{\eta}_\alpha^{\mathrm{plg}}(\mathcal{X}_{n+1}))^2\right] .$$

In a first-order approximation, valid if $\hat{\eta}_\alpha^{\mathrm{plg}}(\mathcal{X}_{n+1})) \simeq \eta_\alpha$, we have

$$\begin{aligned} F_{n+1}(\alpha, \hat{\eta}_\alpha^{\mathrm{plg}}(\mathcal{X}_{n+1})) \simeq & F_{n+1}(\hat{\alpha}_\eta^\star(\mathcal{X}_{n+1}), \eta_\alpha) \\ & + \left.\frac{\partial F_{n+1}}{\partial \alpha}\right|_{\hat{\alpha}_\eta^\star(\mathcal{X}_{n+1}),\eta_\alpha} (\alpha - \hat{\alpha}_\eta^\star(\mathcal{X}_{n+1})) \\ & + \left.\frac{\partial F_{n+1}}{\partial \eta}\right|_{\hat{\alpha}_\eta^\star(\mathcal{X}_{n+1}),\eta_\alpha} (\hat{\eta}_\alpha^{\mathrm{plg}}(\mathcal{X}_{n+1}) - \eta_\alpha) , \end{aligned}$$

where η_α denotes the true percentile. Since

$$\left.\frac{\partial F_{n+1}}{\partial \alpha}\right|_{\hat{\alpha}_\eta^\star(\mathcal{X}_{n+1}),\eta_\alpha} = 1, \quad \left.\frac{\partial F_{n+1}}{\partial \eta}\right|_{\hat{\alpha}_\eta^\star(\mathcal{X}_{n+1}),\eta_\alpha} = \mathbb{E}_{x\sim p_{\mathcal{A}}}\left[p_{f|\mathcal{X}_{n+1}}(\eta_\alpha)\right] ,$$

we obtain a relation between the error in the estimation of η and the error in the estimation of α:

$$\left(\hat{\eta}_\alpha^{\mathrm{plg}}(\mathcal{X}_{n+1}) - \eta_\alpha\right) \simeq \left(\left.\frac{\partial F_{n+1}}{\partial \eta}\right|_{\hat{\alpha}_\eta^\star(\mathcal{X}_{n+1}),\eta_\alpha}\right)^{-1} \left(\alpha - \hat{\alpha}_\eta^\star(\mathcal{X}_{n+1})\right) .$$

Applying expectation to the previous equation conditioned on the available observations $\mathcal{X}_n$, leads us to the following approximate relation between $J_n^{\mathrm{plg}}(x;\alpha)$ and $J_n(x;\eta)$:

$$J_n^{\mathrm{plg}}(x;\alpha) \simeq \mathbb{E}_{f|\mathcal{X}_n}\left[\left(\mathbb{E}_{x\sim p_{\mathcal{A}}}\left[p_{f|\mathcal{X}_{n+1}}(\eta_\alpha)\right]\right)^{-2}\left(\alpha - \hat{\alpha}_\eta^\star(\mathcal{X}_{n+1})\right)^2\right] . \quad (1.15)$$

The expression above indicates that when estimating η the error is the smallest for points x which inclusion in the design lead to large values of $\mathbb{E}_{x\sim p_{\mathcal{A}}}\left[p_{f|\mathcal{X}_{n+1}}(\eta_\alpha)\right]$, i.e., that may increase the average value (over $x \sim p_{\mathcal{A}}$) of the value at η of future posterior densities.

Points for which $p_{f(x)|\mathcal{X}_n}(\eta) \simeq 0$ will most probably lead to a value $\mathbb{E}_{x\sim p_{\mathcal{A}}}\left[p_{f|\mathcal{X}_{n+1}}(\eta_\alpha)\right] \simeq \mathbb{E}_{x\sim p_{\mathcal{A}}}\left[p_{f|\mathcal{X}_n}(\eta_\alpha)\right]$. However, the addition of points x at which the uncertainty of $p_{f(x)|\mathcal{X}_n}$ is small while its value at η_α is large (meaning that x belongs to region of the level line $f(x) = \eta_\alpha$ where the function is well known) only marginally increase the value of $\mathbb{E}_{x\sim p_{\mathcal{A}}}\left[p_{f|\mathcal{X}_{n+1}}(\eta_\alpha)\right]$, by slightly improving knowledge about the function values in that region. On the contrary, addition to X_n of points x belonging to regions where $p_{f(x)|\mathcal{X}_n}(\eta_\alpha)$ is bounded away from zero but the uncertainty of $p_{f(x)|\mathcal{X}_n}$ is large may result in a significant increase with respect to $\mathbb{E}_{x\sim p_{\mathcal{A}}}\left[p_{f|\mathcal{X}_n}(\eta_\alpha)\right]$.

Criteria based on relation (1.15) appropriately target, thus, the exploration of regions whose possible inclusion (or not) in the region contributing to α is highly uncertain.

The conditional expectation in the right-hand-side of (1.15) does not define a proper design algorithm, for two reasons: first, the true value of η_α in $p_{f|\mathcal{X}_{n+1}}(\eta_\alpha)$ and $\hat{\alpha}^\star_\eta(\mathcal{X}_{n+1})$ is unknown. Second, even η_α was known, the expression involves the expectation of a complex nonlinear function of f without closed-form expression.

Below, four criteria motivated by (1.15) are be presented by assuming progressively strong hypotheses and/or simplifications.

1.4.1 Monte-Carlo approximation

Approximation of $J_n^{\rm plg}(x;\alpha)$ in equation (1.15) depends on the unknown η_α, and thus is not computable. A common solution is to replace the η_α by a current estimator $\hat{\eta}_\alpha(\mathcal{X}_n)$. Note that any of the estimators previously presented can be used: the optimal $\hat{\eta}^\star_\alpha(\mathcal{X}_n)$, the empirical $\hat{\eta}^{\rm emp}_\alpha(\mathcal{X}_n)$, or the plug-in $\hat{\eta}^{\rm plg}_\alpha(\mathcal{X}_n)$ estimator. The expectation with respect to $x \sim p_{\mathcal{A}}$ in (1.15) can then be approximated using Monte-Carlo.

Similarly to what has been presented for $J_n^\star(x;\alpha)$, see equation (1.8), the approximation of $J_n^{\rm plg}(x;\alpha)$ for each candidate point $x = x_{n+1}$ requires N_f i.i.d. samples $\{f^k\}_{k=1}^{N_f}$ drawn from the distribution of $f(x)|\mathcal{X}_n$. For each observation set $\mathcal{X}^k_{n+1}$ completed with the sample (x_{n+1}, f^k), we can compute $\mathbb{E}_{x\sim p_{\mathcal{A}}}[p_{f|\mathcal{X}^k_{n+1}}(\hat{\eta}_\alpha(\mathcal{X}_n))]$ and $\alpha^\star_{\hat{\eta}_\alpha(\mathcal{X}_n)}(\mathcal{X}^k_{n+1})$. Both of these quantities, which are expectations over $x \sim p_{\mathcal{A}}$, can be evaluated using Monte-Carlo, by sampling M_x samples from $x \sim p_{\mathcal{A}}$, leading to

$$J_n^{\rm plg}(x;\alpha) \simeq J_n^{\rm MC}(x;\alpha) \triangleq \frac{1}{N_f}\sum_{k=1}^{N_f}\left[\frac{(\alpha - \alpha^\star_{\hat{\eta}_\alpha(\mathcal{X}_n)}(\mathcal{X}^k_{n+1}))}{\mathbb{E}_{x\sim p_{\mathcal{A}}}[p_{f|\mathcal{X}^k_{n+1}}(\hat{\eta}_\alpha(\mathcal{X}_n))]}\right]^2 . \qquad (1.16)$$

1.4.2 Monte-Carlo approximation assuming independency

A simpler expression can be obtained by neglecting the statistical dependency between the two factors $(\mathbb{E}_{x\sim p_{\mathcal{A}}}[p_{f|\mathcal{X}_{n+1}}(\eta_\alpha)])^{-2}$ and $(\alpha - \hat{\alpha}^\star_\eta(\mathcal{X}_{n+1}))^2$ in (1.15):

$$J_n^{\rm plg}(x;\alpha) \simeq \mathbb{E}_{f|\mathcal{X}_n}\left[\left(\mathbb{E}_{x\sim p_{\mathcal{A}}}[p_{f|\mathcal{X}_{n+1}}(\eta_\alpha)]\right)^{-2}\right] J_n^\star(x;\eta_\alpha) = L_n(x;\alpha) J_n^\star(x;\eta_\alpha) \ ,$$

where we implicitly defined $L_n(x;\alpha)$. Note that both factors depend on η_α, which, again, must be replaced by its estimate $\hat{\eta}_\alpha(\mathcal{X}_n)$.

Factor $J_n^\star(x;\hat{\eta}_\alpha(\mathcal{X}_n))$ coincides with one of the SUR criteria studied in [2, 3], see section 1.2.2, which can be efficiently replaced by its upper-bound $J_n^{\mathrm{SUR}}(x,\hat{\eta}_\alpha(\mathcal{X}_n))$. Factor $L_n(x;\alpha)$ must be approximated by Monte-Carlo, as done for $J_n^{\mathrm{plg}}(x;\alpha)$:

$$L_n(x;\alpha) \simeq L_n^{\mathrm{MC}}(x;\alpha) \stackrel{\triangle}{=} \frac{1}{N_f}\sum_{k=1}^{N_f}(\mathbb{E}_{x\sim p_{\mathcal{A}}}[p_{f|\mathcal{X}_{n+1}^k}(\hat{\eta}_\alpha(\mathcal{X}_n))])^{-2} \ . \tag{1.17}$$

Replacing $L_n(x;\alpha)$ by this Monte-Carlo approximation, leads to a second design criterion ("independent / Monte-Carlo") for the estimation of the α-percentile of f:

$$J_n^{\mathrm{iMC}}(x;\alpha) \stackrel{\triangle}{=} L_n^{\mathrm{MC}}(x;\alpha)J_n^\star(x;\hat{\eta}_\alpha(\mathcal{X}_n)) \ . \tag{1.18}$$

1.4.3 Assuming independency and neglecting uncertainty

If we further neglect the uncertainty about the function values, as predicted by $f(x)|\mathcal{X}_n$, when evaluating the outer expectation in $L_n(x;\alpha)$, i.e., consider that

$$\mathcal{X}_{n+1} \simeq \tilde{\mathcal{X}}_{n+1} \stackrel{\triangle}{=} \{(x_1,f(x_1)),\cdots,(x_n,f(x_n)),(x,\mu_{f|\mathcal{X}_n}(x))\} \ , \tag{1.19}$$

where in the last element the unknown value $f(x)$ is approximated by $\mu_{f|\mathcal{X}_n}(x)$, $L_n(x;\alpha)$ can be further approximated by

$$\begin{aligned}\mathbb{E}_{f|\mathcal{X}_n}\left[\left(\mathbb{E}_{x\sim p_{\mathcal{A}}}[p_{f|\mathcal{X}_{n+1}}(\eta_\alpha)]\right)^{-2}\right] &\simeq \left(\mathbb{E}_{x\sim p_{\mathcal{A}}}[p_{f|\tilde{\mathcal{X}}_{n+1}}(\hat{\eta}_\alpha(\mathcal{X}_n))]\right)^{-2} \\ &\stackrel{\triangle}{=} K_n^{-2}(x;\hat{\eta}_\alpha(\mathcal{X}_n)) \ .\end{aligned} \tag{1.20}$$

