Optimization of Noisy Computer Experiments with Tunable Precision

Victor Picheny, David Ginsbourger, Yann Richet, Grégory Caplin

Optimization of Noisy Computer Experiments with Tunable Precision

Victor Picheny

Ecole Centrale de Paris, Chatenay-Malabry, France

email: victor.picheny@ecp.fr

David Ginsbourger

University of Bern, Switzerland

email: david.ginbourger@stat.unibe.ch

Yann Richet

Institut de Radioprotection et de Sûreté Nucléaire, Fontenay-aux-Roses, France

email: yann.richet@irsn.fr

Gregory Caplin

Institut de Radioprotection et de Sûreté Nucléaire, Fontenay-aux-Roses, France

email: gregory.caplin@irsn.fr

August 15, 2011

Abstract

This article addresses the issue of kriging-based optimization of stochastic simulators. Many of these simulators depend on factors that tune the level of precision of the response, the gain in accuracy being at a price of computational time. The contribution of this work is two-fold: firstly, we propose a quantile-based criterion for the sequential choice of experiments, in the fashion of the classical Expected Improvement criterion, which allows an elegant treatment of heterogeneous response precisions. Secondly, we present a procedure to allocate on-line the computational time given to each measurement, allowing a better distribution of the computational effort and increased efficiency. Finally, the optimization method is applied to an original application in nuclear criticality safety.

Keywords: Noisy optimization, Kriging, Tunable fidelity.

1. INTRODUCTION

Using metamodels for facilitating optimization and statistical analysis of computationally expensive simulators has become commonplace. In particular, the kriging-based EGO algorithm (Jones, Schonlau and Welch 1998) and its underlying expected improvement (EI) criterion have been recognized as efficient tools for deterministic black-box optimization.

The way a simulator response follows the function of interest is called *fidelity*. Oftentimes, a large range of response fidelities is available by tuning factors that control the complexity of numerical methods. For instance, the precision of a finite element analysis can be controlled by the discretization technique or the solver convergence. When the response stems from Monte Carlo methods (which is often referred to as *stochastic simulators*), accuracy (measured by response variance) is proportional to sample size.

Such simulators are often called *noisy simulators*, since they return approximate solutions that depart from the exact value by an error term that can be considered as a random quantity. Optimization in this context raises critical issues. Having noise in the responses requires a proper adaptation of criteria and algorithms. Furthermore, for each simulation run, the user has to set a trade-off between computational cost and response precision. This additional degree of freedom may greatly improve the efficiency of the optimization, but requires appropriate tools to choose this trade-off and the ability to work with heterogeneous precisions.

Using metamodels for noisy optimization has been already addressed by several authors. In particular, Huang, Allen, Notz and Zeng (2006) and Forrester, Keane and Bressloff (2006) proposed kriging-based strategies for sequential optimization of uniformly noisy functions. However, little work can be found in the case of heterogeneous noise. Most approaches combining optimization and variable precision are found in the *multifidelity* framework (Alexandrov, Lewis, Gumbert, Green and Newman 2000; Gano, Renaud, Martin and Simpson 2006), but consider only two fidelity levels, the low-fidelity model being then used as a helping tool to choose the high-fidelity evaluations. In a more integrated approach, Huang, Allen, Notz and Miller (2006) proposed a criterion for hierarchical kriging models with finitely many levels of fidelity, that chooses at the same time the observation point and the fidelity.

This article proposes two contributions to this framework. First, we define an extension of EI based on quantiles that enables an elegant treatment of both continuous or discrete fidelities. The proposed criterion not only depends on the noise variances from the past, but also on the fidelity of the new candidate measurement. Hence, this criterion allows to choose both an input space point and a fidelity level at each iteration. Second, we study a procedure taking advantage of this additional degree of freedom. Once an input space point has been selected, computation time is invested on it until a stopping criterion is met. One of the advantages of such procedure is that it prevents from allocating too much time to poor designs, and allows spending more credit on the best ones.

In the next section, we first define the "noisy" framework we are considering and present briefly the kriging model. Section 3 describes the classical kriging-based optimization procedure, and its limitation

with noisy functions. Then, the quantile-based EI criterion is proposed, followed by the on-line allocation procedure, which is compared to existing kriging-based methods. Finally, an original application in nuclear criticality safety is implemented in the Promethee workbench and applied to the Monte Carlo criticality simulator MORET5 (Fernex, Heulers, Jacquet, Miss and Richet 2005).

2. NOTATIONS AND CONCEPTS

2.1 The noisy optimization problem

We consider a single objective, unconstrained optimization problem over a compact set D. The deterministic objective function $y : \mathbf{x} \in D \subset \mathbb{R}^d \longrightarrow y(\mathbf{x}) \in \mathbb{R}$ is here observed in noise. For a measurement at some $\mathbf{x} \in D$, the user does not have access to the exact $y(\mathbf{x})$, but to an approximate response $y(\mathbf{x}) + \epsilon$. ϵ is assumed to be one realization of a "noise" random variable ε, whose probability distribution may depend on $\mathbf{x}$ and other variables, and which realizations might differ for different measurements of y at the same $\mathbf{x}$. So instead of referring to the measurements of y in terms of $\mathbf{x}$'s, we will denote by $\widetilde{y_i} = y(\mathbf{x}^i) + \epsilon_i$ the noisy measurements, where the $\mathbf{x}^i$'s are not necessarily all distinct.

In the rest of this article, we make the assumption that the observation noises are normally distributed, centered and independent from one run to each other:

$$\varepsilon_i \sim \mathcal{N}(0, \tau_i^2) \text{ independently.} \tag{1}$$

2.2 Noise in computer experiments

In classical experiments, noise usually accounts for a large number of uncontrolled variables (variations of the experimental setup, measurement precision, etc.). In computer experiments, noise can have many sources, including modeling and discretization error, incomplete convergence, and finite sample size for Monte-Carlo methods, see for instance Forrester, Keane and Bressloff (2006) for a detailed discussion. Also, in the framework of robust design, a noisy objective function is often defined from a deterministic simulator, considering that some of the inputs are non-controllable and modeled as random variables. The objective function is then a statistic of the output (typically, a mean or quantile), evaluated using Monte-Carlo methods, hence with noise.

The nature of the noise depends on the associated simulator. When classical Monte-Carlo simulations are involved in the output evaluation, error is independent from one run to each other, even for measurements with the same input variables. Such simulators are often referred to as *stochastic*, and are the main target for the method presented here.

In the framework of *multi-fidelity evaluations*, error is due to a simplification of the physics equations, geometry, or discretization (e.g. meshing in Finite Elements models). In that case, errors are likely to be strongly correlated, especially for simulations with similar fidelities, and repeated experiments provide the

same observations. This situation has been addressed in the literature (see Kennedy and O'Hagan (2000), Santner, Williams and Notz (2003) and Qian and Wu (2008) for modeling, Alexandrov et al. (2000) and Huang, Allen, Notz and Zeng (2006) for optimization) and will not be considered here, although many of the concepts presented here may apply with a proper adaptation of the kriging model.

When error is due to incomplete convergence, errors are also likely to be correlated. In the problem considered in Forrester, Bressloff and Keane (2006), simulations across the design space tend to converge in unison (errors are almost equal for two measurements with the same convergence level), which makes the partial convergence equivalent to a multi-fidelity problem. However, when the output convergence behavior varies substantially across the design space, the hypothesis of independence of the error between runs may become reasonable, especially if experiments are well spread in the design space and different convergence levels are used.

2.3 Experiments with tunable precision

As mentioned in the introduction, the precision of many simulators can be tuned by the user, for instance by changing the number of solver steps for incomplete convergence or the sample size for Monte-Carlo methods. Of course, computational time increases with precision.

Hence, we consider that for every measurement i ($1 \leq i \leq n$), the noise variance $\tau_i^2 = \tau(t_i)$ is a monotonically decreasing function of computation time t_i, with:

$$\tau^2 : t \in [0, +\infty[\longrightarrow \tau^2(t) \in [0, +\infty[\tag{2}$$

Then, the actual (inaccessible) objective function y is the response given by the simulator with an infinite computational time allocated at every $\mathbf{x} \in D$. The difference between the simulation and the actual phenomenon is not considered here.

A perhaps "canonical" example of tunable precision is when the response considered is obtained by averaging an arbitrary number b_i of independent drawings (or repeated experiments):

$$\widetilde{Y}_i = \frac{1}{b_i} \sum_{j=1}^{b_i} y(\mathbf{x}_i) + \varepsilon_{i,j}, \tag{3}$$

when $\varepsilon_{i,j} \sim \mathcal{N}(0, \nu^2)$. We have then $\widetilde{Y}_i \sim \mathcal{N}\left(y(\mathbf{x}_i), \frac{\nu^2}{b_i}\right)$, so $\tau^2(t) = \rho\nu^2/t$, ρ being the time needed for a single drawing. The value of b_i chosen by the user tunes the precision of $\widetilde{Y}_i$.

In this work, we make two strong assumptions: (a) the computation time, and hence the error variance, is controllable, and (b) the function $\tau(t)$ is accurately known. Although some stochastic simulators, such as the one described in section 7.3, directly provide an accurate estimate of the output uncertainty, in most real applications a learning study is necessary, typically assuming a (simple) parametric form for the variance. In

the case of Monte-Carlo simulators (or repeated experiments) and assuming small variations of the output across the design space, we have $\tau^2(t) = C/t$, where C is an unknown constant which can be estimated when building the kriging model, as described in section 2.4.

Finally, for simulators relying on Monte Carlo or on iterative solvers, the response corresponding to a given precision is not obtained directly but more as a limit of intermediate responses of lower precisions. For each measurement, the noisy response $\widetilde{y}_i$ is thus obtained as last term of a sequence of measurements $\widetilde{y}_i[1], \ldots, \widetilde{y}_i[b_i]$ with decreasing noise variances, $\tau_i^2[1] > \ldots > \tau_i^2[b_i]$, where $b_i \in \mathbb{N}$ is the number of calculation steps at the i^{th} measurement. Furthermore, each step is assumed here to correspond to one elementary computation time $t_e \in]0, +\infty[$, so that $\forall j \in \{1, \ldots, b_i\},\ \tau_i^2[j] = \tau^2(j \times t_e)$.

Figure 1 represents two examples of response convergence. First, the convergence of the output of the stochastic simulator of section 7.3 is drawn for its nominal design values. Here, the variance is known accurately, and depicted by the 95% confidence interval. The curve *ytilde* represents the sequence $\widetilde{y}[j], j = 1 \ldots 100$. The second figure is taken from Forrester, Bressloff and Keane (2006) and represents the convergence of an objective function (namely the L/D ratio) calculated using an Euler simulation of an aerofoil, as a function of the number of solver steps. The response oscillates around its final value with decreasing amplitude. Here, error variance is not available directly and requires to infer a parametric model for τ based on a couple of trial responses such as this one.

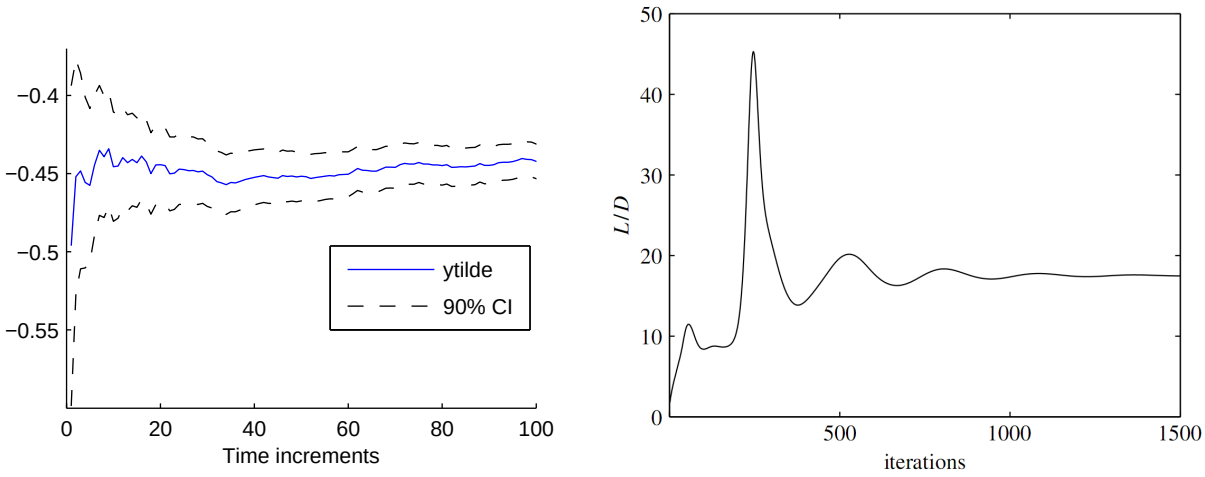

Figure 1: Examples of tunable precision responses. Left: convergence of the output of the MORET simulator for its nominal design values; right: partially converged response of a CFD code.