Contrary to $L_n(x;\alpha)$, computation of the factor $K_n^{-2}(x;\hat{\eta}_\alpha(\mathcal{X}_n))$ does not require sampling from $f|\mathcal{X}_n$ at each candidate design point, but only a Monte-Carlo approximation of the expectation over $x\sim p_{\mathcal{A}}$. This leads to a third, simpler, design criterion ("independent / deterministic") for the estimation of the percentile:

$$J_n^{\mathrm{id}}(x;\alpha) \stackrel{\triangle}{=} K_n^{-2}(x;\hat{\eta}_\alpha(\mathcal{X}_n))\ J_n^\star(x;\hat{\eta}_\alpha(\mathcal{X}_n)) \ . \tag{1.21}$$

1.4.4 Using SUR design criterion for exceedance probability

A fourth design criterion $J_n^{\mathrm{SUR}}(x;\alpha)$ is obtained by simply dropping factor $\left(\mathbb{E}_{x\sim p_{\mathcal{A}}}\left[p_{f|\mathcal{X}_{n+1}}(\eta_\alpha)\right]\right)^{-2}$ within the expectation in equation (1.15), keeping only factor $\left(\alpha-\hat{\alpha}^\star_\eta(\mathcal{X}_{n+1})\right)^2$. This is equivalent to neglecting the variation of $\left(\mathbb{E}_{x\sim p_{\mathcal{A}}}\left[p_{f|\mathcal{X}_{n+1}}(\eta_\alpha)\right]\right)^{-2}$ with the candidate point x, which is approximately valid when the remaining uncertainty is small, i.e., when design X_n is rich., and further observation points do not significantly decrease the posterior uncertainty.

This leads to the use, for percentile estimation, of a SUR criterion $J_n(x;\alpha_\eta)$ for the estimation of the probability of exceedance of the percentile η_α. The dependency on η_α is handled as before, replacing it by the current estimate of η_α:

$$J_n(x;\alpha) \stackrel{\triangle}{=} J_n^\star(x;\hat{\eta}_\alpha(\mathcal{X}_n)) = \mathbb{E}_{f|\mathcal{X}_n}\left[(\alpha-\hat{\alpha}^\star_{\hat{\eta}_\alpha(\mathcal{X}_n)}(\mathcal{X}_{n+1}))^2\right] . \tag{1.22}$$

As previously discussed, this criterion can be efficiently approximated by the upper bound $J_n^{\mathrm{SUR}}(x;\hat{\eta}_\alpha(\mathcal{X}_n))$.

1.5 Numerical Implementation

We address now possible implementations of the two approximations on which the criteria presented in the last section are based: Monte-Calor approximation of expected values, and replacement of the true percentile by its current estimate.

All the alternative criteria presented in the previous section extensively ressort to Monte-Carlo to approximate expected values (over $x\sim p_{\mathcal{A}}$ and over $f(x)\sim f(x)|\mathcal{X}_n$). A major problem may affect the integration over the input factors, which we illustrate using $J_n^{\mathrm{ID}}(x;\alpha)$, given by the product of $J_n(x;\alpha)$ and $K_n(x;\alpha)$, see equation 1.20). Unless $\mu_{f|\mathcal{X}_{n+1}}(x^i)-\hat{\eta}_\alpha(\mathcal{X}_n)$ is within a few posterior standard deviations ($\rho\geq 4$), $p_{f(x^i)|\mathcal{X}_{n+1}}(\hat{\eta}_\alpha(\mathcal{X}_n))\simeq 0$, i.e. x^i does not significantly contribute to the expected value, and

$$K_n(x;\alpha)\simeq \frac{M_x^2}{\left(\sum_{x^i\in\mathcal{I}(\mathcal{X}_n)} p_{f(x^i)|\mathcal{X}_{n+1}}(\hat{\eta}_\alpha(\mathcal{X}_n))\right)^2} ,$$

where (remember that $r_{f|\mathcal{X}_n}(x)$ is the posterior variance at point x)

$$\mathcal{I}(\mathcal{X}_n)\stackrel{\triangle}{=}\{x\ :\ (\mu_{f|\mathcal{X}_n}(x)-\eta_\alpha)^2\leq\rho^2\cdot r_{f|\mathcal{X}_n}(x)\} .$$

When $r_{f|\mathcal{X}_n}(x)$ is small – and in particular in regions of fast variation of f – the set of Monte-Carlo points falling in set $\mathcal{I}(\mathcal{X}_n)$ can be empty unless a very dense sampling from $p_{\mathcal{A}}$ is done, i.e., unless M_x is very large. If the number of Monte Carlo samples inside $\mathcal{I}(\mathcal{X}_n)$ is not large enough, the numerical estimate of $\mathbb{E}_{x\sim p_{\mathcal{A}}}\left[p_{f(x)|\mathcal{X}_{n+1}}(\hat{\eta}_\alpha(\mathcal{X}_n))\right]$ will be close to zero for all candidate design points, failing to correctly indicate the expected error in the estimation of η_α if the points are added to the design. The same problem also affects, although to a lesser extent, the numerical determination of J_n^{SUR}, which also involves an expectation over $x \sim p_{\mathcal{A}}$.

Increasing M_x such that $\mathcal{I}(\mathcal{X}_n)$ is non-empty with large probability would require infeasibly large values. Instead, the problem can be overcome by using an importance sampling scheme, that increases the density of the samples in the regions where the integrated function may have values away from zero. We implemented an importance sampling method that adds new N_{rand} samples normally distributed around each candidate point that belongs to the set $\mathcal{I}(\mathcal{X}_n)$ defined above[2]:

$$p_{is} = \frac{M_x}{N_{rand}N_{is} + M_x} p_{\mathcal{A}} + \frac{N_{rand}}{N_{rand}N_{is} + M_x} \sum_{k=1}^{N_{is}} \mathcal{N}(x_k^{mc}; \sigma_{mc}^2) \ .$$

Above, $\{x_k^{mc}\}_{k=1}^{N_{is}}$ is the set of grid points in $\mathcal{I}(\mathcal{X}_n)$, and σ_{mc}^2 is chosen such that the density of Monte Carlo points is increased by N_{rand} times relative to $p_{\mathcal{A}}$ at each point in $\mathcal{I}(\mathcal{X}_n)$. The importance weights for the computation of the integral are thus $p_{\mathcal{A}}(x^i)/p_{is}(x^i)$.

Criteria J_n^{ID} and J_n^{iMC} both weight J_n^{SUR} by a multiplicative factor, L_n^{MC} and K_n^{-2}, respectively, that depend on the (posterior) density for the field values at the current percentile estimate $\hat{\eta}_\alpha$. As discussed before, these factors should induce concentration of future design points in the neighbourhood of the currently detected level-set $\{x \in \mathcal{A} \ : \mu_{f|\mathcal{X}_n}(x) = \hat{\eta}_\alpha\}$. This set cannot be empty when $\hat{\eta}_\alpha$ is the empirical estimate $\hat{\eta}_\alpha^{\mathrm{emp}}$, guaranteeing that the criteria will target the most relevant regions of the input space according to the observations made so far. However, this level-set can be empty for the plug-in estimate $\hat{\eta}_\alpha^{\mathrm{plg}}$, in particular when the value of α is close to either 0 or 1. If this happens, the numerical evaluation of the criterion becomes highly sensitive to the assumed probabilistic model as well as to the effectiveness of the Monte-Carlo integration. We expect thus that replacing η_α by $\hat{\eta}_\alpha^{\mathrm{emp}}$, instead of $\hat{\eta}_\alpha^{\mathrm{plg}}$ to be a more robust choice when implementing these criteria.

[2] $\mathcal{N}(\mu, \sigma^2)$ denotes the normal density with mean μ and variance σ^2.

1.6 Numerical study

1.6.1 Comparison study

The paper presents a comparative study of percentile estimators and adaptive design algorithms. The set of compared estimators is $\mathcal{E} \triangleq \{\hat{\eta}_\alpha^{\rm emp}, \hat{\eta}_\alpha^{\rm plg}, \hat{\eta}_\alpha^\star\}$,

- the plug-in estimate $\hat{\eta}_\alpha^{\rm plg}$, see eq. (1.12);
- the empirical estimate $\hat{\eta}_\alpha^{\rm emp}$, see eq. (1.9);
- the optimal Bayesian estimate $\hat{\eta}_\alpha^\star$ given by eq. (1.1).

The set $\mathcal{D} \triangleq \{J_n^{\rm SUR}, J_n^{\rm ID} J_n^{\rm iMC}, J_n^\star\}$ of the following four adaptive design criteria are considered

- the SUR criterion $J_n^{\rm SUR}(x;\alpha)$, see page 7 in section 1.2.2;
- the independent/deterministic criterion $J_n^{\rm ID}(x;\alpha)$, equation (1.21) in section 1.4.3;
- the independent/Monte-Carlo criterion $J_n^{\rm iMC}(x;\alpha)$, equation (1.18) in section 1.4.2;
- the optimal Bayesian criterion $J_n^\star(x;\alpha)$, equation 1.8 of section 1.2.2.

Four distinct implementations of criteria $J_n^{\rm SUR}, J_n^{\rm ID}$ and $J_n^{\rm iMC}$ are tested (see section 1.5)

- using $\hat{\eta}_\alpha^{\rm plg}$ with no importance sampling;
- using $\hat{\eta}_\alpha^{\rm emp}$ with no importance sampling;
- using $\hat{\eta}_\alpha^{\rm plg}$ with importance sampling;
- using $\hat{\eta}_\alpha^{\rm emp}$ with importance sampling.

We will occasionally denote by $\mathcal{I}$ the set of these four implementation choices.

The combination (d, i) of a design criterion $d \in \mathcal{D}$ and an implementation $i \in \mathcal{I}$ will be designated by "design algorithm". A total of $3 \times 4 + 1$ distinct design algorithms, collectively denoted by $\mathcal{A}_D$ are thus studied.

We designate by "solution" a combination of estimator and design algorithm. There are thus $4 \times 13 = 52$ different solutions, collectively represented by $\mathcal{S}$, under comparison.

1.6.2 Methodology

Comparison of design methods faces the same difficulties as the comparison of optimisation methods: the performance of each method depends strongly on the characteristics of the function it is applied to. Our study relies on the observation of the performance of the set of solutions $\mathcal{S}$ on a set of problems $\mathcal{C}$ with diverse characteristics, in particular of their robustness with respect to particular challenges attached to each problem $p \in \mathcal{C}$.

As far as possible, solutions are compared under the same conditions, adapted to the dimension of the input space of the case-study. In particular, the same Gaussian process model is used, with a constant trend and a Matèrn 5/2 isotropic kernel [7], and the same initial design is used by all solutions in all problems. Also, the geometry of the set of candidate design points is the same for all functions with the same dimension d of the input space of the function. Further details are given in section 1.6.3.

1.6.2.1 Case studies

The 52 distinct solutions in $\mathcal{S}$ described above will be compared in the same set of problems $\mathcal{C} = \left\{(f, \alpha), \alpha \in \alpha^{(f)}, f \in \mathcal{F}\right\}$ where

$$\mathcal{F} \stackrel{\triangle}{=} \{\text{Ackley}, \text{F1}, \text{Gramacy}, \text{Branin}, \text{Goldprice}\} \ ,$$

is the set of (deterministic, one- and two-dimensional) functions considered in the study and $\alpha^{(f)}$ defines the set of percentiles η_α estimated for function f, see Table 1.1. Each problem (or case-study) in $\mathcal{C}$ is thus a combination of a function and a value of α, generically denoted by f_α. There are a total of $|\mathcal{C}| = 11$ distinct problems (or case-studies) on which the solutions are tested. We will occasionally use notation $\mathcal{C}_d, d = 1, 2$ to denote the set of case-studies for d-dimensional functions, $|\mathcal{C}_1| = 7$ and $|\mathcal{C}_2| = 4$

Figure 1.1 gives a graphical representation of $\mathcal{C}$ (for $d = 2$ a 3D plot is given at the right of the corresponding rows). Analytical expressions of the test functions are given in Appendix 2.

Table 1.1 Test functions

function $\alpha^{(f)}$	d	Challenges
Ackley $\{0.3,\ 0.8\}$	1	Simple monotone function. The two values of α considered enable observation of the performance in regions of different gradient values.
F1 $\{0.1, 0.17, 0.32\}$	1	Quasi-oscillating function with several peaks. Precision up to three decimals is required to correctly identify all modes contributing to $\alpha^\cdot$
Gramacy $\{0.037, 0.1\}$	1	Non-stationary function.
Branin $\{0.02, 0.9\}$	2	Non-stationary function. Level sets with very different topologies for the values of α considered.
Goldprice $\{0.005, 0.14\}$	2	Non-stationary function. Level-sets with several connected components at distinct gradient values.

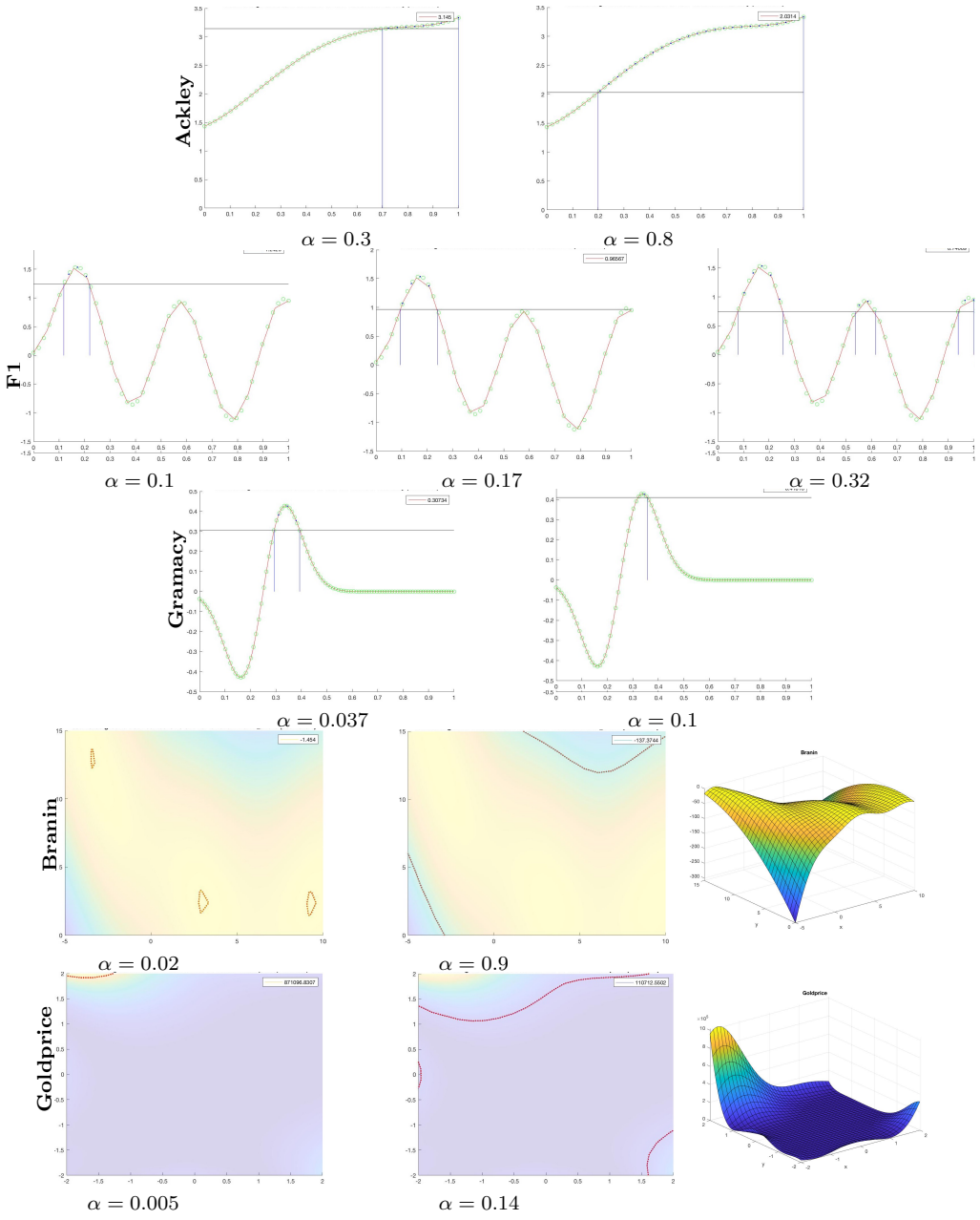

Fig. 1.1 Set of case-studies $\mathcal{C}$. Functions Ackley, F1, Gramacy, Branin and Goldprice (top to bottom) and corresponding level-sets for the set of percentiles indicated in Table 1.1.

1.6.2.2 Performance indicators.

Performance of both estimators in $\mathcal{E}$ and design algorithms in $\mathcal{A}_D$ is assessed through the errors in the estimation of the percentiles η_α:

- the relative performance of estimators is assessed by comparing the errors of the estimates obtained using *the same design points*, i.e., for the same problem $f_\alpha \in \mathcal{C}$ and the same $d_a \in \mathcal{A}_D$.
- the relative merit of the possible implementations of design criteria is assessed by comparing statistics of the errors, on the same case-study $f_\alpha \in \mathcal{C}$, of the same estimate $e \in \mathcal{E}$ using *designs of the same size* $k \in \{N_0 + 1, \ldots, N\}$, identified by same design criterion $d \in \mathcal{D}$.
- the relative performance of design criteria is assessed comparing the errors of the same estimate $e \in \mathcal{E}$ in the same case-study $f|_\alpha \in \mathcal{C}$, using the same implementation choice $i \in \mathcal{I}$.

The *performance profile* (of an estimator, a criterion or a design algorithm) is an aggregated plot that allows visualisation of overall relative performance [11]. Let $\epsilon_{s,p}(k)$ denote the errors observed when applying solution $s \in \mathcal{S}$ to case-study $f|_\alpha$ for a design of size k. Consider factorisations of the solution set the form $\mathcal{S} = \mathcal{T} \times \mathcal{G}$, where $\mathcal{T}$ is the set of solution choices that we want to compare. The performance profile of choice $t \in T$ for the (finite) set of case-studies $\mathcal{C}$ is the longitudinal curve indexed by k defined by

$$P_{t,\mathcal{C}}(k;t) \stackrel{\triangle}{=} \# \left\{ f|_\alpha \in \mathcal{C}, g \in \mathcal{G} \ : \ \epsilon_{(t,g),f|_\alpha} \ = \min_{t' \in \mathcal{T}} \epsilon_{(t',g),f|_\alpha}(k) \right\}, \qquad k = 1, 2, \ldots \tag{1.23}$$

i.e., $P_{t,\mathcal{C}}(k,t)$ counts the number of problems in $\mathcal{C}$ for which the solutions $s = (t, g)$ are the best over the set of solutions that only differ in t. For individual values of k we will designate $P_{t,\mathcal{C}}(k;t)$ as the *score* of t in $\mathcal{C}$.

1.6.3 Numerical results

This section presents the actual numerical comparison of design criteria, estimators and implementations, using the methodology outlined above. We consider first (section 1.6.3.1) the relative performance of the three estimators $\hat{\eta}_\alpha^{\text{emp}}, \hat{\eta}_\alpha^{\text{plg}}$ and $\hat{\eta}_\alpha^\star$. In a second step, we will address (section 1.6.3.2) the impact of the implementation of each of the criteria: which (current) estimator should be used, does importance sampling actually prevent possible numerical integration problems? Once the best (or preferable) estimator and most appropriate implementation choices are identified, we finally address (section 1.6.3.3) the comparison of the design criteria.

In all numerical experiments, the following choices and parameters have been used

- $p_{\mathcal{A}}$ is always the uniform distribution over the domain of f.
- Initial designs are regular rectangular grids (for all case-studies the input domain is an interval in $\mathbb{R}^d$) with $N_0 = 4$ points for $d = 1$ and $N_0 = 9$ points for $d = 2$.
- Design points are chosen amongst the elements of a regular grid covering the function domain. We use $N_C = 40$ for $d = 1$ and $N_C = 30 \times 30 = 900$ for $d = 2$.
- Importance sampling is done as presented in section 1.5, using $N_{rand} = 10$.
- The determination of an error requires the knowledge of the true value of the estimated η_α, which are unknown for the set of problems considered. We replace them by "ground truth" values $\eta_\alpha^{\rm gt}$ obtained, for case-study $f|_\alpha$, by using the entire set of N_C grid points. Whenever "error" is mentioned below, it means deviation with respect to $\eta_\alpha^{\rm gt}$.
- Design criterion $J_n^\star$ does not require an estimate of η_α, and is not affected by the numerical problems that motivate the use of importance sampling. There is thus a single implementation of thus criterion. Being based on Monte-Carlo, this estimator is random. To assess its variability, we performed four independent executions of $J_n^\star$ for each $f|_\alpha$. In the paper we only present results of the application of $J_n^\star$ to the one-dimensional case-studies $\mathcal{C}_1$, (using $M_f = N_f = 20$), since its application to problems in $\mathcal{C}_2$ would require larger values of M_f and N_f leading to impractically large computational times.
- All computations of field estimation and uncertainty characterisation based on Gaussian Process models rely on the Matlab package STK (Small Toolbox for Kriging) [8]. All models consider an isotropic Matèrn correlation and a linear trend.