2.4 The Kriging metamodel

Kriging is a functional approximation method originally coming from geosciences (Krige 1951), and having been popularized in machine learning (Gaussian Process paradigm, see e.g. Rasmussen and Williams (2006)) and in numerous application fields. In the noiseless framework, Kriging simultaneously provides an interpolator of the partially observed function y, *the Kriging mean predictor* $m(.)$, and a measure of prediction

uncertainty at every $\mathbf{x}$, the *Kriging variance* $s^2(.)$. The basic idea is to see y as one realization of a square-integrable real-valued random process indexed by D, and to make optimal linear predictions of $Y(\mathbf{x})$ given the Y values at the already evaluated input points $\mathbf{X}^n := \{\mathbf{x}^i, 1 \leq i \leq n\}$. Of course, this prediction depends on the two first moments of the process Y, which are generally assumed to be known up to some coefficients. Here we assume that Y has an unknown constant trend $\mu \in \mathbb{R}$, and a stationary covariance kernel k, i.e. of the form $k : (\mathbf{x}, \mathbf{x}') \in D^2 \longrightarrow k(\mathbf{x}, \mathbf{x}') = \sigma^2 r(\mathbf{x} - \mathbf{x}'; \psi)$ for some admissible correlation function r with parameters ψ. This is the framework of *Ordinary Kriging* (OK) (Matheron 1969). Additionally, assuming further that $Y|\mu$ is a Gaussian Process (GP) and that μ is independent of Y and follows an improper uniform distribution over $\mathbb{R}$ leads to the convenient result that OK amounts to conditioning Y on the measurements, i.e. ensuring that $m(.)$ and $s^2(.)$ coincide respectively with the conditional mean and variance functions. We stick here to this set of assumptions, in order to get explicit (Gaussian) conditional distributions for $Y(\mathbf{x})$ knowing the observations, and to be in position to use generalizations of this to the heterogeneously noisy case.

Let us indeed come back to our noisy observations $\widetilde{y_i} = y(\mathbf{x}^i) + \epsilon_i$ $(1 \leq i \leq n)$. If we suppose that y is a realization of a GP following the OK assumptions above, the $\widetilde{y_i}$'s can now be seen as realizations of the random variables $\widetilde{Y_i} := Y(\mathbf{x}^i) + \varepsilon_i$, so that Kriging amounts to conditioning Y on the heterogeneously noisy observations $\widetilde{Y_i}$ $(1 \leq i \leq n)$. As shown earlier in Ginsbourger, Picheny, Roustant and Richet (2008), provided that the process Y and the Gaussian measurement errors ε_i are stochastically independent, the process Y is still Gaussian conditionally on the noisy observations $\widetilde{Y_i}$ $(1 \leq i \leq n)$, and its conditional mean and variance functions are given by the following slightly modified OK equations:

$$m_n(\mathbf{x}) = \mathbb{E}[Y(\mathbf{x})|\widetilde{A_n}] = \widehat{\mu}_n + \mathbf{k}_n(\mathbf{x})^T (K_n + \Delta_n)^{-1} (\widetilde{\mathbf{y}}^n - \widehat{\mu}_n \mathbf{1}_n), \quad (4)$$

$$s_n^2(\mathbf{x}) = \mathrm{Var}[Y(\mathbf{x})|\widetilde{A_n}] = \sigma^2 - \mathbf{k}_n(\mathbf{x})^T (K_n + \Delta_n)^{-1} \mathbf{k}_n(\mathbf{x}) + \frac{\left(1 - \mathbf{1}_n^T (K_n + \Delta_n)^{-1} \mathbf{k}_n(\mathbf{x})\right)^2}{\mathbf{1}_n^T (K_n + \Delta_n)^{-1} \mathbf{1}_n}, \quad (5)$$

where | means "conditional on", $\widetilde{\mathbf{y}}^n = (\widetilde{y_1}, \ldots, \widetilde{y_n})^T$, $\widetilde{A_n}$ is the event $\left\{Y(\mathbf{x}^i) + \varepsilon_i = \widetilde{y_i}, 1 \leq i \leq n\right\}$, $K_n = \left(k(\mathbf{x}^i, \mathbf{x}^j)\right)_{1 \leq i,j \leq n}$, $\mathbf{k}_n(\mathbf{x}) = (k(\mathbf{x}, \mathbf{x}^1), \ldots, k(\mathbf{x}, \mathbf{x}^n))^T$, Δ_n is a diagonal matrix of diagonal terms $\tau_1^2 \ldots \tau_n^2$, $\mathbf{1}_n$ is a $n \times 1$ vector of ones, and $\widehat{\mu}_n = \mathbf{1}_n^T (K_n + \Delta_n)^{-1} \widetilde{\mathbf{y}}^n / \mathbf{1}_n^T (K_n + \Delta_n)^{-1} \mathbf{1}_n$ is the best linear unbiased estimate of μ. $m(.)$ and $s^2(.)$ are indexed by n in order to bring to light the dependence on the design of experiments, and to prepare the ground for the algorithmic developments needing sequential Kriging updates.

The only difference compared to OK equations is the replacement of K_n by $K_n + \Delta_n$ at every occurrence. Specific properties of this generalization of OK include that $m_n(.)$ is not interpolating noisy measurements, that $s_n^2(.)$ does not vanish at that points and is globally inflated compared to the noiseless case. Note that although $s_n^2(.)$ now depends on both the design $\mathbf{X}^n$ and the noise variances $\boldsymbol{\tau}^2 := \{\tau_1^2, \ldots, \tau_n^2\}$, it still does not depend on the observations.

Such model resembles to the *kriging with (heterogeneous) nugget effect* of the geostatistic literature

(Matheron 1969), with the notable difference that the error variance term does not appear in the covariance vector $\mathbf{k}_n(\mathbf{x})$, which would make the model interpolant and thus not suited for repeated measurements.

Figure 2 shows an example of kriging based on noisy observations with heterogeneous noise. For small observation noise, the model is almost interpolating the data (e.g. at $x = 0.5$), while for large noise the confidence interval remains large and the best predictor can be far from the observation.

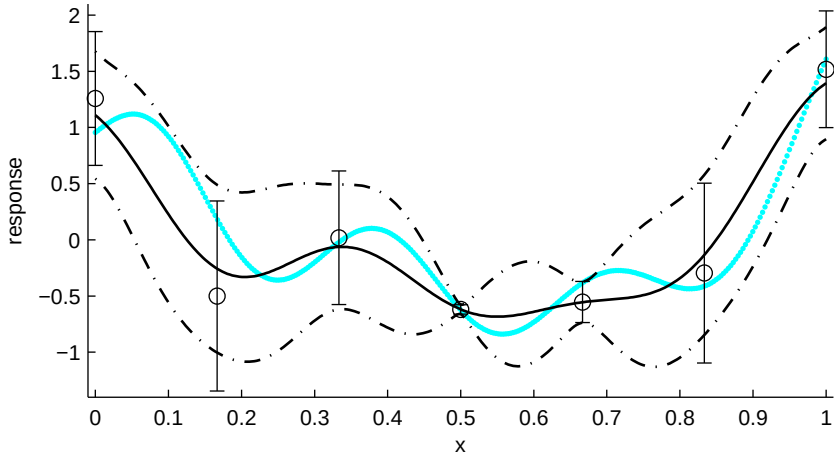

Figure 2: Actual function (bold gray), Kriging mean (bold black) and 90% confidence intervals (mixed line); the circles are the observation values $\tilde{y}_i$, the bars show the noise amplitude ($\pm 2 \times \tau_i$).

As for a classical kriging model, the covariance parameters σ^2 and ψ need to be estimated. Several alternatives exist, including variogram techniques, cross-validation and maximum likelihood. In the applications covered here, the method used is essentially maximum likelihood. In cases where the noise variances are considered as known, σ^2 and ψ may be estimated by maximizing the probability density function of $\tilde{\mathbf{Y}}^n$ seen as a function of the latter parameters:

$$\begin{aligned} L(\psi, \sigma^2, \tilde{\mathbf{y}^n}) &= p(\psi, \sigma^2, \tilde{\mathbf{y}^n}) \\ &= (2\pi)^{-\frac{n}{2}} \det\left[(K_n(\psi, \sigma^2) + \Delta_n)\right]^{-\frac{1}{2}} \\ &\quad \exp\left(-\frac{1}{2}\left(\tilde{\mathbf{y}} - \widehat{\mu}_n(\psi, \sigma^2)\mathbf{1}_n\right)^T (K_n(\psi, \sigma^2) + \Delta_n)^{-1} \left(\tilde{\mathbf{y}} - \widehat{\mu}_n(\psi, \sigma^2)\mathbf{1}_n\right)\right) \end{aligned} \tag{6}$$

or equivalently by minimizing:

$$\begin{aligned} l(\psi, \sigma^2, \tilde{\mathbf{y}^n}) &= \log\left(\det\left[(K_n(\psi, \sigma^2) + \Delta_n)\right]\right) + \\ &\quad \left(\tilde{\mathbf{y}} - \widehat{\mu}_n(\psi, \sigma^2)\mathbf{1}_n\right)^T (K_n(\psi, \sigma^2) + \Delta_n)^{-1} \left(\tilde{\mathbf{y}} - \widehat{\mu}_n(\psi, \sigma^2)\mathbf{1}_n\right) \end{aligned} \tag{7}$$

Note that contrarily to the noiseless case, no explicit expression for the optimal σ^2 as a function of ψ could be derived. Hence the optimization of l needs to be performed with respect to the whole vector of parameters (ψ, σ^2).

If the noise variances are not known but a simple parametric functional relashionship is assumed between the τ_i^2's and the t_i's, the corresponding parameters may be embedded within the ML procedure. For instance, assuming a Monte-Carlo-type behavior of the form $\tau_i^2 = C/t_i$, with $C \in \mathbb{R}^+$ independent of $\mathbf{x}_i$, l would simply become:

$$\begin{aligned} l(\psi, \sigma^2, \tilde{\mathbf{y}^n}, C) &= \log\left(\det\left[(K_n(\psi,\sigma^2) + \Delta_n(C))\right]\right) + \\ &\quad \left(\tilde{\mathbf{y}} - \widehat{\mu}_n(\psi,\sigma^2,C)\mathbf{1}_n\right)^T (K_n(\psi,\sigma^2) + \Delta_n(C))^{-1} \left(\tilde{\mathbf{y}} - \widehat{\mu}_n(\psi,\sigma^2,C)\mathbf{1}_n\right) \end{aligned} \quad (8)$$

3. KRIGING-BASED OPTIMIZATION; LIMITATIONS WITH NOISY FUNCTIONS

Optimization (say minimization) based on Kriging with noiseless observations has truly become a hit following the publication of the EGO algorithm (Jones et al. 1998). EGO consists in sequentially evaluating y at a point maximizing a figure of merit relying on Kriging, the *Expected Improvement* criterion (EI), and updating the metamodel at each new observation. As illustrated in Jones (2001), directly minimizing $m_n(.)$ is inefficient since it may lead the sequence of good points to get trapped in an artificial basin of minimum, whereas maximizing EI provides a right trade-off between exploitation and exploration in order to converge to a global minimizer. Our goal here is to adapt EI to the heterogeneously noisy case. Let us previously recall the definition and analytical expression of EI in the noiseless case.