1.6.3.1 Estimators performance

This section compares the performance of the estimators $\mathcal{E}$. Figure 1.2 shows the performance profiles – separately for $d = 1$ (left), and $d = 2$ (right) – of the three estimators: $\hat\eta_\alpha^{\rm emp}$ (red), $\hat\eta_\alpha^{\rm plg}$ (green) $\hat\eta_\alpha^\star$ (blue). i.e., for $\mathcal{T} \equiv \mathcal{E}$ in (1.23). For each design size k the errors of the three estimates are compared on a total of $N_{\rm pts} = N_{\mathcal{I}} \times N_{\mathcal{D}} \times N_{\mathcal{C}}$ distinct designs, where $N_{\mathcal{I}}$ is the number of distinct criteria implementations, $N_{\mathcal{D}}$ is the number of criteria and $N_{\mathcal{C}}$ is the number of case-studies Table 1.2 presents the values of $N_{\mathcal{I}}$, $N_{\mathcal{D}}$, $N_{\mathcal{C}}$ and $N_{\rm pts}$ for $d = 1, 2$.

The figure reveals a clear overall superiority of $\hat\eta_\alpha^{\rm emp}$, which, except for very small design sizes and $d = 1$, provides the most often the estimate with lowest error. No noticeable preference can be established between $\hat\eta_\alpha^{\rm plg}$ and $\hat\eta_\alpha^\star$.

Table 1.2 Size of test cases for performance profiles of Figure 1.2.

d	$N_{\mathcal{I}}$	$N_{\mathcal{D}}$	$N_{\mathcal{C}}$	N_{pts}	$N_{\text{pts}}^{f\|\alpha}(all)$	$N_{\text{pts}}^{f\|\alpha}(large)$
1	4	4	7	112	176	64
2	4	3	4	48	132	48

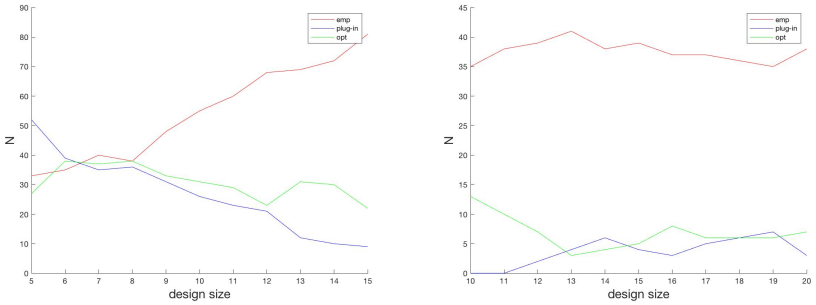

Fig. 1.2 Performance profiles of estimators (all implementations of all criteria, all case-studies, all adaptive designs). Left: 1D functions, right: 2D functions.

The performance plots in Figure 1.2 do not enable to appreciate how the relative performance may depend on the characteristics of each case-study. Table 1.3 presents a dual result, aggregating all design sizes, but computing separately for each $f|_\alpha$, the number of times $N_{\hat{\eta}_\alpha}^{f|_\alpha}$ that each $\hat{\eta}_\alpha \in \mathcal{E}$ had the lowest error. To enable observation of the behaviour at convergence a score N_s^L which considers only the largest 4 designs is also presented. The best estimate is shown in bold. The total number of design algorithms for each entry of this table, $N_{\text{pts}}^{f|\alpha} = N_{\mathcal{I}} \times N_{\mathcal{D}} \times N_{des}$, is indicated in Table 1.2 for $d = 1$ and $d = 2$ ($N_{des} = 11$ for all designs and $N_{des} = 4$ when only the four largest designs are considered).

The table confirms the large-sample superiority of $\hat{\eta}_\alpha^{\text{emp}}$ for all case-studies, in particular when $d = 2$ (last four lines of Table 1.3). This observation holds when all design sizes are considered, except for the markedly oscillating function F1 for which $\hat{\eta}_\alpha^{\text{plg}}$ has the lowest error best slightly more often than the other two estimates. Indeed, for this type of functions, the series predicted from small designs may span only a small subset of the actual function values, preventing the order statistic $\hat{\eta}_\alpha^{\text{emp}}$ from reflecting the actual distribution of function values. The other two estimates, by taking into account the uncertainty in the predicted field partially overcome this limitation, integrating the possibility of function values outside the observed range.

Quite surprisingly, Table 1.3 reveals a rather disappointing behaviour of $\hat{\eta}_\alpha^\star$. This may either indicate lack of robustness of this model-based estimator, which is heavily dependent on the prior probabilistic model for f, or that our

Table 1.3 Number of times each estimate has lowest error (over all implementations of all criteria, considering several design sizes).

Function/α	N^L_{plg}	N^L_{emp}	$N^L_\star$		N^A_{plg}	N^A_{emp}	$N^A_\star$
Ackely/0.3	0	**59**	5		16	**124**	36
Ackely/0.8	0	**52**	12		4	**123**	49
F1/0.1	6	**33**	25		**65**	51	60
F1/0.17	16	**37**	11		**82**	47	47
F1/0.32	13	**39**	12		**69**	54	53
Gramacy/0.1	11	**33**	20		37	**85**	54
Gramacy/0.037	6	**37**	21		21	**115**	40
Branin/0.02	16	**22**	10		22	**96**	14
Branin/0.9	4	**36**	8		5	**115**	12
Goldprice/0.005	1	**40**	7		12	**70**	49
Goldprice/0.14	0	**48**	0		0	**132**	0

choice of parameters M_f and N_f does not enable a sufficiently rich sampling from the posterior distributions.

The analysis above establishes the relative merits of the estimators. However, it does not enable assessment of how their relative merits may depend on the characteristics of f. For each $f \in \mathcal{F}$ let $\delta_{e,e';f}(k)$ denote, for each $f \in \mathcal{F}$ and pair of estimators (e, e'), the set of differences of their absolute errors on the designs of size k produced by the same design algorithm:

$$\delta_{e,e';f}(k) \stackrel{\triangle}{=} \left\{ \left|\epsilon_{(e,d),f_\alpha}(k)\right| - \left|\epsilon_{(e',d),f_\alpha}(k)\right|, \qquad d \in \mathcal{A}_D \ , \alpha \in \alpha^{(f)} \right\},$$

where $\mathcal{A}_D$ is the set of design algorithms, and $\alpha^{(f)}$ the set of values of α studied for function f.

Each histogram in Figures 1.3 and 1.4 uses sets $\{\delta_{e,e';f}(k)\}_{k=n}^{N}$. In Figure 1.3, $n = N_0 + 1$, while in Figure 1.4 $n = N - 3$, such that only the four largest designs are considered. When $f_\alpha \in \mathcal{C}_1$, the histograms in Figure 1.3 collect $N_{\text{des}} \times |\mathcal{A}_D| \times \alpha^{(f)}$ and thus 352 (528) values of $\delta_{e,e';f_\alpha}$ for functions Ackley and Gramacy (for function F1), and 264 data points for Branin and Goldprice. The histograms in Figure 1.4 use, respectively, 128 (192) and 96 error differences.

The height of the histograms around the origin reflects how frequently the error levels of estimates e and e' are similar, while their queues indicate situations when one estimator is significantly better than the other. The blue histograms correspond to $\delta_{\text{plg,emp};f}$: positive queues indicate that $\hat{\eta}^{\text{plg}}_\alpha$ has much larger error than $\hat{\eta}^{\text{emp}}_\alpha$, and negative queues the reverse situation. The green histograms, for $\delta_{\text{plg},\star;f_\alpha}$, compare the two model-based estimators $\hat{\eta}^{\text{plg}}_\alpha$ and $\hat{\eta}^{\star}_\alpha$: positive queues indicate outliers of $\hat{\eta}^{\text{plg}}_\alpha$ with respect to $\hat{\eta}^{\star}_\alpha$ and negatives ones outliers of $\hat{\eta}^{\star}_\alpha$. Finally, the red histograms collect values of $\delta_{\text{emp},\star;f_\alpha}$,with positive queues indicating outliers of $\hat{\eta}^{\text{emp}}_\alpha$ and strong negatives queues outliers of $\hat{\eta}^{\star}_\alpha$.

We present below a detailed analysis of these histograms for each function. Notation $\succeq$ reads "has slightly better performance then", $\succ$ "has better performance than", and $\simeq$ "has similar performance as".

- ($\hat{\eta}_\alpha^{\text{emp}} \succeq \hat{\eta}_\alpha^\star \succ \hat{\eta}_\alpha^{\text{plg}}$) Function Ackley is the "easiest" function considered: it is smooth and monotone. Indeed, the support of the corresponding histograms is the most narrow around the origin. For all data sizes $\hat{\eta}_\alpha^{\text{emp}}$ and $\hat{\eta}_\alpha^\star$ have **consistently similar errors**, with small probability of relative outliers, the red histograms having a strong peak at the origin and very weak queues. In both Figures, the other two (blue and green) histograms are very similar, both exhibiting a mode at positive values near the origin and a second mode at larger values, indicating that $\hat{\eta}_\alpha^{\text{plg}}$ **has larger errors than the other two estimates**, showing that neglecting the joint distribution of the predicted field does indeed lead to a degraded performance when the field is highly correlated as in this case. These two clusters correspond to the two values of α considered, that have distinct levels of difficulty (different derivatives of the function at the level set): the mode at larger values to $\alpha = 0.3$ and the other mode, at values almost one order of magnitude smaller, for $\alpha = 0.8$. This shows that the **percentile errors increase with the strength of the gradient of the function at the target percentile value**.
- ($\hat{\eta}_\alpha^{\text{plg}} \succeq \hat{\eta}_\alpha^\star \succ \hat{\eta}_\alpha^{\text{emp}}$) For function F1, it is the green histogram of Figure 1.3 that presents a large peak at the origin, showing a **strong agreement between** $\hat{\eta}_\alpha^{\text{plg}}$ and $\hat{\eta}_\alpha^\star$. The fact that only the green and blue histograms have non-zero values for the negative semi-axis confirms the tendency for smaller errors of $\hat{\eta}_\alpha^{\text{plg}}$ indicated in Table 1.3. The differences attenuate when the designs are larger, see Figure 1.4. The shape of the blue and red histograms for all designs sizes, shows that $\hat{\eta}_\alpha^{\text{emp}}$ may, for this function, **perform worse than the two other estimators** for designs of small size. Figure 1.4 shows that this difference becomes weaker for larger designs.
- ($\hat{\eta}_\alpha^{\text{plg}} \succeq \hat{\eta}_\alpha^\star \succeq \hat{\eta}_\alpha^{\text{emp}}$) For function Gramacy, the histograms reveal an **overall similar behaviour of all estimates** (all red, blue and green curves showing a major central peak), in particular of $\hat{\eta}_\alpha^\star$ and $\hat{\eta}_\alpha^{\text{plg}}$ (in green). This, together with the symmetry of the red and the blue histograms around the origin confirms that for this function **the other two estimates may sometimes be better** than $\hat{\eta}_\alpha^{\text{emp}}$. We verified that the two modes that can be identified in the blue (red) histogram, one for small positive (negative) values and the other for large (positive) negative values, do not correspond to the two values of α considered for this function, indicating rather that poor designs are produced by some design algorithms, leading to poor performance of the two model-based estimates. Figure 1.4 shows that for larger designs all estimates present similar errors (notice the difference in scale), and that no estimate seems to presents a systematically better behaviour.

- ($\hat{\eta}_\alpha^{\rm emp} \succ \hat{\eta}_\alpha^\star \succeq \hat{\eta}_\alpha^{\rm plg}$) The histograms for function Branin show that $\hat{\eta}_\alpha^{\rm emp}$ often produces **estimates with smaller errors**: the red (blue) histogram is zero for negative (positive) values. Note that $\hat{\eta}_\alpha^{\rm plg}$ occasionally produces (with low probability) slightly worse estimates than $\hat{\eta}_\alpha^\star$, as the lobe at small positive values of the green histogram shows, this difference remaining at large designs. We verified that the histograms for different values of α have a similar shape, with heavier queues for $\alpha = 0.02$, for which the gradient is larger over the corresponding level-set.
- ($\hat{\eta}_\alpha^{\rm emp} \succ \hat{\eta}_\alpha^\star \succeq \hat{\eta}_\alpha^{\rm plg}$)The histograms for function Goldprice indicate a lack of convergence of some design algorithms/estimates, revealed by the almost equal scale in Figures 1.3 and 1.4: all histograms basically retain the same shape as in 1.3 slightly more concentrated towards the origin. This important remark flags a failure of the designs to provide enough information for identifying the models on which these estimators are based, even for the largest designs considered, when $\hat{\eta}_\alpha^{\rm emp}$ is **always the best estimate**. The relative symmetry of the red and blue histograms together with the concentration of the green histogram around the origin indicates that the deviations of $\hat{\eta}_\alpha^{\rm plg}$ and $\hat{\eta}_\alpha^\star$ with respect to $\hat{\eta}_\alpha^{\rm emp}$ are similar. The positive skewness of the green histogram indicates that $\hat{\eta}_\alpha^{\rm plg}$ **nearly always performs worst** than $\hat{\eta}_\alpha^\star$, even for larger designs (and even worse with respect to $\hat{\eta}_\alpha^{\rm emp}$).

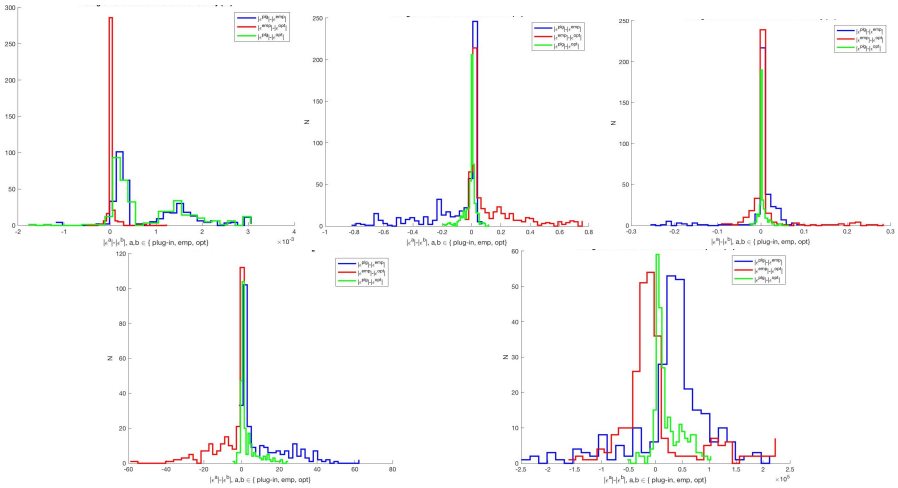

Fig. 1.3 Histograms of differences of absolute errors of estimators (by function, for all implementations of all criteria, for all designs sizes). Left to right, top to bottom: Ackely, F1, Gramacy, Branin and Goldprice.

In summary,

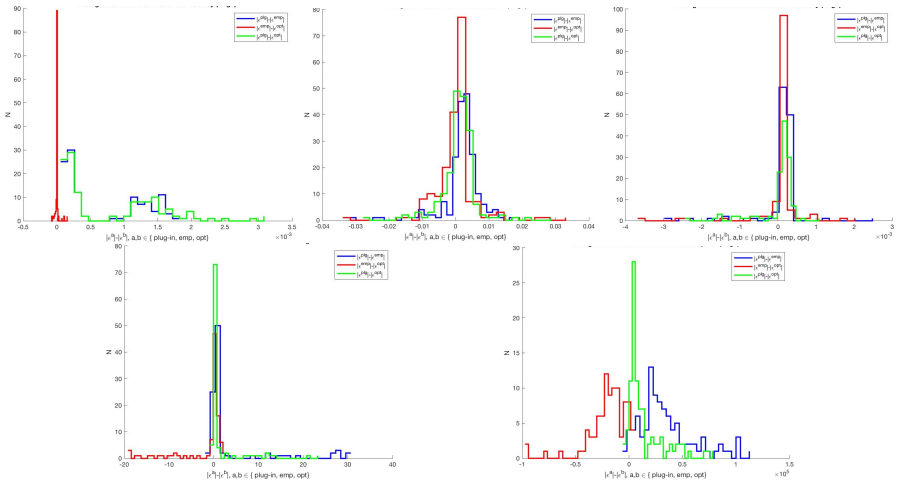

Fig. 1.4 Histograms of differences of absolute errors of estimators (by function, for all implementations of all criteria, four largest designs). Left to right, top to bottom: Ackley, F1, Gramacy, Branin and Goldprice.

- Unless – like it is the case for function F1 – the observed function matches the characteristics of the fitted Gaussian Process model, $\hat{\eta}_\alpha^{\text{emp}}$ is a robust and numerically efficient alternative to $\hat{\eta}_\alpha^{\text{plg}}$ and $\hat{\eta}_\alpha^\star$.
- The new estimate introduced in the paper, $\hat{\eta}_\alpha^{\text{plg}}$, offers for the majority of case-studies a performance comparable to the optimal estimate $\hat{\eta}_\alpha^\star$ at a much lower computational complexity, showing that full reliance on the Gaussian Process based uncertainty predictions does not necessarily lead to better percentile estimates.
- As it might be expected, our experiments confirm the impact of the gradient of f at the level-set defined by the estimated percentile on its estimation error.

1.6.3.2 Implementation

We address now the impact of the two implementation choices discussed before: *(a)* which estimate of η_α should be used in the design criteria J_n^{SUR}, J_n^{ID} and J_n^{iMC}? *(b)* does importance sampling, as proposed in section 1.5 lead to numerically stable implementations?

Figure 1.5 addresses the first question, showing performance profiles comparing the errors of solutions that differ only on the estimate ($\hat{\eta}_\alpha^{\text{plg}}$ in the red line, and $\hat{\eta}_\alpha^{\text{emp}}$ for the blue line) is used in the computation of the design criteria. The scores in the top plots are computed over the entire set of estimates $\mathcal{E}$, and the bottom only for the most robust estimator $\hat{\eta}_\alpha^{\text{emp}}$. The plots show that the designs found by the implementations that use $\hat{\eta}_\alpha^{\text{emp}}$ lead in

general to better designs for both $d = 1$ and $d = 2$, large and small designs, and rather independently of the estimator that produces the estimate.

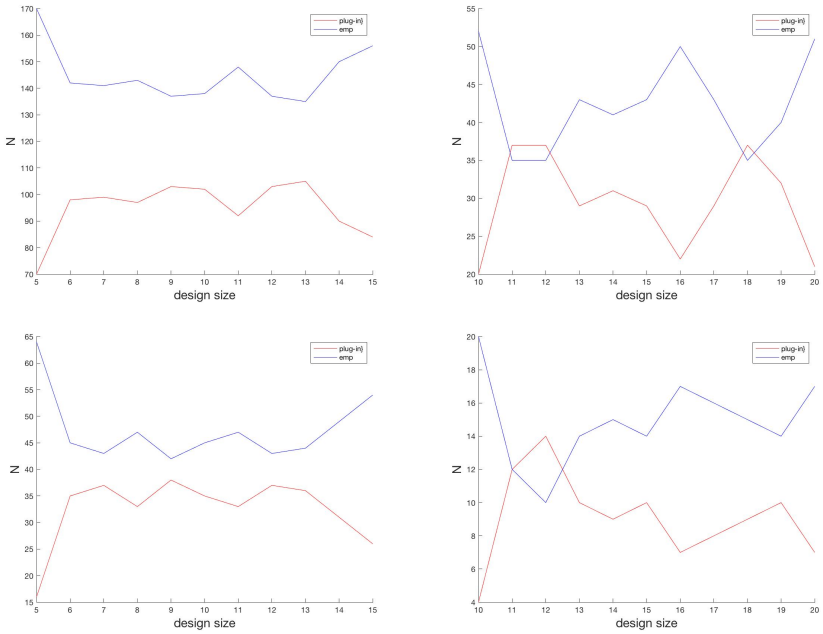

Fig. 1.5 Performance profile for estimate used in the computation of the criteria. Left: $d = 1$, right: $d = 2$. Top: all estimates in $\mathcal{E}$, bottom: only $\hat{\eta}_\alpha^{\text{emp}}$.

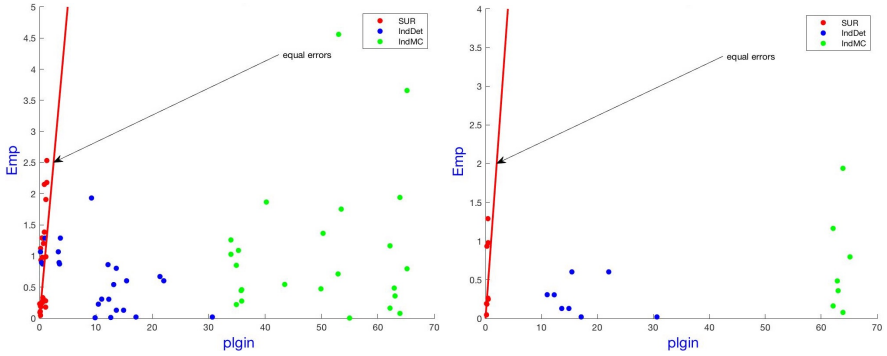

Fig. 1.6 Impact of the estimate used in the implementation of the criteria. Branin function, $\alpha = 0.02$. Left: all estimates, right: empirical estimate only. The largest four designs are considered.