Let $y_i = y(\mathbf{x}^i)$ $(1 \leq i \leq n)$, $\mathbf{y}^n = (y_1, \ldots, y_n)^T$, A_n denote the event $\left\{Y(\mathbf{x}^i) = y_i, 1 \leq i \leq n\right\}$, and m_n and s_n^2 still refer to the Kriging mean and variance. The idea underlying EI is that sampling at $\mathbf{x}$ will bring an improvement of $\min(y(\mathbf{X}^n)) - y(\mathbf{x})$ if $y(\mathbf{x})$ is below the current minimum $\min(y(\mathbf{X}^n))$, and 0 otherwise. Of course, this quantity cannot be known in advance since $y(\mathbf{x})$ is unknown. However, the GP model and the available information A_n make it possible to define and derive the following conditional expectation:

$$EI_n(\mathbf{x}) := \mathbb{E}\left[\left(\min(Y(\mathbf{X}^n)) - Y(\mathbf{x})\right)^+ | A_n\right] = \mathbb{E}\left[\left(\min(\mathbf{y}^n) - Y(\mathbf{x})\right)^+ | A_n\right] \quad (9)$$

An integration by parts yields the well-known analytical expression:

$$EI_n(\mathbf{x}) := \left(\min(\mathbf{y}^n) - m_n(\mathbf{x})\right) \Phi\left(\frac{\min(\mathbf{y}^n) - m_n(\mathbf{x})}{s_n(\mathbf{x})}\right) + s_n(\mathbf{x})\phi\left(\frac{\min(\mathbf{y}^n) - m_n(\mathbf{x})}{s_n(\mathbf{x})}\right), \quad (10)$$

where Φ and ϕ are respectively the cumulative distribution function and the probability density function of the standard Gaussian law. The latter analytical expression is very convenient since it allows fast evaluations of EI, and even analytical calculation of its gradient and higher order derivatives. This is used in particular in the DiceOptim R package (Roustant, Ginsbourger and Deville 2009) for speeding up EI maximization.

Let us now state why the classical EI is not well adapted to Kriging with noisy observations. Coming

back to the previous notations, we have indeed:

$$EI_n(\mathbf{x}) = \mathbb{E}\left[\left(\underbrace{\min(Y(\mathbf{X}^n))}_{\text{unknown}} - \underbrace{Y(\mathbf{x})}_{\text{unreachable}}\right)^+ \middle| \widetilde{A_n}\right], \tag{11}$$

which is not very satisfactory for at least two reasons. The first one is that the current minimum $\min(Y(\mathbf{X}^n))$ is not deterministically known conditionally on the noisy observations, contrarily to the noiseless case. The second reason is that the EI is based on the improvement that could bring a deterministic evaluation of y at the candidate point $\mathbf{x}$. Now, if the next evaluation is noisy, $Y(\mathbf{x})$ will remain non-exactly known. It would hence be more adapted to have a criterion taking the precision of the next measurement into account.

The first alternative to use the EI in the noisy case is to plug in a surrogate value for the unknown $\min(Y(\mathbf{X}^n))$. However, the natural choice $\min(\widetilde{\mathbf{y}^n})$ can be highly risky since the noisy minimum is a biased estimator of the noiseless minimum, and it suffices to have one highly noisy observation with a low value to deeply underestimate $\min(\mathbf{y}^n)$ for the rest of the optimization.

A rule of thumb proposed by Vazquez, Villemonteix, Sidorkiewicz and Walter (2008) is to plug in the minimum of the Kriging mean predictor $\min(m_n(\mathbf{X}^n))$ instead of $\min(\widetilde{\mathbf{y}^n})$, which seems to be a more sensible option in order to smooth out the noise fluctuations. In the same fashion, Huang, Allen, Notz and Zeng (2006) plug in the mean predictor at the training point with smallest kriging quantile.

Forrester, Keane and Bressloff (2006) proposed a *reinterpolation technique* that replaces the noisy observations by the kriging mean predictor $m_n(\mathbf{X}^n)$, and fits a noise-free kriging on such data, which is used for the standard EGO algorithm. However, this heuristic was designed for deterministic errors and does not allow repeated observations, which can be desirable in our framework.

However, on both cases the noise in the future response is not taken into account. In Huang, Allen, Notz and Zeng (2006), a variant of the EI called *Augmented Expected Improvement* (AEI) is proposed for uniformly noisy observations, the EI being multiplied by a penalization function $1 - \frac{\tau}{\sqrt{s_n^2(\mathbf{x})+\tau^2}}$ to account for the diminishing return of observation replicates, which is a first answer to that question.

A more rigorous alternative, as proposed in Gramacy and Lee (2010) and Gramacy and Polson (2011) consists of computing the EI based on the joint distribution of $(\min(Y(\mathbf{X}^n), Y(\mathbf{x}))$ conditional on $\widetilde{A_n}$; however, the EI is not analytically tractable in this form and must be estimated by expensive Monte-Carlo simulations, which makes the EI maximization challenging.

In the next section, we present an alternative infill criterion that takes into account past and future noises with transparent probabilistic foundations, and which can be derived analytically.

4. EXPECTED QUANTILE IMPROVEMENT

We now introduce a variant of EI for the case of a deterministic objective function with noisy measurements with heterogeneous variances. Our aim is to get a Kriging-based optimization criterion measuring which

level of improvement can be statistically expected from sampling y at a new $\mathbf{x}$ with a noise of given variance τ^2. A first question to be addressed is of decision-theoretic nature: what does the term "improvement" mean when comparing two sets of noisy observations? According to what kind of criterion should we judge that a set of noisy observations, or the associated metamodel, is better (in terms of minimization) after the $(n+1)^{\text{th}}$ measurement than before it?

Obviously, using only the noisy observations $\widetilde{\mathbf{y}^n}$ and $\widetilde{y}^{n+1}$ for that matter is a highly risky strategy, since the noise may introduce errors in the ranking of the observations. Here we propose to use the β-quantiles given by the Kriging conditional distribution, for a given level $\beta \in [0.5, 1[$: a point is declared "best" over a set of candidates $\mathbf{X}^n$ whenever it has the lowest β-quantile:

$$\mathbf{x}^* = argmin_{\mathbf{x}\in\mathbf{X}_n} q_n(\mathbf{x}) \tag{12}$$

with $q_n(\mathbf{x}) := \inf\{u \in \mathbb{R} : \mathbb{P}(Y(\mathbf{x}) \leq u|\widetilde{A_n}) \geq \beta\} = m_n(\mathbf{x}) + \Phi^{-1}(\beta)s_n(\mathbf{x})$, which is the decision criterion also used in Huang, Allen, Notz and Zeng (2006).

Now, we propose to define an improvement that is *consistent* with our decision criterion: improvement I is here the decrease of the lowest β-quantile, between the present step n and the forthcoming step $n+1$:

$$I = \left(\min\left(q_n\left(\mathbf{X}^n\right)\right) - q_{n+1}\left(\mathbf{x}^{n+1}\right)\right)^+ \tag{13}$$

Of course, like in the noiseless case, this improvement cannot be known in advance, because $q_{n+1}\left(\mathbf{x}_{n+1}\right)$ depends on the future observation $\tilde{y}_{n+1}$. However, thanks to the particular form of the kriging equations, the future quantile q_{n+1} can be predicted, and consequently the EI calculated, based on the GP model at step n, as we show below.

One can remark that we restrict here our choice to the observed points ($\mathbf{X}^n$ and $\mathbf{x}^{n+1}$), even though a similar criterion could be defined over the entire design space: $I = \left(\min_D\left(q_n\left(\mathbf{x}\right)\right) - \min_D\left(q_{n+1}\left(\mathbf{x}\right)\right)\right)^+$. Such a restriction greatly simplifies calculations, and is a reasonable conservative measure: indeed, who would trust a metamodel so much to propose a final candidate minimizer without any measurement at that point?

Let us denote by $Q_i(\mathbf{x})$ the kriging quantile $q_i(\mathbf{x})$ ($i \leq n+1$) where the measurements are still in their random form, and define the Expected Quantile Improvement (EQI) as:

$$EQI_n(\mathbf{x}^{n+1}, \tau^2_{n+1}) := \mathbb{E}\left[\left(\min_{i\leq n}(Q_n(\mathbf{x}^i)) - Q_{n+1}(\mathbf{x}^{n+1})\right)^+ \middle| \widetilde{A_n}\right] \tag{14}$$

where the dependence on the future noise τ^2_{n+1} appears through $Q_{n+1}(\mathbf{x})$'s distribution.

The randomness of $Q_{n+1}(\mathbf{x})$ conditional on $\widetilde{A_n}$ is indeed a consequence from $\widetilde{Y_{n+1}} := Y(\mathbf{x}^{n+1}) + \varepsilon_{n+1}$ having not been observed yet at step n. However, following the fact that $\widetilde{Y_{n+1}}|\widetilde{A_n}$ is Gaussian with known mean and variance, one can show that $Q_{n+1}(.)$ is a GP conditional on $\widetilde{A_n}$ (see proof and details in appendix).

Furthermore, $\min_{i\leq n}(Q_n(\mathbf{x}^i))$ is known conditional on $\widetilde{A_n}$. As a result, the proposed EQI is analytically tractable, and we get by a similar calculation as in Eq. 10:

$$EQI_n(\mathbf{x}^{n+1}, \tau^2_{n+1}) = \left(\min(\mathbf{q}^n) - m_{Q_{n+1}}\right) \Phi\left(\frac{\min(\mathbf{q}^n) - m_{Q_{n+1}}}{s_{Q_{n+1}}}\right) + s_{Q_{n+1}} \phi\left(\frac{\min(\mathbf{q}^n) - m_{Q_{n+1}}}{s_{Q_{n+1}}}\right) \quad (15)$$

where:
$$\begin{cases} \mathbf{q}^n := \{q_n(\mathbf{x}^i), i \leq n\} \text{ is the set of current quantile values at the already visited points,} \\ m_{Q_{n+1}} := \mathbb{E}[Q_{n+1}(\mathbf{x}^{n+1})|\widetilde{A_n}] \text{ is } Q_{n+1}(\mathbf{x}^{n+1})\text{'s conditional expectation —seen from step } n, \\ s^2_{Q_{n+1}} := Var[Q_{n+1}(\mathbf{x}^{n+1})|\widetilde{A_n}] \text{ is its conditional variance, both derived in appendix.} \end{cases}$$

As in the noiseless case, the EQI criterion is hence known in closed form, which is a desirable feature for its maximization. τ^2 and β are to be considered here as parameters, and EQI maximization is done with respect to $\mathbf{x}^{n+1}$ only. Making a measurement where EQI is maximum is then the optimal strategy (in expectation) with respect to the decision rule, which is here choosing the lowest kriging quantile.

Note that the EQI statistic shares some concepts with the so-called *Expected Conditional Improvement* (ECI) proposed by Gramacy and Lee (2010), which uses the knowledge that a (noiseless) measurement will be made at a point $\mathbf{u}$ to deduce an improvement at another point $\mathbf{v}$, using the distribution of $M_{n+1}(.|\widetilde{A}_n)$ (as defined in appendix). ECI and EQI coincide in some special case (using $\beta = 0.5$, $\mathbf{u} = \mathbf{v}$, $f_{min} = \min(m_n(\mathbf{X}^n))$ and modifying ECI to account for noise).

The EQI criterion has the following important properties:

- in absence of future noise ($\tau^2_{n+1} = 0$), the future quantile at $\mathbf{x}^{n+1}$ coincides with the observation $\tilde{y}^{n+1} = y(\mathbf{x}^{n+1})$, since the prediction variance at this point will be null; it follows directly that $Q_{n+1}(\mathbf{x}^{n+1})|\widetilde{A_n} = Y_{n+1}|\widetilde{A_n}$, so the EQI is then equal to the classical EI with a plugin of the kriging quantiles for $\min(\mathbf{y}^n)$
- in absence of past noise (for the n first observations), $\min(\mathbf{q}^n)$ is equal to the minimum of the observations, $\min(\mathbf{y}^n)$
- in absence of both past and future noise, the EQI is then equal to the classical EI.

β tunes the level of reliability wanted on the final result (which acts thus somehow similarly to the power parameter of the generalized improvement of Schonlau, Welch and Jones (1998)). With $\beta = 0.5$, the design points are compared based on the kriging mean predictor only, without taking into account the prediction variance at those points, while high values of β (i.e. near to 1) penalize designs with high uncertainty, which is a more conservative approach. Hence, with a high β, the criterion is more likely to favor observation repetitions or clustering, in order to locally decrease the prediction variance, while with $\beta = 0.5$, the criterion can be expected to be tendentially more exploratory.

The future noise τ^2_{n+1} also strongly affects the shape of the EQI. Indeed, a very noisy future observation can only have a very limited influence on the kriging model. Then, the only possibility to have a non-null improvement is either to sample where $q_n(\mathbf{x})$ is minimum if this point is not in $\mathbf{X}^n$ (the lowest quantile will then be chosen on the set $\mathbf{X}^{n+1}$ instead of $\mathbf{X}^n$, which will bring an improvement even if $q_{n+1} = q_n$), or to sample at the current best point, which may decrease its uncertainty and brings a small but measurable improvement. On the contrary, if τ^2_{n+1} is very small, the EQI behaves like the classical EI, making the well-known trade-off between exploration and exploitation.