Figure 1.6 illustrates this considering function Branin with $\alpha = 0.02$. The points $(\epsilon^{plg}(k), \epsilon^{emp}(k))$ in this plot enable the comparison of the errors $\epsilon^{plg}(k)$ of the estimates obtained with the largest four designs produced by implementations that use $\hat{\eta}_\alpha^{\text{emp}}$ against $\epsilon^{emp}(k)$, the errors of solutions that only differs in using an implementation based on $\hat{\eta}_\alpha^{\text{plg}}$. The red line represents $(\epsilon^{plg}(k) = \epsilon^{emp}(k))$. The color of the symbols codes the criterion used, as indicated in the legend of the figure. The left plot considers all three estimates, while the right plot shows only $\hat{\eta}_\alpha^{\text{emp}}$. We can see that except for J_{SUR} (red dots) use of $\hat{\eta}_\alpha^{\text{plg}}$ almost always leads to larger errors than use of $\hat{\eta}_\alpha^{\text{emp}}$: all blue and green points fall to the right of the red line, where $\epsilon^{plg}(k) > \epsilon^{emp}(k)$.

Figure 1.7 shows the (20 points) designs for this case-study generated using the two implementations of J_n^{iMC} that use importance sampling. In the left, the implementation using $\hat{\eta}_\alpha^{\text{plg}}$, in the right the one that uses $\hat{\eta}_\alpha^{\text{emp}}$ (see Figure 1.1 for the target region for this case-study). We see that the implementation that uses $\hat{\eta}_\alpha^{\text{plg}}$ placed the design points far from the level curves corresponding to the correct η_α. In this – rather extreme – case-study, the equation $\hat{\mu}_{f|\mathcal{X}_n}(x) = \hat{\eta}_\alpha^{\text{plg}}$ has no solution (explaining the absence of contour lines in the plot), clearly violating the small error assumption behind both J_n^{ID} and J_n^{iMC}. Use of the $\hat{\eta}_\alpha^{\text{emp}}$ in their implementation leads to design points closer to the regions contributing to the exceedance probability, as shown in the right plot.

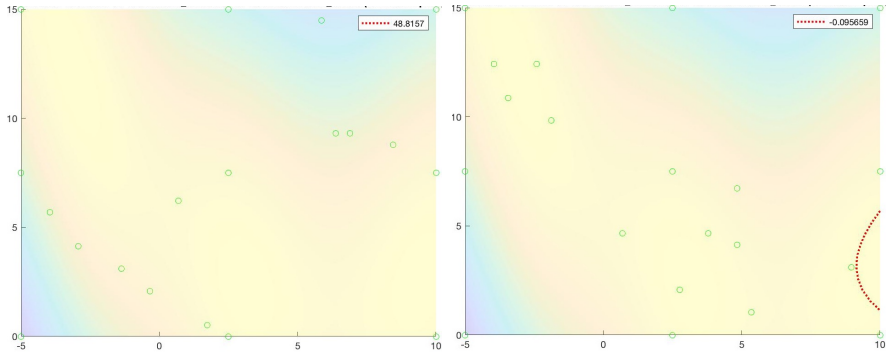

Fig. 1.7 Design build using J_n^{iMC} using $\hat{\eta}_\alpha^{\text{plg}}$ (left) and $\hat{\eta}_\alpha^{\text{emp}}$ (right) and importance sampling for the function Branin, with $\alpha = 0.02$.

We consider now the impact of the second implementation choice, concerning the use of importance sampling. We illustrate this problem considering one-dimensional functions only, for which, given the smaller uncertainty affecting the field posteriors, its impact may be the largest.

Figure 1.8 plots the evolution of the estimation errors of $\hat{\eta}_\alpha^{\text{emp}}$ (as designs are adaptively increased) for functions F1 (with $\alpha = 0.1$ and $\alpha = 0.32$) and Gramacy (with $\alpha = 0.1$) for designs produces by implementations that use $\hat{\eta}_\alpha^{\text{emp}}$. The legends detail the color/line-style/symbol codes used.

Leaving aside the cyan lines, which correspond to $J_n^\star$, comparison of lines with the same color (red, gree or blue) but with different line-styles, which differ only on the use (or not) of importance sampling, show a real impact of importance sampling for $J_n^{\rm ID}$ and $J_n^{\rm iMC}$, for which it leads to faster convergence to the ground truth value. The stochastic variability it induces in $J_n^{\rm SUR}$ often degrades performance presenting no clear advantage.

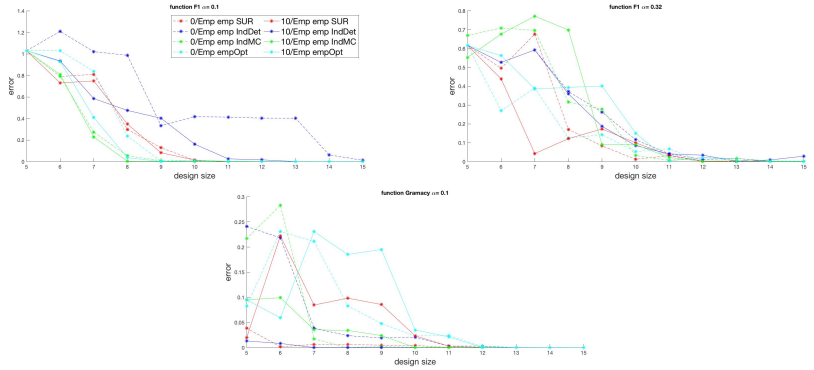

Fig. 1.8 Evolution of the estimation errors. Top: function F1/$\alpha = 0.1$ and $\alpha = 0.32$. Bottom Gramacy/$\alpha = 0.1$.

Figure 1.9 (right) shows the designs found by implementations of $J_n^{\rm ID}$ using $\hat{\eta}_\alpha^{\rm emp}$ for the top plot of Figure 1.8, with (left) and without (right) importance sampling (numbers indicate the order by which design points have been chosen): without importance sampling no design points are placed in the vicinity of the true level-set.

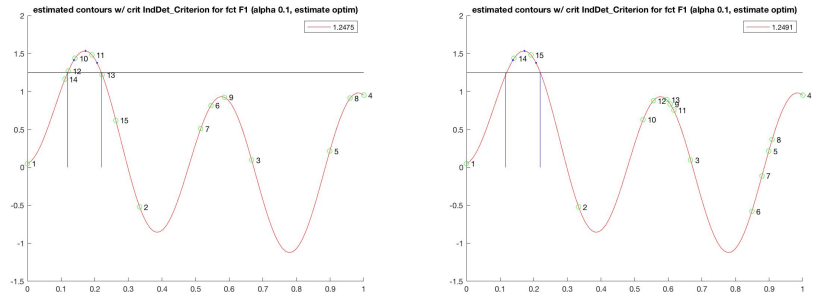

Fig. 1.9 Design produced by $J_n^{\rm ID}$ implemented with (left) and without (right) importance sampling and using $\hat{\eta}_\alpha^{\rm plg}$. Function F1, $\alpha = 0.1$.

In summary,

- As it might be anticipated, implementations based on $\hat{\eta}_\alpha^{\text{emp}}$ are more robust than those that use $\hat{\eta}_\alpha^{\text{plg}}$, in particular when α is close to 0 or 1.
- Importance sampling is able to attenuate numerical problems associated to J_n^{ID} and J_n^{iMC} with limited computational complexity. It is particularly important in situations of low posterior uncertainty. It presents no advantage for J_n^{SUR}.

1.6.3.3 Criteria

We finally address the relative merits of the different criteria studied. Since even our moderate choices of M_f and N_f lead to infeasibly large computation times for $d = 2$, we compare only the other three criteria: J_n^{SUR}, J_n^{ID} and J_n^{iMC}.

Figure 1.10 shows performance profiles of J_n^{SUR}, J_n^{ID} and J_n^{iMC} – for $d = 1$ (left) and $d = 2$ (right) – computed only from the errors of $\hat{\eta}_\alpha^{\text{emp}}$. The top row considers the two implementations of the criteria based on $\hat{\eta}_\alpha^{\text{emp}}$, while the bottom row retains only the implementation of each criterion that uses $\hat{\eta}_\alpha^{\text{emp}}$ and importance sampling. The overall evolutions of the relative performance of the criteria as design size grows are identical for both the top and the bottom rows, but differ largely between $d = 1$ and $d = 2$. For one-dimensional functions J_n^{SUR} and J_n^{ID} show a similar ability of producing the best designs, while for $d = 2$ the J_n^{SUR} leads to better final (larger) designs and J_n^{ID} provides the best mid-size designs. J_n^{iMC} is rarely the best method, irrespective of the value of d.

To assess how different the errors of the estimates obtained with the designs produced by distinct criteria can differ, Figures 1.11 and 1.12 present, for each function — respectively for all designs and for the four larger designs – empirical cumulative distribution functions of the differences of absolute errors of designs of the same size produced by the three criteria, considering only implementations that use $\hat{\eta}_\alpha^{\text{emp}}$. The black vertical line indicates the zero value, and concentration of the cumulative distribution functions around this line indicate that the corresponding estimates agree.

We can first notice that no criteria leads to systematically smaller errors than the other two for all case-studies.

When all designs are considered, see Figure 1.11, we can see that for $d = 1$ all criteria lead to similar performance, but $J_n^{\text{SUR}} \simeq J_n^{\text{ID}} \succ J_n^{\text{iMC}}$ for function Ackley, no clear classification can be established for function F1 (the different peaks of the function being discovered at distinct design sizes for the different criteria), and $J_n^{\text{ID}} \simeq J_n^{\text{iMC}} \succ J_n^{\text{SUR}}$ for function Gramacy. For $d = 2$ $J_n^{\text{ID}} \succeq J_n^{\text{iMC}} \succeq J_n^{\text{SUR}}$ for both functions. The behaviour for function Ackley when only the errors in the four largest designs are considered does not change, see Figure 1.11, and $Jsur \simeq J_n^{\text{iMC}} \succeq J_n^{\text{ID}}$ for F1, while all three criteria perform similarly for Gramacy, with J_n^{SUR} performing slightly better than J_n^{iMC}. For $d = 2$ $J_n^{\text{ID}} \succeq J_n^{\text{iMC}} \succeq J_n^{\text{SUR}}$ for Branin and $J_n^{\text{ID}} \succ J_n^{\text{iMC}} \succeq J_n^{\text{SUR}}$ for function Goldprice.