Figure 3 illustrates the dependence of the EQI on both τ^2_{n+1} and β. The actual function, DoE and kriging model are taken from section 7.1.

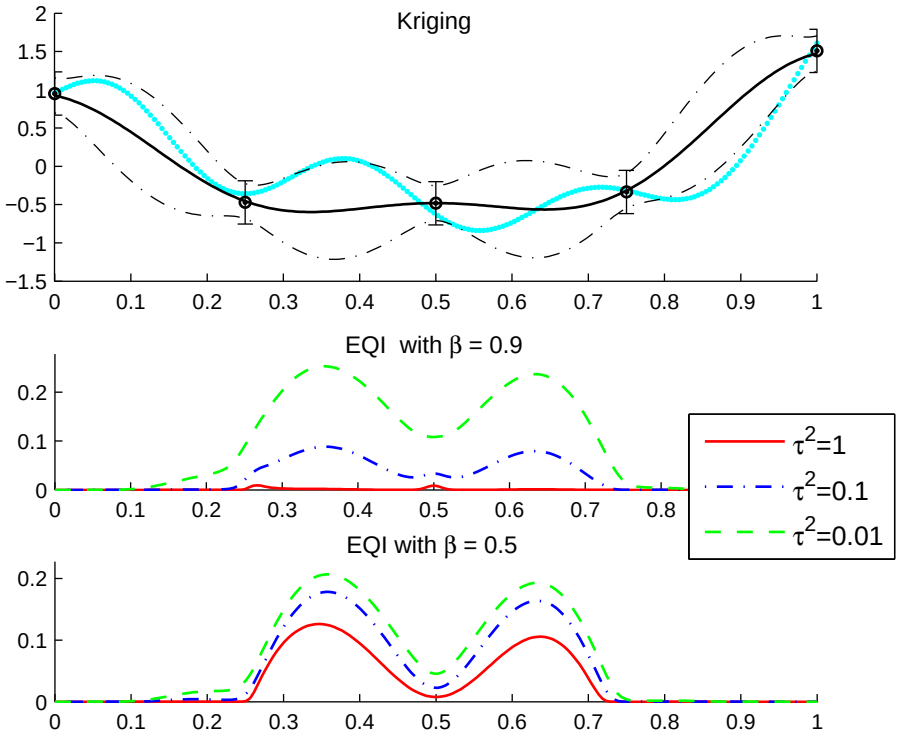

Figure 3: *Upper graph*: Actual function and Kriging (presented as in Figure 2). *Middle and lower graphs*: Corresponding EQI for three different future noise levels and two different quantiles.

We can see that the choice of the future noise level has a great influence on the criterion. With small noise variance, the EQI behaves like the classical EI, with highest values in regions with high uncertainty and low mean predictions. With higher noise variances and high quantiles, the criterion becomes very conservative since is it high only in the vicinity of existing measurements. With $\beta = 0.5$, the EQI is high even for the largest variance.

The proposed EQI criterion seems to be conservative compared to other EI variants, since it intends to measure the *actual* improvements on the final decision criterion, which can be very limited for a single step. In the next section, we propose a numerical trick that turns this property into a substantial asset, and results in an optimization strategy less myopic than the classical EGO.

5. OPTIMIZATION WITH FINITE COMPUTATIONAL BUDGET

It is well-known that the EGO algorithm is a so-called *myopic* strategy, since its criterion EI always considers the next step as if it were the last one. However, for most computer experiments, the total computational budget is bounded, and prescribed by industrial constraints such as time and power limitations. In the deterministic framework, this results in a limited (given) number of observations for optimization. It has been shown in that case in Mockus (1988) followed by Ginsbourger and Le Riche (2010) that taking into account the finite budget may modify the optimization strategy and improve significantly its efficiency.

In our framework of simulators with tunable precision, the concept of finite budget is particularly critical, since each observation requires a trade-off between accuracy and rapidity, and in general, the user has to trade off between the total number of observations and their precision. In linear modeling, this problem is typical of the theory of optimal designs (Fedorov and Hackl 1997), with the notable difference that we face it here within a sequential strategy.

Coming back to the notations of section 2, we defined a noisy measurement as depending on both a location $\mathbf{x}_i$ and a computational time t_i, and which variance is given by $\tau^2(t_i)$ (for the clarity of the writing the potential dependence on $\mathbf{x}_i$ is not considered here). The computational constraint implies that the sum of all computational times is fixed to a given budget, say T_0. At step n, the remaining budget for optimization is $T_{n+1} = T_0 - \sum_{i=1}^{n} t_i$.

The EQI criterion allows taking into account such computational boundedness in the choice of the new candidate observations. Indeed, the future noise level τ^2_{n+1}, which is a parameter of the EQI criterion, will stand here for the finite resource. Given a computational budget T_{n+1}, the smallest noise variance achievable (i.e. the largest precision) for a new measurement is $\tau^2(T_{n+1})$, assuming that all the remaining budget will be attributed to this measurement. Note that in the course of the optimization process, the remaining budget decreases, so $\tau^2(T_{n+1})$ increases with n.

Then, we propose to set $\tau^2_{n+1} = \tau^2(T_{n+1})$ for the EQI calculation within our sequential optimization procedure, meaning that the EQI will measure the potential improvement if all the remaining budget would be attributed to the next observation. Of course, the actual budget for the next observation may be a lot smaller than T_{n+1} so the optimization does not stop after one step. With this setting, the new experiment is chosen knowing that even if all the budget was used for a single observation, its noise variance would not decrease below a certain value.

Consequently, the EQI will behave differently at the beginning and at the end of the optimization. When the budget is high, EQI will be tendencially higher in unexplored regions, since it is where accurate measurements are likely to be most efficient (the EQI will actually be almost similar to a classical EI). At the end of the optimization, however, when the remaining time is small, the EQI will be small in unexplored regions since even if the actual function is low, there is not enough computational time to obtain a lower quantile than the current best one. In that case, the EQI will be highest close to the current best point(s)

and favor local search.

6. ALLOCATION OF RESOURCE

In the previous sections, we proposed an infill criterion for choosing sequentially sample points, and a specific tuning of the criterion to account for finite computational resources. However, the question of the choice of the computational budget for a single observation has yet been left open. For clarity purpose, this section is divided in two parts: first, the update schemes are described in the case of constant (or uniform) allocation; then, a strategy is proposed to take advantage of the response convergence monitoring to dynamically adapt the budget to each measurement.

6.1 Constant allocation

We assume first that the computational budget T_0 can be divided in elementary time steps $t_e \in]0, +\infty[$, so that $T_0 = N \times t_e$. An elementary step can correspond for example to a given number of solver iterations for partial convergence, or a number of drawings for stochastic simulators. An algorithm with constant allocation will then attribute one by one the N elementary time steps to either generate new measurements or improve accuracy on existing ones.

At step n, a budget $n \times t_e < T_0$ as already been spent on the measurements. At unsampled locations, the criterion is simply evaluated with $\tau^2_{n+1} = \tau^2(T_{n+1})$ (where $T_{n+1} = T_0 - nt_e$). At existing observations, a different value has to be used; if not, the EQI would estimate the value of a *new* measurement with variance $\tau^2(T_{n+1})$, instead of the value of improving the existing measurement.

To compute this value, we use the fact that it is equivalent for the kriging model to have at the same point several measurements with independent noises or a single equivalent measurement that is the weighted average of the observations. For instance, let $\widetilde{y}_{i,1}$ and $\widetilde{y}_{i,2}$ be two measurements with respective noise levels $\tau^2_{i,1}$ and $\tau^2_{i,2}$. They are equivalent to a single measurement

$$\widetilde{y}_{i,eq} = \frac{\tau_{i,1}^{-2}\widetilde{y}_{i,1} + \tau_{i,2}^{-2}\widetilde{y}_{i,2}}{\tau_{i,1}^{-2} + \tau_{i,2}^{-2}} \tag{16}$$

with variance $\tau^2_{i,eq}$ the harmonic mean of $\tau^2_{i,1}$ and $\tau^2_{i,2}$, namely:

$$\frac{1}{\tau^2_{i,eq}} := \frac{1}{\tau^2_{i,1}} + \frac{1}{\tau^2_{i,2}} \Longrightarrow \tau^2_{i,eq} = \frac{\tau^2_{i,1}\tau^2_{i,2}}{\tau^2_{i,1} + \tau^2_{i,2}} \tag{17}$$

Now, for the EQI calculation, we want to measure the effect of carrying a measurement with current error variance $\tau^2(t_i)$ until the variance $\tau^2(t_i + T_{n+1})$ is reached. This is equivalent, in terms of the kriging model, to make a new measurement with noise variance:

$$\tau^2(t_i \to t_i + T_{n+1}) := \frac{\tau^2(t_i)\tau^2(t_i + T_{n+1})}{\tau^2(t_i) - \tau^2(t_i + T_{n+1})} \tag{18}$$

This formula is obtained from Eq. 17, with $\tau^2_{i,eq} = \tau^2(t_i + T_{n+1})$, $\tau^2_{i,1} = \tau^2(t_i)$ and $\tau^2_{i,2} = \tau^2(t_i \to t_i + T_{n+1})$. If $t_i = 0$, we have $\tau^2(t_i \to T_{n+1}) = \tau^2(T_{n+1})$.

Once the new point $\mathbf{x}^{n+1}$ is chosen and the measurement is made, the kriging model has to be updated. If the new point does not belong to the current DoE ($\mathbf{x}^{n+1} \notin \mathbf{X}_n$), the point is added to the DoE and the kriging equations modified accordingly; otherwise, it only requires to replace the previous values of response and noise $\widetilde{y}_i[b_i]$ and $\tau^2_i[b_i]$ by $\widetilde{y}_i[b_i + 1]$ and $\tau^2_i[b_i + 1]$ in the kriging equations (5). The remaining time is then updated and the next point can be chosen. The algorithm with constant allocation is presented in pseudo-code form in table 1.

Table 1: EQI algorithm with constant allocation

- Choose T_0, β, t_e, n_0
- Build initial DoE $\mathbf{X}^{n_0}$, generate observations $\widetilde{\mathbf{y}}^{n_0}$, fit Kriging model
- Set $n = n_0$, $b_i = 1 (1 \leq i \leq n_0)$ and $T_n = T_0 - n_0 t_e$

while $T_n > 0$

- Choose new design point $\mathbf{x}^{n+1}$ that maximizes $EQI_n\left(., \tau^2_{n+1}\right)$, with:
  - $\tau^2_{n+1} = \tau^2(T_{n+1})$ if $\mathbf{x} \notin \mathbf{X}_n$
  - $\tau^2_{n+1} = \tau^2(b_i \times t_e \to b_i \times t_e + T_{n+1})$ if $\mathbf{x} = \mathbf{x}^i$
- **If** $\mathbf{x}^{n+1} \in \mathbf{X}_n$:
  - Update corresponding $\widetilde{y}_i$ and τ^2_i, set $b_i = b_i + 1$
- **Otherwise**
  - Generate $\widetilde{y}_{n+1}$, set $\tau^2_{n+1} = \tau^2(t_e)$ and $b_{n+1} = 1$
  - Add $\mathbf{x}^{n+1}$ to the DoE and $\widetilde{y}_{n+1}$ to the observations
  - Set $n = n + 1$
- **end if**
- Update kriging model
- Set $T_{n+1} = T_{n+1} - t_e$

end while

- Choose $\mathbf{x}^*$ based on the Kriging quantiles at the measurement points.

Remark:

The previous procedure allows to address the problem of optimization of a homogeneously noisy function (as in Huang, Allen, Notz and Zeng (2006) for instance), which can be seen as a particular case of constant allocation. We consider that each observation requires a constant time t_e, and has a constant noise variance ν^2. At each optimization step, the user has the possibility to either sample at a new location or duplicate an existing measurement, so attributing all the remaining budget to a measurement means performing $N - n$ replications at this point, hence leading to the situation described in 3. In that case, it is straightforward to get $\tau^2(t_i \to t_i + T_{n+1}) = \frac{\nu^2}{N-n} = \tau^2(T_{n+1})$, so the criterion writes similarly at sampled and unsampled locations. This, in this framework, the procedure simplifies to maximizing at each step $EQI_n(., \frac{\nu^2}{N-n})$.

6.2 On-line allocation

The constant allocation strategy of the previous section performs $N - n_0$ EGO iterations, and each requires running an inner optimization loop for the maximization of the EQI, which can be very time-consuming. Hence, the elementary time step t_e must be chosen large enough to limit the number of EQI optimizations.

Typically, with partial convergence or stochastic simulators, t_e must be chosen a lot larger than a single solver iteration or drawing, respectively. This limitation can greatly hinder the flexibility and potential of tunable precision, since it reduces the possibilities of a quasi-continuum of fidelities to a few discrete precision levels.

Here, we propose to overcome this problem by using a heuristic for dynamically choosing the computational resource given to an experiment. A simple way to do so is to monitor the evolution of the EQI at the current observation point. Indeed, instead of maximizing the EQI after each t_e is spent, we will choose an observation point, and allocate several time steps on it until a criterion is met. As for the constant allocation case, the EQI is updated after each step, by replacing the previous values of response and noise $\widetilde{y}_i[b_i]$ and $\tau_i^2[b_i]$ by $\widetilde{y}_i[b_i+1]$ and $\tau_i^2[b_i+1]$ in the kriging equations, and by replacing the future noise level $\tau^2(t_i \to t_i + T_{n+1})$ by $\tau^2(t_i + t_e \to t_i + T_{n+1} - t_e)$.

The updated EQI tends by construction to decrease when computation time is added, since (a) the kriging uncertainty reduces at the observation point, and (b) EQI decreases when $\tau^2(t_i \to t_i + T_{n+1})$ increases. However, if the measurement converges to a good (small) value, EQI can increase temporarily. Inversely, if the measurement converges to a high value, EQI decreases faster. Hence, we can define a ("point switching") stopping criterion for resource allocation based on EQI. If the EQI decreases below a certain value, carrying on the calculations is not likely to help the optimization, so the observation process should stop and another point be chosen. Here, we propose to interrupt a measurement and search for a new point when the current value of the EQI is less than a proportion of the initial EQI value (that is, the value of EQI when starting the measurement process at that point), for instance 50%.

The sequence of this new procedure is as follow: first, choose the point with highest EQI given the whole remaining budget, store the corresponding value as reference, and then invest new elementary measurements at this point until the EQI with updated data falls under a given proportion $\gamma \in]0,1[$ of the reference value. The operation of choosing the most promising point is then started again, and so on until the total computational budget has been spent. Note that the final number of measurements and EQI maximizations are not determined beforehand but adapts automatically to the budget and resource distribution, and may be a lot smaller than the number of steps N. The algorithm is presented in pseudo-code form in table 2. For the sake of conciseness, this algorithm does not consider the case where $\mathbf{x}^{n+1} \in \mathbf{X}^n$, which requires to be treated differently, as in 1.

6.3 Practical issues

Budget discretization: The total computational budget T_0 needs to be defined before optimization, and discretized in incremental steps $\{t_e, 2 \times t_e, \ldots, N \times t_e\}$, where $N = T_0/t_e$. Smaller steps (i.e. a smaller t_e) result in increased precision, but requires more Kriging updates and probably more EQI maximizations, which can become computationally intensive. A prescribed fraction T_0 of this budget is allocated to build an

Table 2: EQI algorithm with on-line resource allocation

- Build initial DoE $\mathbf{X}^{n_0}$, generate observations $\widetilde{\mathbf{y}}^{n_0}$ using T_{n_0} computational time, fit Kriging model
- Set $n = n_0$ and $T_n = T_0 - T_{n_0}$

while $T_n > 0$

- Choose new design point $\mathbf{x}^{n+1}$ that maximizes $EQI_n\left(., \tau^2(T_n)\right)$
- Generate $\widetilde{y}_{n+1}[1]$ with one time increment
- Augment DoE: $\mathbf{X}^{n+1} = \left\{\mathbf{X}^n, \mathbf{x}^{n+1}\right\}$
- Update Kriging model with $\widetilde{y}_{n+1} = \widetilde{y}_{n+1}[1]$ and $\tau^2_{n+1} = \tau^2(t_e)$
- Set $T_{n+1} = T_n - t_e$, $j = 1$, and $t = t_e$

while $EQI_{n+1}\left(\mathbf{x}^{n+1}, \tau^2(t_{n+1} \to t_{n+1} + T_{n+1})\right) > \gamma EQI_n\left(\mathbf{x}^{n+1}, \tau^2(T_n)\right)$

- Generate $\widetilde{y}_{n+1}[j+1]$ by adding one time increment
- Set $T_{n+1} = T_{n+1} - t_e$, $j = j + 1$, and $t_{n+1} = t_{n+1} + t_e$
- Update Kriging model with: $\widetilde{y}_{n+1} = \widetilde{y}_{n+1}[j]$, $\tau^2_{n+1} = \tau^2(t)$

end while

- Set $n = n + 1$

end while

- Choose final design based on the Kriging quantile

initial DoE, which should be designed in order to fit a realistic Kriging model. Based on previous numerical experiments, it has been found that using 10% to 30% of the total budget on a space-filling DoE (for instance, an LHS design) with uniform observation variances is a reasonable option.

EQI maximization: Although the EQI criterion is analytical, its maximization with respect to $\mathbf{x}^{n+1}$ is potentially complex and time-consuming. Indeed, the EQI (like the classical EI) is highly multimodal and requires the use of global search algorithms (population-based techniques for instance). Also, each EQI evaluation requires the inversion of a $(n+1) \times (n+1)$ covariance matrix, which can be expensive for large n. A valuable computational shortcut can be achieved in the update of the inverse of the covariance matrix when adding an observation (see Marrel (2008) or Gramacy and Polson (2011)). It is to be noted that the gradients of EQI can be calculated analytically for a given covariance kernel (in the fashion of Ginsbourger (2009) (Chapter 4)).

Variance dependence on $\mathbf{x}$: On the two algorithms proposed in this section, the variance level depends only on computational time; however, the algorithms write similarly if the variance is design-dependent, by replacing $\tau^2(t)$ by $\tau^2(t; \mathbf{x})$. Although not considered in the rest of this article, it is the authors' belief that such framework would be particularly adapted for the EQI criterion, since it would automatically take the ratio cost/precision into account for the optimization strategy.

Setting a bound for minimum noise: Also, a minimum achievable noise can be set by the user. In that case, τ^2 should be bounded by a $\tau^2_{min} = \tau^2(T_{max})$, and T_n replaced by $\min(T_n, T_{max})$ in all EQI expressions. In that case, we would have $\tau^2(t \to T_i) = \tau^2(t \to T_{max})$ for $T_i > T_{max}$, and $\tau^2(T_{max} \to T_i) = +\infty$, which would result in $EQI = 0$, so once T_{max} time is spent on an observation, it is never chosen again.

Sensitivity to model accuracy: The performance of the EQI strategy strongly relies on the quality of the kriging model. It has been observed that the EQI is particularly sensitive to underestimation of the activity of the actual function, that is underestimation of the process variance and overestimation of range parameters. This issue also observed for the classical EI, see Forrester and Jones (2008).

Time-variance relashion knowledge: Finally, a crucial point is that the EQI criterion requires the relation between the error variance and computational time to be known. As noted in section 2, a preliminary study may be needed to calibrate an error model, for instance by studying the convergence of a small number of simulations and inferring a parametric model for the noise. Those simulations can then be integrated in the initial DOE. Such error model might of course be updated during the optimization with the help of the new observations.

7. EXPERIMENTS

In the first part of this section, we show a one-dimensional analytical example for illustrative purpose; then, the proposed method is compared to existing algorithms using five and six-dimensional functions. Finally, the EQI is applied to a two-dimensional nuclear safety problem.

7.1 One-dimensional example

The objective function is defined over $[0, 1]$ by:

$$y(x) = \frac{1}{2}\left(\frac{sin(20x)}{1+x} + 3x^3 cos(5x) + 10(x-0.5)^2 - 0.6\right) \tag{19}$$

The noise is here inversely proportional to computational time, and independent of $\mathbf{x}$: $\tau^2(t) = 0.1/t$.

The initial DoE consists of five equally-spaced measurements, with noise variances equal to 0.02 (the 95% confidence interval at a measurement point is approximately 25% of the range of y). The kriging model has a gaussian covariance kernel with parameters $\sigma = 1$ and $\theta = 0.1$. The optimization is performed with a total computational budget of $T_0 = 100$, starting from the DoE described above. T_0 is divided in 100 time increments. Each initial DoE measurement has required five time units ($t_e = 1$), so the DoE used 25% of the computational budget. Here the kriging parameters are assumed to be known and are not re-evaluated at each iteration.

Figure 4 represents the final DoE and kriging model. Nine measurement points have been added, with computational times varying from one to 41. The final DoE consists of highly noisy observations space-filling the design region and a cluster of accurate observations in the region of the global optimum.

7.2 Comparison to the Augmented Expected Improvement (AEI) procedure

The proposed strategy is compared to the AEI method as proposed in Huang, Allen, Notz and Zeng (2006) for the optimization of homogeneously noisy experiments, which has been found to be already very competitive

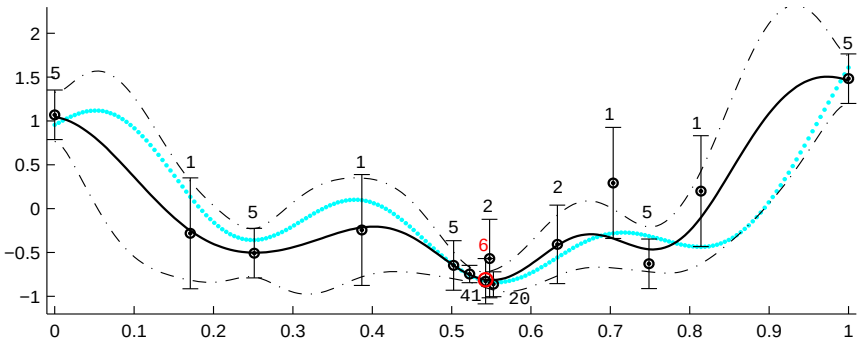

Figure 4: Observations and kriging after optimization. The numbers are the times steps for each observation. The chosen best observation is represented with the larger circle.

compared to non-kriging based local or global optimizers such as revised simplex search (Humphrey and Wilson 2000) or DIRECT (Gablonsky and Kelley 2001). Both EQI and AEI heuristics are compared to the classical EI, with the minimal value of the observations replaced by the minimum of the kriging mean at the observations, which can be considered as the baseline approach. As test problems, we employed two analytical benchmark problems, the six-dimensional *Hartman* function (Dixon and Szego 1978) and the five-dimensional *Ackley* function (Ackley 1987).

Hartman:

$$y(\mathbf{x}) = \frac{-1}{1.94}\left[2.58 + \sum_{i=1}^{4} C_i \exp\left(-\sum_{j=1}^{6} a_{ji}\left(x_j - p_{ji}\right)^2\right)\right] \tag{20}$$

with: $\mathbf{C} = [1.0, 1.2, 3.0, 3.2]$, $\mathbf{a} = \begin{bmatrix} 10.00 & 0.05 & 3.00 & 17.00 \\ 3.00 & 10.00 & 3.50 & 8.00 \\ 17.00 & 17.00 & 1.70 & 0.05 \\ 3.50 & 0.10 & 10.00 & 10.00 \\ 1.70 & 8.00 & 17.00 & 0.10 \\ 8.00 & 14.00 & 8.00 & 14.00 \end{bmatrix}$, $\mathbf{p} = \begin{bmatrix} 0.1312 & 0.2329 & 0.2348 & 0.4047 \\ 0.1696 & 0.4135 & 0.1451 & 0.8828 \\ 0.5569 & 0.8307 & 0.3522 & 0.8732 \\ 0.0124 & 0.3736 & 0.2883 & 0.5743 \\ 0.8283 & 0.1004 & 0.3047 & 0.1091 \\ 0.5886 & 0.9991 & 0.6650 & 0.0381 \end{bmatrix}$.

Ackley:

$$y(\mathbf{x}) = -20\exp\left(-0.2\sqrt{\frac{1}{d}\sum_{i=1}^{d} x_i^2}\right) - \exp\left(\frac{1}{d}\sum_{i=1}^{d}\cos(2\pi x_i)\right) + 20 + \exp(1) \tag{21}$$

Both functions are normalized so their design region D is $[0,1]^d$ and their standard deviation is one over D. Their minima are zero for Ackley and -1.94 for Hartman. A Gaussian noise $\epsilon \sim \mathcal{N}(0, 10\tau^2)$ is added to the analytical functions.