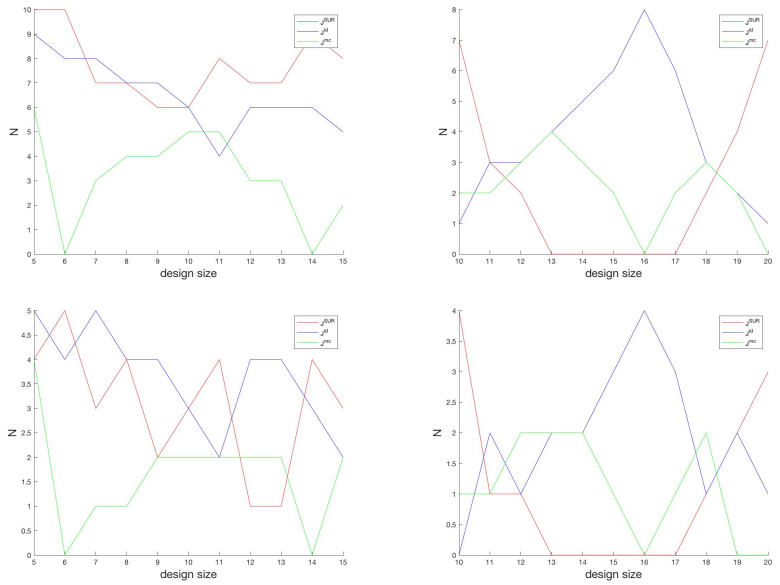

Fig. 1.10 Performance profile for J_n^{SUR}, J_n^{ID} and J_n^{iMC} criteria. Left: $d = 1$, right: $d = 2$. Top: $\hat{\eta}_\alpha^{\mathrm{emp}}$, implementations using $\hat{\eta}_\alpha^{\mathrm{emp}}$. Bottom: $\hat{\eta}_\alpha^{\mathrm{emp}}$, implementation importance sampling and $\hat{\eta}_\alpha^{\mathrm{emp}}$.

We conclude thus that even if there is no marked preference for the new criteria J_n^{ID} and J_n^{iMC} relative to J_n^{SUR}, they often lead to similar and improved performance, while being prone, for all functions, to (rare) larger errors than those of the simpler criterion J_n^{SUR}. The designs found with J_n^{ID} and J_n^{iMC} lead to similar error levels (for the same design size) except for for F1, for which our implementation of J_n^{iMC} tends to perform the worst.

Finally, we present in Figure 1.13 the designs found by the implementations of the three criteria that use $\hat{\eta}_\alpha^{\mathrm{emp}}$ together with importance sampling for one difficult case-study: the function Branin with $\alpha = 0.02$. The plots show a color code of the interpolated field and (in red) the inferred level-lines $\mu_{f|\mathcal{X}_n}(x) = \hat{\eta}_\alpha$. This figure confirms that, as we might expect, that although not necessarily leading to better percentile estimates, J_n^{SUR} is better able to correctly locate the three separate components of the level-line. Figure 1.14 shows the evolution of the absolute errors $of\hat{\eta}_\alpha^{\mathrm{emp}}$ as the design size is increased. In the top plot, which corresponds to the designs shown in Figure 1.13, we can see that the error of the estimate produced by either J_n^{ID} or J_n^{iMC} is lower (except for the final design point) demonstrating that the problems of estimation of excursion sets, probability of exceedance and percentile are indeed different problems. While J_n^{SUR} seems to outperform the other two criteria for the estimation of excursion sets, in many situations, as for the case-studies in this Figure, it leads to worse estimations of the percentile.

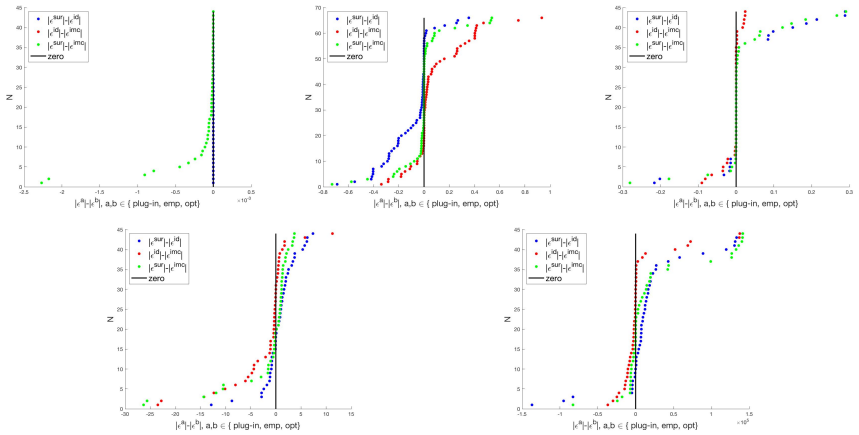

Fig. 1.11 Empirical cumulative distribution functions of differences of absolute errors of the criteria (all designs, implementations that use $\hat{\eta}_\alpha^{\mathrm{emp}}$). Left to right, top to bottom: Ackley, F1, Gramacy, Branin and Groldprice functions.

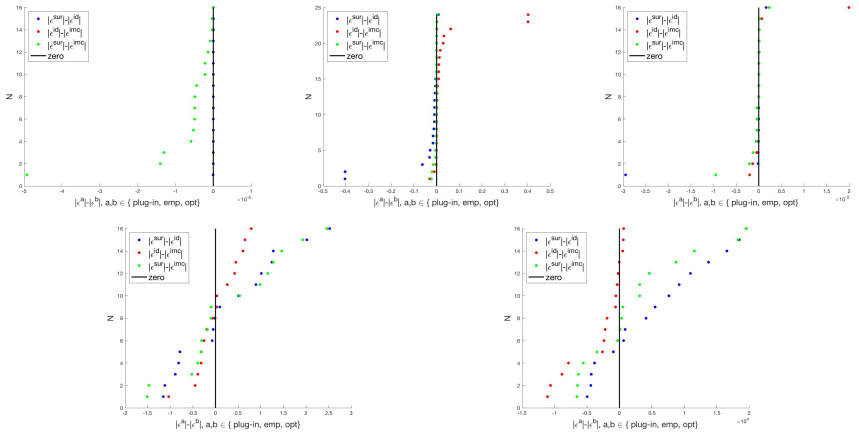

Fig. 1.12 Empirical cumulative distribution functions of differences of absolute errors of the criteria (four largest designs, implementations that use $\hat{\eta}_\alpha^{\mathrm{emp}}$). Left to right, top to bottom: Ackley, F1, Gramacy, Branin and Groldprice functions.

1.7 Conclusions

The paper proposes two new new adaptive design criteria for the estimation of η_α, the α-percentile of a function. Both criteria are modifications of criterion J_n^{SUR} [2], which quantifies the expected error in the dual problem of estimation of the exceedance probability α_η. Design construction proceeds, when using these methods, by alternating between the choice of a new point by optimising the criterion for a current estimate of the percentile η_α, and

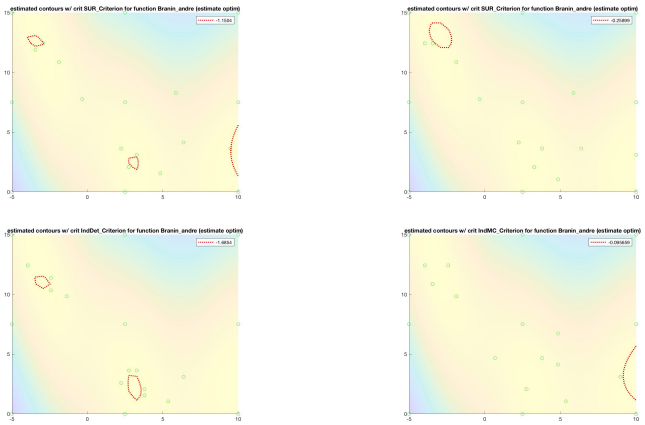

Fig. 1.13 Designs found for function Branin with $\alpha = 0.02$. Top: J_n^{SUR} with (left) and without (right) importance sampling.Bottom, J_n^{ID} (left) and J_n^{iMC} (right),implementations using $\hat{\eta}_\alpha^{\mathrm{emp}}$ and importance sampling.

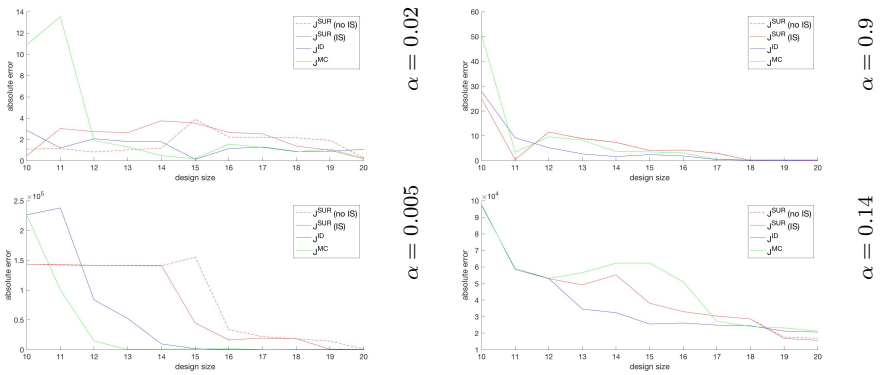

Fig. 1.14 Evolution of the absolute errors of $\hat{\eta}_\alpha^{\mathrm{emp}}$ during construction of the designs. Top: function Branin ($\alpha = 0.3$ and $\alpha = 0.8$). Bottom: function Goldprice ($\alpha = 0.005$ and $\alpha = 0.14$).

update of the percentile estimate, additionally integrating the observation at the added point. The case-studies considered gives evidence that this alternating procedure converges in a few design points to the true value of the estimated percentile.

At the root of the derivation of the new criteria proposed is a the (novel) plug-in estimator $\hat{\eta}_\alpha^{\mathrm{plg}}$ proposed, which is based on the duality between η_α and α_η, initially expected to achieve a compromise between the simplicity of the empirical estimate of the percentile $\hat{\eta}_\alpha^{\mathrm{emp}}$ and the complexity of the full optimal Bayesian estimate $\hat{\eta}_\alpha^\star$. The numerical experiments presented reveal, on a number of distinct case-studies, that $\hat{\eta}_\alpha^{\mathrm{plg}}$ has performance similar to our

implementation of optimal Bayes estimator, but no better error performance – in the set of deterministic functions used – than the numerically simple and robust empirical estimate $\hat{\eta}_\alpha^{\mathrm{emp}}$. We are convinced that this "negative result" is interesting in itself, showing in a number of concrete examples that application of model-based estimators to very small designs as the ones considered here must be done with care, being prone to performance degradation when the function does not match the model assumptions.

The two new criteria for percentile estimation derived in the manuscript, J_n^{ID} and J_n^{iMC}, both implement a multiplicative correction of criterion J_n^{SUR}. The computation of these multiplicative factors is computationally demanding, requiring a delicate integration over $x \sim p_{\mathcal{A}}$ of the posterior density at the current percentile estimate $\hat{\eta}_\alpha$. We observed the influence of the choice of the estimate – either the new plug-in estimate $\hat{\eta}_\alpha^{\mathrm{plg}}$ or the classic empirical estimate $\hat{\eta}_\alpha^{\mathrm{emp}}$ – used in the evaluation of the design criterion in the convergence of this alternating process. Our numerical experiments show that the empirical estimate $\hat{\eta}_\alpha^{\mathrm{emp}}$, which always defines a non-empty level-set in the predicted field, should be preferred to $\hat{\eta}_\alpha^{\mathrm{plg}}$ to ensure convergence in difficult problems. The computation of both multiplicative correction factors, of J_n^{ID} and J_n^{iMC}, is prone to numerical instabilities, in particular when the uncertainty of the predicted field is small. We verified in a number of concrete examples that importance sampling may attenuate the problem, at the cost of increased computational complexity.