In order to model a tunable fidelity framework while allowing a fair comparison between methods, each noisy measurement $\tilde{y}_i$ is taken as the average of several function evaluations as described in section 2.3. For the AEI procedure, which is designed for homoscedastic noise, ten time steps are used for each observation, so the noise variance is τ^2. For the EQI procedure, the noise variance potentially varies between $10\tau^2$ and

$10\tau^2/T_0$.

For both methods, the initial DoEs are chosen as LHS designs with maximin criterion, and are generated using ten time steps for each observation. The total optimization budget is chosen equal to two times the budget needed to generate the initial DOE. Two versions of EQI are tested: with $\beta = 0.5$ (decision based on kriging mean only) and with $\beta = 0.9$; the criteria are referred to as $EQI.50$ and $EQI.90$, respectively.

Several budgets, noise levels and initial DOE sizes are tested. The different configurations are summarized in Table 3. The noise level τ can be compared to objective function standard deviation (SD), which is one for both functions. With $\tau = 0.2$, the optimization problem can be considered as a very noisy. The total budget is deliberately chosen very small since it may correspond to typical situations in real-life applications.

Table 3: Summary of the test problem configurations

Function	**Initial DoE size** n_0	**Budget** T_0	τ
Ackley	25 points	500 steps	0.05
Ackley	50 points	1000 steps	0.2
Hartman	60 points	1200 steps	0.2

For each configuration, 40 initial DoEs and observations are generated to account for randomness in the LHS designs and the observations. The kriging parameters are estimated only at the initial step, using the R package *DiceKriging* (Roustant et al. 2009). The results are presented using boxplots on Figure 5.

For the Ackley function, with $\tau = 0.05$, the EQI procedure outperforms the AEI procedure in terms of actual value and kriging uncertainty at the best design. The choice of $\alpha = 0.5$ for the quantile level provides the best results. With $\tau = 0.2$, AEI provides the best results in terms of optimization. However, the standard deviation at current best point is significantly lower for $EQI.90$ than for the other methods, which illustrates the tendency of this method to reduce uncertainty at the expense of exploration.

For the Hartman function, EQI slightly outperforms the two other methods in terms of actual value at best design. In terms of kriging uncertainty at best design, EI and AEI are clearly less efficient than EQI. The difference between the strategies $\alpha = 0.5$ and $\alpha = 0.9$ appear clearly: indeed, with $\alpha = 0.5$, the choice of the best design is made on the kriging mean only; on the other hand, with $\alpha = 0.9$, observations with high uncertainty are penalized so the kriging standard deviation at best design is always small.

Table 4 shows the average number of distinct measurements and the average number of time steps at best design. Even though EI and AEI use uniform allocation, their values are not constant because some measurements are repeated (i.e. criteria are maximal at existing measurements during optimization). For instance, the first row and last column of the table indicates that for EI, there is on average four measurements at best design.

For the small budget and small noise on the Ackley function, the online allocation resulted with more measurements than for AEI and EI. Here, online allocation was used to improve exploration by having more measurement locations and resulted in better performance in terms of optimization (see Figure 5).

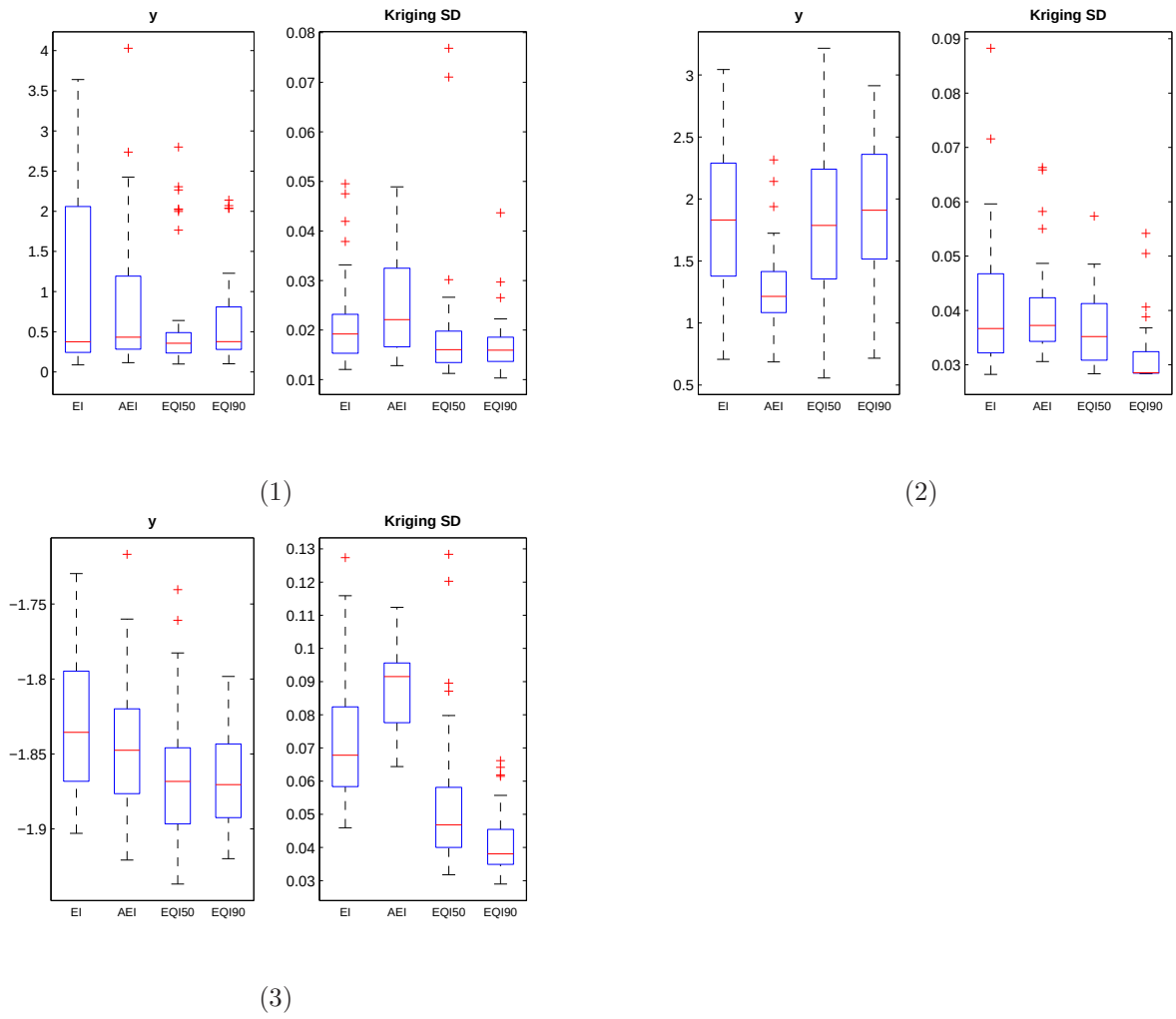

Figure 5: Optimization results for the different methods on (1) the Ackley function with $\tau = 0.05$ and 500 steps (top left), (2) the Ackley function with $\tau = 0.2$ and 1000 steps (right), (3) the Hartman function $\tau = 0.2$ and 1200 steps (bottom).

On the contrary, for the large budget and large noise, it resulted with accurate measurements at best design to the detriment of optimization. For the Hartman function, the number of measurements are almost equivalent for all methods.

It is interesting to note that for the first configuration on the Ackley function and for the Hartman function, the average number of time steps at best design is smaller for EQI than for EI and AEI but the kriging standard deviation is also smaller (Figure 5), which is counter-intuitive. For EQI, the small standard deviation is obtained because the measurements form a cluster around the best design.

Table 4: Computational time allocation during optimization

Configuration	Number of distinct measurements				Time steps at best design			
(Function, Initial DOE, Budget, τ)	EQI.50	EQI.90	AEI	EI	EQI.50	EQI.90	AEI	EI
Ackley, 25 points, 500 steps, 0.05	76	70	45	44	6	6	43	40
Ackley, 50 points, 1000 steps, 0.2	72	63	87	98	21	64	15	11
Hartman, 60 points, 1200 steps, 0.2	130	118	117	103	5	8	23	38

7.3 Application to a 2D benchmark from nuclear criticality safety assessments

In this section, the optimization algorithm is applied to the problem of safety assessment of a nuclear system involving fissile materials. The benchmark system used is an interim storage of dry PuO_2 powder into a regular array of storage tubes. The criticality safety of this system is evaluated through the neutron multiplication factor (called k-effective or k_{eff}), which models the nuclear chain reaction trend:

- $k_{\text{eff}} > 1$ is an increasing neutrons production leading to an uncontrolled chain reaction,
- $k_{\text{eff}} = 1$ means a stable neutrons population as required in nuclear reactors,
- $k_{\text{eff}} < 1$ is the safety state required for all unused fissile materials, like for fuel storage.

The neutron multiplication factor depends on many parameters such as the composition of fissile materials, operation conditions, geometry, etc. For a given set of parameters, the value of k_{eff} can be evaluated using the MORET stochastic simulator (Fernex et al. 2005), which is based on Markov Chain Monte-Carlo (MCMC) simulation techniques. The precision of the evaluation depends on the amount of simulated particles (neutrons), which is tunable by the user.

When assessing the safety of a system, one has to ensure that, given a set of admissible values D for the parameters $\mathbf{x}$, there are no physical conditions under which the k_{eff} can reach the critical value of 1.0 (minus a margin, usually chosen as 0.05):

$$\max_{\mathbf{x}\in D} \quad k_{\text{eff}}(\mathbf{x}) \leq 1.0 - \text{margin} \tag{22}$$

The search for the worst combination of parameters $\mathbf{x}$ defines a noisy optimization problem which is often challenging in practice, due to the possible high computational expense of the MORET simulator. An efficient resolution technique of this problem is particularly crucial since this optimization may be done numerous times. To account for the noise, classically the actual k_{eff} is replaced by its estimate $\hat{k}_{\text{eff}}$ plus three output standard deviations. Here this conservative measure is replaced by the kriging quantile approach.

In this article, we focus on the maximization of k_{eff} with respect to two parameters, the other possible inputs being fixed to their most penalizing values (based on expert knowledge):

- *d.puo*2, the density of the fissile powder, with original range [0.5, 4] $g.cm^{-3}$, rescaled to the [0, 1] interval,
- *d.brouiscale*, the density of water between storage tubes, with range [0, 1], which accounts for the possible flooding of the storage (leading to an interstitial moderation of the neutrons interacting from a storage tube to another one).

Hence, to agree with previous notations, we set: $\mathbf{x} = (d.puo2, d.brouiscale)$, and $y(\mathbf{x}) = -k_{\text{eff}}(\mathbf{x})$. Simulation time is assumed proportional to the number of particles simulated (the entry cost of a new simulation being neglected). Since the simulator is based on MCMC method, the variance of the k_{eff} estimate is exactly inversely proportional to the number of particles. The variance slightly varies with input parameters, but this dependence can be considered negligible here.

For practical considerations, the optimization space D is discretized in a 75×75 grid, and for each new measurement the EQI maximization is performed by exhaustive search on the grid. The incremental time step t_e is defined by the simulation of 4000 particles, which takes about half of minutes on a 3 GHz CPU. The response noise standard deviation can take values between 5.23×10^{-2} (one time step) and 4.01×10^{-3} (200 time steps).

To evaluate the efficiency of our algorithm, all the 5625 points of the grid have been evaluated with highest precision, which gives an accurate estimation of the shape (represented in figure 6), minimal value and minimizer x^* of the function. With this accurate dataset we find that $x^* = [0.1892, 0.0811]$ and $f(x^*) = -0.9847$.

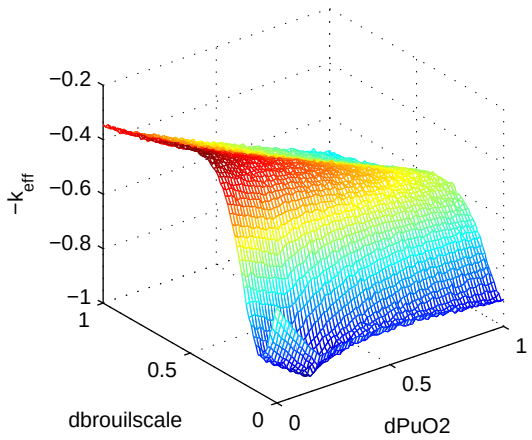

Figure 6: Accurate evaluation of k_{eff} (based on 200 time steps for each point) as a function of the design variables.

The k_{eff} range is approximately $[0.3, 1.0]$, so with one time step, the measurement 95% confidence interval length is $4 \times 0.0523 = 0.209$, which is about 30% of the response range. With 200 time steps, this length is 2% of the range. We consider two computational budgets here: $T_0 = 30$ and $T_0 = 100$, which correspond to a single observation with standard deviation respectively equal to 7.3×10^{-3} and 5.7×10^{-3}. Both can be considered as very small budgets regarding the problem complexity. The initial DoEs consist of respectively 6 and 20-point random designs (with optimized *maximin* distance), with one time step used for each measurement (so respectively 17 and 20% of the budget is allocated to the initial DoE).

The kriging fit is made using the R package *DiceKriging* (Roustant et al. 2009). The chosen model has a constant trend (ordinary kriging) and Matern 5/2 anisotropic covariance function. The covariance parameters are re-evaluated after each new observation.

The optimization results are given in tables 5 and 6. Figures 7 and 8 show the final kriging after optimization (mean, standard deviation and 90^{th} quantile); figure 9 represents the final designs of experiments along with contour lines of the actual response.

For the budget $T_0 = 30$, during optimization ten measurements have been added with time steps varying

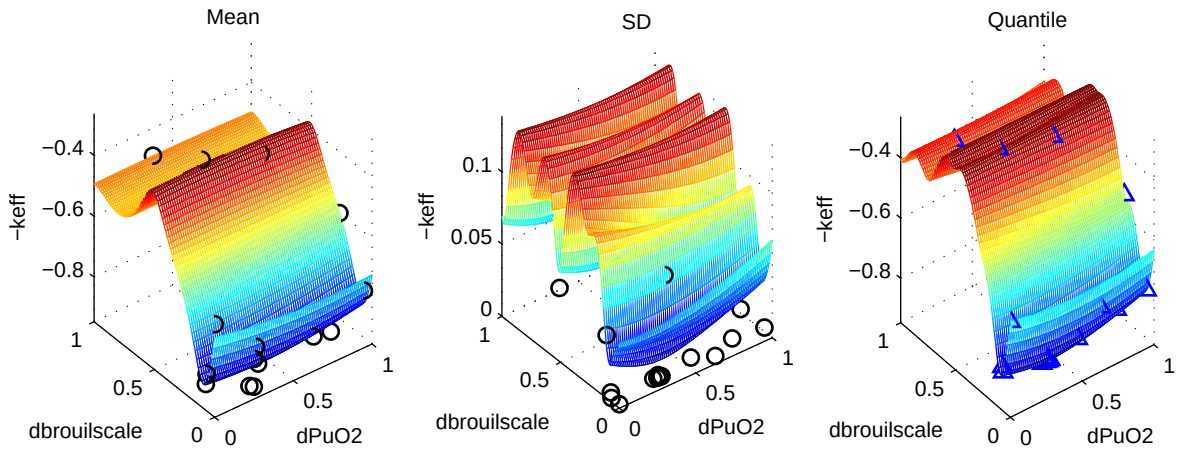

Figure 7: Final kriging after optimization for a $T = 30$ budget.

from one to six. The best design $\mathbf{x} = [0.3378, 0.0811]$ (based on the kriging quantile) has a kriging standard deviation of 0.0123 and is relatively close to the actual minimizer. From figure 9 (left), we see that five points were used for exploration (with only one time step) and the other five form a cluster of more accurate points in the optimal region. With such a low budget, the resulting kriging model (figure 7) is quite imprecise since the best predictor differs substantially in shape from the actual function and the kriging standard deviation is high over most of the design region. However, the points added during optimization reduce the error locally which results in a low standard deviation at the best design.

For the budget $T_0 = 100$, during optimization 14 measurements have been added with time steps varying from one to 36. The best design $\mathbf{x} = [0.1892, 0.0676]$ has a kriging standard deviation of 0.0071 and is almost equal to the actual minimizer. Here, approximately one third of the computational budget is allocated to the best design itself. The final kriging (figure 8) is more accurate than in the previous case, even though the standard deviation remains high in all the regions with high response values. In the region of the optimum, the kriging quantile is almost similar to the actual function.

Table 5: Sequential measurements for a $T = 30$ budget. (bold: best value)

Iteration	$\mathbf{x}$	Time allocated	$\widetilde{y}$	Kriging SD	Kriging quantile	Distance to actual optimum
0	[0.3108, 0.5135]	1	-0.3360	0.0515	-0.2920	0.4492
0	[0.6486, 0.0270]	1	-0.8462	0.0380	-0.7453	0.4626
0	[1 , 0.2703]	1	-0.6010	0.0513	-0.5342	0.8326
0	[0.8649, 0.7432]	1	-0.5217	0.0515	-0.4558	0.9460
0	[0.0270, 0.1081]	1	-0.8783	0.0335	-0.8193	0.1644
0	[0.3649, 0.9865]	1	-0.4846	0.0516	-0.4250	0.9223
1	[0 , 0.0676]	1	-0.8263	0.0279	-0.8180	0.1897
2	[0.2838, 0.0811]	2	-0.9382	0.0127	-0.8615	0.0946
3	[0.8108, 0.0946]	1	-0.8900	0.0298	-0.8133	0.6218
4	[0.5405, 0.0946]	1	-0.8065	0.0198	-0.8459	0.3516
5	[0.3108, 0.0811]	4	**-0.9469**	0.0124	-0.8621	**0.1216**
6	[0.3243, 0.0811]	4	-0.8183	0.0123	-0.8623	0.1351
7	[0 , 0]	1	-0.6408	0.0452	-0.6290	0.2058
8	[1 , 0.0676]	1	-0.7926	0.0359	-0.7866	0.8109
9	[0.3243, 0.0946]	3	-0.8624	0.0160	-0.8568	0.1358
10	[0.3378, 0.0811]	**6**	-0.8769	**0.0123**	**-0.8623**	0.1486

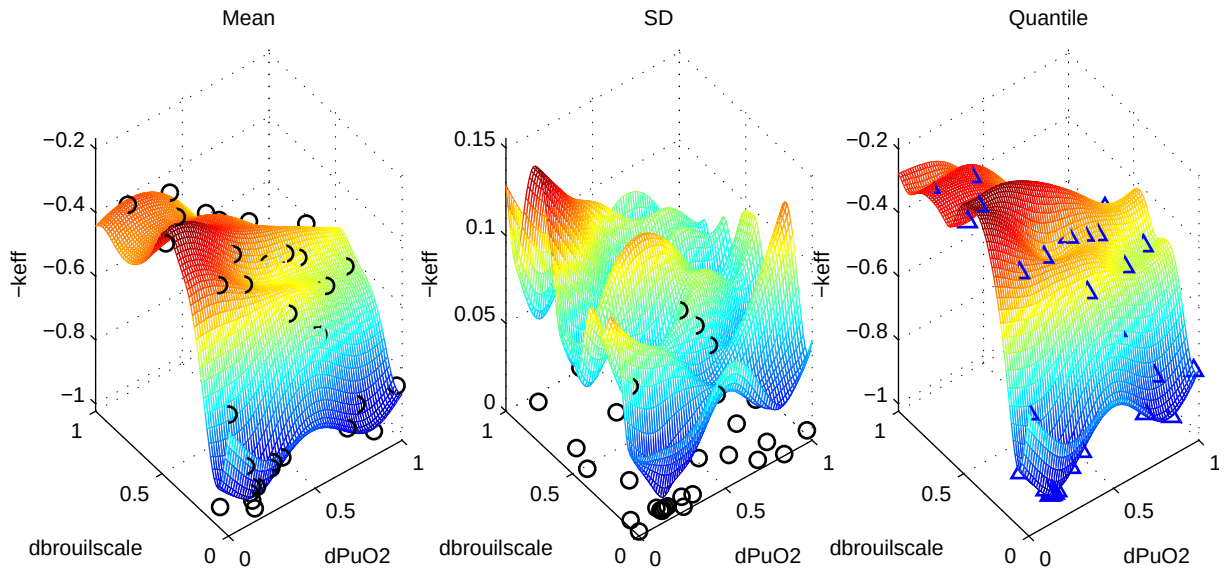

Figure 8: Final kriging after optimization for a $T = 100$ budget.

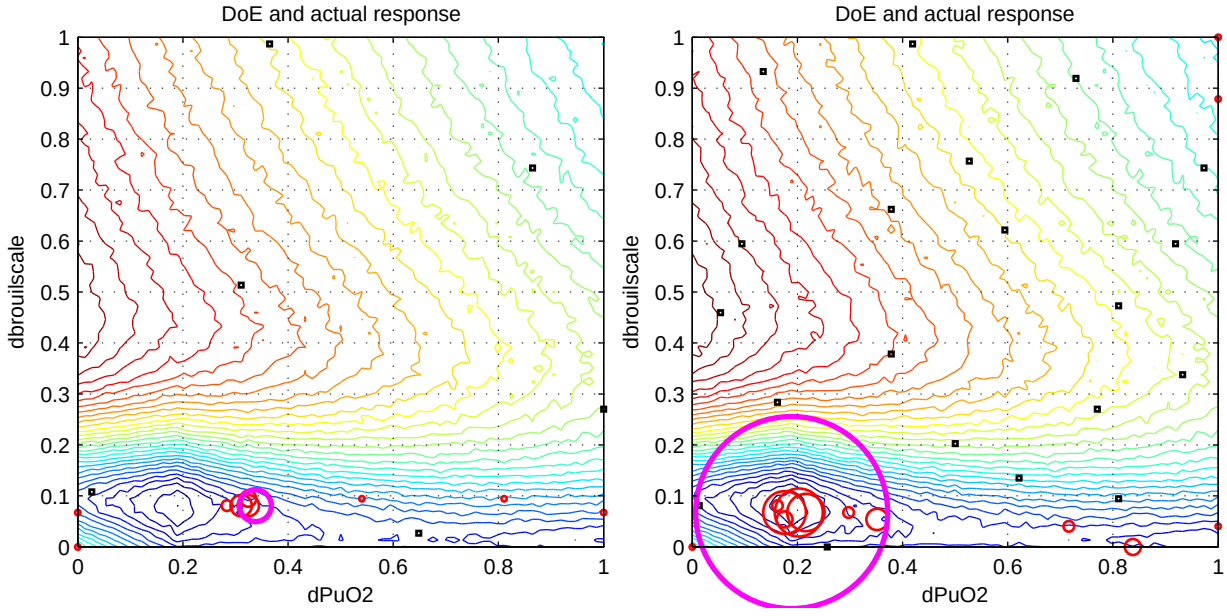

Figure 9: Final DoEs for the budgets $T = 30$ (left) and $T = 100$ (right) and contour lines of the actual function. Squares represent the inital measurements and circles the added ones; the markers are proportional to computational time.

Table 6: Sequential measurements for a $T = 100$ budget.