The performance of designs build by alternating estimation of η_α and application of criteria J_n^{SUR}, J_n^{ID} and J_n^{iMC} is compared on a set of 11 case-studies, derived from 3 one-dimensional and 2 two-dimensional functions. In all cases the estimates converged in a few design points to the correct percentiles. Our experiments show no clear evidence that the more complex new criteria J_n^{ID} and J_n^{iMC} outperforms the direct application of J_n^{SUR}, making this latter criterion a competitive choice for adaptive estimation of percentile estimation: its numerical complexity is smaller, it is less prone to numerical problems, and leads to designs with nearly identical performance. In fact, its performance comes close to – or even improves on – the performance of practical implementations of the ideal model-based criterion $J_n^\star$. Figure 1.8 illustrates this. The four lines cyan lines in each plot of this figure show the evolution of the (absolute) error during four statistically independent definitions of designs of size 15 using J_n^{SUR}. To keep computation time within acceptable limits, the implementation of $J_n^\star$ uses only $M_f = 20$ realisations from the posterior $p_{f|\mathcal{X}_n}$ and draws only $N_f = 20$ independent samples at each evaluation of $J_n^\star$. The significant variability of the generated designs, which is apparent from the fluctuation of the corresponding estimates, is at least comparable to the error level of the other criteria (similar results were obtained for the other case-studies), showing that the values of M_f and N_f used are not sufficiently large to guarantee that $J_n^\star$ leads to smaller error than the other tested criteria. Given the much heavier computational complexity

of $J_n^\star$, we conclude that, in practice, $J_n^\star$ is an unattractive alternative with respect to any of other three criteria.

To the best of our knowledge, the possibility of using J_n^{SUR} for adaptive percentile estimation had not been demonstrated in the past. Given the much lower numerical complexity of J_n^{SUR} when compared to the design algorithms that have been proposed in the literature for the problem of percentile estimation, e.g. in [5], this finding provides a practical alternative under limited computational budgets.

Acknowledgements This work benefited from the support of the project INDEX (INcremental Design of EXperiments) ANR-18-CE91-0007 of the French National Research Agency (ANR).

Appendix 1

Posterior mean and variance of f under the Gaussian procees assumption

For a Gaussian process with mean function $\mu(x)$ and covariance function $C(x, x')$, its mean and variance conditioned on n observations at point x, $\mu_{f|\mathcal{X}_n}(x)$ and $r_{f|\mathcal{X}_n}(x)$ respectively, are given by the following expressions [7]:

$$\mu_{f|\mathcal{X}_n}(x) = \mu(x) + r(x, X_n)^{\mathrm{T}} R_{X_n}^{-1}(f(X_N) - \mu(X_N)) \ , \tag{1.24}$$

$$r_{f|\mathcal{X}_n}(x) = C(x, x) - r(x, X_n)^{\mathrm{T}} R_{X_n}^{-1} r(x, X_n) \ , \tag{1.25}$$

where $r(x, X_n) = [C(x, x_1) \cdots C(x, x_n)]^{\mathrm{T}}$, R_{X_n} is the covariance matrix of the n observations, $f(X_n) = [f(x_1) \cdots f(x_n)]^{\mathrm{T}}$ and $\mu(X_n) = [\mu(x_1) \cdots \mu(x_n)]^{\mathrm{T}}$.

SUR design criteria for exceedance probability estimation

A SUR criterion introduced in [2] for sampling x_{n+1}, when we aim at estimating a probability of exceedance α_η is the conditional mean square error of the optimal estimator:

$$J_n^\star(x; \eta) = \mathbb{E}_{f|\mathcal{X}_n}\left[(\alpha_\eta - \hat{\alpha}_\eta^\star(\mathcal{X}_{n+1}))^2\right] \ . \tag{1.26}$$

Notice that $\hat{\alpha}_\eta^\star(\mathcal{X}_{n+1})$ given $\mathcal{X}_{n+1}$ is a random variable depending on $f(x)$, therefore the above expectation is an expectation with respect to (w.r.t.) $f(x)$ conditioned on $\mathcal{X}_n$.

By developing the expectation, it can be shown [3] that (1.26) can be rewritten as follows

$$J_n^\star(x;\eta) = \gamma_n - \mathbb{E}_{z_1\sim\mathcal{A}}\{\mathbb{E}_{z_2\sim\mathcal{A}}\{\mathbb{E}_{f|\mathcal{X}_n}\left[\mathbb{P}_{f|\mathcal{X}_{n+1}}(f(z_1)\geq\eta)\mathbb{P}_{f|\mathcal{X}_{n+1}}(f(z_2)\geq\eta)\right]\}\}\ , \tag{1.27}$$

where γ_n does not depend on $x = x_{n+1}$. Therefore, minimizing $J_n^\star(x;\eta)$ is equivalent to the maximization of the double expectation of

$$\mathbb{E}_{f|\mathcal{X}_n}\left[\mathbb{P}_{f|\mathcal{X}_{n+1}}(f(z_1)\geq\eta)\mathbb{P}_{f|\mathcal{X}_{n+1}}(f(z_2)\geq\eta)\right]\ .$$

In [3], it was shown that the conditional expectation above can be written analytically as a function of the bivariate normal cumulative distribution. As a consequence, evaluation of $J_n^\star(x;\eta)$ can be carried out without requiring computationally expensive conditional sampling of the trajectories of f.

By applying the generalized Minkowski inequality and the Cauchy-Schwarz inequality, it can also be shown [2] that $J_n^\star(x;\eta)$ can be upper-bounded as follows:

$$J_n^\star(x;\eta) \leq J_n^{\mathrm{SUR}}(x;\eta) \stackrel{\triangle}{=} \mathbb{E}_{z\sim\mathcal{A}}\{\mathbb{E}_{f|\mathcal{X}_n}[\mathbb{P}_{f|\mathcal{X}_{n+1}}(f(z)\geq\eta)(1-\mathbb{P}_{f|\mathcal{X}_{n+1}}(f(z)\geq\eta))]\}\ . \tag{1.28}$$

In a similar way as for $J_n^\star(x;\eta)$, the inner conditional expectation can be evaluated analytically as a function of bivariate normal cumulative distribution. For details on the analytical expressions of the conditional expectations in $J_n^\star(x;\eta)$ and $J_n^{\mathrm{SUR}}(x;\eta)$ see [3].

Its important to notice that to approximate $J_n^{\mathrm{SUR}}(x;\eta)$ with Monte-Carlo, a set of i.i.d. draws from a single random variable $z\sim\mathcal{A}$ is required, while for $J_n^\star(x;\eta)$ sampling from the tuple (z_1,z_2) is required. Thus, for the same approximation accuracy a much smaller number of samples must be required to evaluate $J_n^{\mathrm{SUR}}(x;\eta)$. This leads to a much smaller computational complexity of approximation of $J_n^{\mathrm{SUR}}(x;\eta)$, when compared to the approximation of $J_n^\star(x;\eta)$.

Appendix 2

We present below the analytical expressions of the functions used in the study, see [9] and [10].

$$\text{Ackley}(x) = a + \exp(1) - a\exp\left(-b\sqrt{\frac{x^2}{d}}\right) - \exp\left(\frac{1}{d}\cos(2\pi x)\right), x \in [0,1],$$
$$a = 20, b = 0.2 \text{ and } d = 4.$$
$$\text{F1}(x) = 2\left(x - 0.75\right)^2 + \sin\left(5\pi x - 0.4\pi\right) - 0.125, \qquad x \in [0,1] \ .$$
$$\text{Gramacy}(x) = \left(a\,x - 2\right)\exp\left(-\left(a\,x-2\right)^2\right), \qquad x \in [0,1] \ \text{, with } a = 8.$$
$$\text{Branin}(x_1, x_2) = a\left(x_2 - bx_1^2 + cx_1 - r\right)^2 + s(1-t)\cos(x_1) + s \ ,$$
$$x_1 \in [-5, 10], x_2 \in [0, 15],$$
$$a = 1, b = 5.1/(4\pi^2), = 5/\pi, = 10, s = 10, \ t = 1/(8\pi).$$
$$\text{Goldprice}(x) = \left[1 + (x_1 + x_2 + 1)^2\left(19 - 14x_1 + 3x_1^2 - 14x_2 + 6x_1x_2 + 3x_2^2\right)\right]$$
$$\times\left(30 + (2x_1 - 3x_2)^2\left(18 - 32x_1 + 12x_1^2 + 48x_2 - 36x_1x_2 + 27x_2^2\right)\right)$$
$$x_1 \in [-2, 2], x_2 \in [-2, 2] \ .$$

References

1. A. Arnaud, J. Bect, M. Couplet, A. Pasanisi, and E. Vazquez. Évaluation d'un risque d'inondation fluviale par planification séquentielle d'expériences. In *42èmes Journées de Statistique*, 2010.
2. J. Bect, D. Ginsbourger, L. Li, V. Picheny, and E. Vazquez. Sequential design of computer experiments for the estimation of a probability of failure. *Statistics and Computing*, 22(3):773–793, 2012.
3. C. Chevalier, J. Bect, D. Ginsbourger, E. Vazquez, V. Picheny, and Y. Richet. Fast parallel kriging-based stepwise uncertainty reduction with application to the identification of an excursion set. *Technometrics*, 56(4):455–465, 2014.
4. M. Jala, C. Lévy-Leduc, É. Moulines, E. Conil, and J. Wiart. Sequential design of computer experiments for the assessment of fetus exposure to electromagnetic fields. *Technometrics*, 58(1):30–42, 2016.
5. T Labopin-Richard and V Picheny. Sequential design of experiments for estimating percentiles of black-box functions. *Statistica Sinica*, 28:853–877, 2018.
6. J. Oakley. Estimating percentiles of uncertain computer code outputs. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53(1):83–93, 2004.
7. C. E. Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.
8. STK a Small (Matlab/Octave) Toolbox for Kriging. `https://sourceforge.net/projects/kriging/`.
9. Derek Bingham, Virtual Library of Simulation Experiments: Test functions and Datasets. `https://www.sfu.ca/~ssurjano/optimization.html`.
10. J. Andre, P. Siarry, and T. Dognon. An improvement of the standard genetic algorithm fighting premature convergence in continuous optimization *Advances in Engineering Software*, 32(1):49–60, 2000.
11. V. Beiranvand, W. Hare and Y. Lucet. Best Practices for Comparing Optimization Algorithms *Optimization and Engineering*, June 2017 `https://arxiv.org/pdf/1709.08242.pdf`