Iteration	$\mathbf{x}$	Time allocated	$\widetilde{y}$	Kriging SD	Kriging quantile	Distance to actual optimum
0	[0.8108, 0.4730]	1	-0.4716	0.0493	-0.4237	0.7348
0	[0.8108, 0.0946]	1	-0.8900	0.0423	-0.8013	0.6218
0	[0.4189, 0.9865]	1	-0.4578	0.0492	-0.4078	0.9341
0	[0.5946, 0.6216]	1	-0.4578	0.0471	-0.4029	0.6757
0	[0.0946, 0.5946]	1	-0.3653	0.0492	-0.2980	0.5222
0	[0.5270, 0.7568]	1	-0.4864	0.0488	-0.4277	0.7554
0	[0.7703, 0.2703]	1	-0.5750	0.0471	-0.5147	0.6111
0	[0.0541, 0.4595]	1	-0.2170	0.0496	-0.1818	0.4018
0	[0.3784, 0.3784]	1	-0.4948	0.0493	-0.4096	0.3524
0	[0.7297, 0.9189]	1	-0.7182	0.0481	-0.6464	0.9971
0	[0.5000, 0.2027]	1	-0.5544	0.0460	-0.4884	0.3338
0	[0.0135, 0.0811]	1	-0.9669	0.0410	-0.8600	0.1757
0	[0.9730, 0.7432]	1	-0.7334	0.0492	-0.6615	1.0260
0	[0.9189, 0.5946]	1	-0.6349	0.0486	-0.5605	0.8923
0	[0.1351, 0.9324]	1	-0.3888	0.0501	-0.3343	0.8531
0	[0.2568, 0]	1	-0.8683	0.0418	-0.8039	0.1055
0	[0.9324, 0.3378]	1	-0.5889	0.0484	-0.5247	0.7863
0	[0.6216, 0.1351]	1	-0.6300	0.0422	-0.6176	0.4358
0	[0.3784, 0.6622]	1	-0.3829	0.0465	-0.3337	0.6111
0	[0.1622, 0.2838]	1	-0.3948	0.0495	-0.3562	0.2045
1	[0, 0]	1	-0.6408	0.0460	-0.6388	0.2058
2	[0.2162, 0.0676]	7	-0.9603	0.0086	-0.9722	0.0302
3	[0.2973, 0.0676]	2	-0.9260	0.0184	-0.9234	0.1089
4	[0.1622, 0.0811]	2	-0.8837	0.0159	-0.9579	0.0270
5	[0.7162, 0.0405]	2	-0.9159	0.0324	-0.8696	0.5286
6	[0.3514, 0.0541]	4	-0.9027	0.0231	-0.8837	0.1644
7	[1, 0.0405]	1	-0.8678	0.0465	-0.8196	0.8118
8	[0.2027, 0.0676]	9	-1.0230	0.0074	-0.9754	0.0191
9	[0.1757, 0.0541]	3	-0.9426	0.0142	-0.9543	0.0302
10	[1, 1]	1	-0.8610	0.0499	-0.7778	1.2255
11	[1, 0.8784]	1	-0.7817	0.0480	-0.7240	1.1371
12	[0.1892, 0.0676]	**36**	-0.9949	**0.0071**	**-0.9760**	**0.0135**
13	[0.1757, 0.0676]	8	-0.9674	0.0080	-0.9735	0.0191
14	[0.8378, 0]	3	-0.9456	0.0298	-0.9000	0.6537

8. CONCLUSION AND PERSPECTIVES

In this paper, we have proposed a quantile-based expected improvement for the optimization of noisy black-box simulators. This criterion allows an elegant treatment of heterogeneous noise and takes into account the noise level of the candidate measurements. In the context of simulators with tunable fidelity, we proposed an on-line procedure for an adapted distribution of the computational effort. One of the advantages of such procedure is that it prevents from allocating too much time to poor designs, and allows spending more credit on the best ones. Another remarkable property of this algorithm is that, unlike EGO, it takes into account the limited computational budget. Indeed, the algorithm is more exploratory when there is much budget left, and favors a more local search when running out of computational credit. The online allocation optimization algorithm was first compared to existing methods on two analytical benchmark functions and was found to be very competitive. Finally, it was applied to an original application in nuclear criticality safety, the Monte Carlo criticality simulator MORET5. The algorithm showed promising results, using coarse measurements for exploration and accurate measurements at best designs. Future work may include deeper comparison of the EQI to other criteria for point selection on an extended benchmark of test functions, analysis of the effect of on-line allocation compared to a uniform allocation strategy, and an adaptation of the algorithm in the case of correlated errors.

Acknowledgements

This work was partially supported by French National Research Agency (ANR) through COSINUS program (project OMD2 ANR-08-COSI-007).

References

Ackley, D. (1987), *A connectionist machine for genetic hillclimbing* Kluwer Boston Inc., Hingham, MA.

Alexandrov, N., Lewis, R., Gumbert, C., Green, L., and Newman, P. (2000), "Optimization with variable-fidelity models applied to wing design," *AIAA paper*, 841(2000), 254.

Dixon, L., and Szego, G. (1978), *Towards Global Optimisation 2* North-Holland.

Fedorov, V., and Hackl, P. (1997), *Model-oriented design of experiments* Springer.

Fernex, F., Heulers, L., Jacquet, O., Miss, J., and Richet, Y. (2005), The MORET 4B Monte Carlo code - New features to treat complex criticality systems,, in *M&C International Conference on Mathematics and Computation Supercomputing, Reactor Physics and Nuclear and Biological Application, Avignon, France.*

Forrester, A., Bressloff, N., and Keane, A. (2006), "Optimization using surrogate models and partially converged computational fluid dynamics simulations," *Proceedings of the Royal Society A*, 462(2071), 2177.

Forrester, A., and Jones, D. (2008), Global optimization of deceptive functions with sparse sampling,, in *12th AIAA/ISSMO multidisciplinary analysis and optimization conference, Victoria, British Columbia, Canada*, pp. 10–12.

Forrester, A., Keane, A., and Bressloff, N. (2006), "Design and Analysis of" Noisy" Computer Experiments," *AIAA journal*, 44(10), 2331.

Gablonsky, J., and Kelley, C. (2001), "A locally-biased form of the DIRECT algorithm," *Journal of Global Optimization*, 21(1), 27–37.

Gano, S., Renaud, J., Martin, J., and Simpson, T. (2006), "Update strategies for kriging models used in variable fidelity optimization," *Structural and Multidisciplinary Optimization*, 32(4), 287–298.

Ginsbourger, D. (2009), Métamodèles Multiples pour l'Approximation et l'Optimisation de Fonctions Numériques Multivariables, PhD thesis, Ecole Nationale Supérieure des Mines de Saint Etienne.

Ginsbourger, D., and Le Riche, R. (2010), "Towards GP-based optimization with finite time horizon," in *mODa 9 Advances in Model-Oriented Design and Analysis*, eds. A. Giovagnoli, A. C. Atkinson, B. Torsney, and C. May, Contributions to Statistics Physica-Verlag HD, pp. 89–96.

Ginsbourger, D., Picheny, V., Roustant, O., and Richet, Y. (2008), Kriging with Heterogeneous Nugget Effect for the Approximation of Noisy Simulators with Tunable Fidelity,, in *Congrès conjoint de la Société Statistique du Canada et de la SFdS, May 25th-29th, Ottawa (Canada)*.

Gramacy, R., and Lee, H. (2010), "Optimization under unknown constraints," *Arxiv preprint arXiv:1004.4027*, .

Gramacy, R., and Polson, N. (2011), "Particle learning of Gaussian process models for sequential design and optimization," *Journal of Computational and Graphical Statistics*, 20(1), 102–118.

Huang, D., Allen, T., Notz, W., and Miller, R. (2006), "Sequential kriging optimization using multiple-fidelity evaluations," *Structural and Multidisciplinary Optimization*, 32(5), 369–382.

Huang, D., Allen, T., Notz, W., and Zeng, N. (2006), "Global optimization of stochastic black-box systems via sequential kriging meta-models," *Journal of Global Optimization*, 34(3), 441–466.

Humphrey, D., and Wilson, J. (2000), "A revised simplex search procedure for stochastic simulation response surface optimization," *INFORMS Journal on Computing*, 12(4), 272–283.

Jones, D. (2001), "A taxonomy of global optimization methods based on response surfaces," *Journal of Global Optimization*, 21(4), 345–383.

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

Kennedy, M., and O'Hagan, A. (2000), "Predicting the output from a complex computer code when fast approximations are available," *Biometrika*, 87(1), 1.

Krige, D. (1951), "A Statistical Approach to Some Mine Valuation and Allied Problems on the Witwatersrand: By DG Krige,".

Marrel, A. (2008), Mise en oeuvre et utilisation du metamodele processus gaussien pour l'analyse de sensibilite de modeles numeriques: Application a un code de transport hydrogeologique, PhD thesis, INSA Toulouse.

Matheron, G. (1969), "Le krigeage universel," *Cahiers du centre de morphologie mathématique*, 1.

Mockus, J. (1988), *Bayesian Approach to Global Optimization* Kluwer academic publishers.

Qian, P., and Wu, C. (2008), "Bayesian hierarchical modeling for integrating low-accuracy and high-accuracy experiments," *Technometrics*, 50(2), 192–204.

Rasmussen, C., and Williams, C. (2006), *Gaussian processes for machine learning* Springer.

Roustant, O., Ginsbourger, D., and Deville, Y. (2009), "The DiceKriging package: kriging-based metamodeling and optimization for computer experiments," *Book of abstract of the R User Conference*, .

Santner, T., Williams, B., and Notz, W. (2003), *The design and analysis of computer experiments* Springer Verlag.

Schonlau, M., Welch, W., and Jones, D. (1998), "Global versus local search in constrained optimization of computer models," *Lecture Notes-Monograph Series*, pp. 11–25.

Vazquez, E., Villemonteix, J., Sidorkiewicz, M., and Walter, É. (2008), Global optimization based on noisy evaluations: an empirical study of two statistical approaches,, in *Journal of Physics: Conference Series*, Vol. 135, p. 012100.

APPENDIX A: ONE-STEP AHEAD CONDITIONAL DISTRIBUTIONS OF THE MEAN, VARIANCE AND QUANTILE PROCESSES

Let $\mathbf{x}^{n+1}$ be the point to be visited at the $(n+1)^{\text{th}}$ step, τ^2_{n+1} and $\widetilde{Y}_{n+1} = Y(\mathbf{x}^{n+1}) + \varepsilon_{n+1}$ the corresponding noise variance and noisy response, respectively. We will now discuss the properties of the Kriging mean and variance at step $n+1$ seen from step n. Let $M_{n+1}(\mathbf{x}) := \mathbb{E}[Y(\mathbf{x})|\widetilde{A_n}, \widetilde{Y}_{n+1}]$ be the kriging mean function at the $(n+1)^{\text{th}}$ step and $S^2_{n+1}(\mathbf{x}) := Var[Y(\mathbf{x})|\widetilde{A_n}, \widetilde{Y}_{n+1}]$ the corresponding conditional variance. Seen from step n, both of them are *ex ante* random processes since they are depending on the not yet observed measurement $\widetilde{Y}_{n+1}$. We will now prove that they are in fact Gaussian Processes $|\widetilde{A_n}$, as well as the associated quantile $Q_{n+1}(\mathbf{x}) = M_{n+1}(\mathbf{x}) + \Phi^{-1}(\beta)S_{n+1}(\mathbf{x})$. The key results are that the Kriging predictor is linear in the observations, and that the Kriging variance is independent of them, as can be seen from Eqs. 5 and 5. Writing

$$M_{n+1}(\mathbf{x}) = \left(\sum_{j=1}^{n} \lambda_{n+1,j}(\mathbf{x})\widetilde{Y}_j\right) + \lambda_{n+1,n+1}(\mathbf{x})(Y(\mathbf{x}^{n+1}) + \varepsilon_{n+1}), \text{ where} \tag{23}$$

$$(\lambda_{n+1,.}(\mathbf{x})) := \left(\mathbf{k}_{n+1}(\mathbf{x})^T + \frac{(1 - \mathbf{k}_{n+1}(\mathbf{x})^T(K_{n+1} + \Delta_{n+1})^{-1}\mathbf{1}_{n+1})}{\mathbf{1}^T_{n+1}(K_{n+1} + \Delta_{n+1})^{-1}\mathbf{1}_{n+1}}\mathbf{1}^T_{n+1}\right)(K_{n+1} + \Delta_{n+1})^{-1}, \tag{24}$$

it appears that M_{n+1} is a GP $|\widetilde{A_n}$, with the following conditional mean and covariance kernel:

$$\mathbb{E}[M_{n+1}(\mathbf{x})|\widetilde{A_n}] = \sum_{j=1}^{n} \lambda_{n+1,j}(\mathbf{x})\tilde{y_i} + \lambda_{n+1,n+1}(\mathbf{x})m_n(\mathbf{x}) \text{ and} \tag{25}$$

$$\text{Cov}[M_{n+1}(\mathbf{x}), M_{n+1}(\mathbf{x}')|\widetilde{A_n}] = \lambda_{n+1,n+1}(\mathbf{x})\lambda_{n+1,n+1}(\mathbf{x}')(s^2_n(\mathbf{x}^{n+1}) + \tau^2_{n+1}). \tag{26}$$

Using that $Q_{n+1}(\mathbf{x}) = M_{n+1}(\mathbf{x}) + \Phi^{-1}(\beta)S_{n+1}(\mathbf{x})$, we observe that seen from the n^{th} step, $Q_{n+1}(.)$ is a GP as sum of a GP and a deterministic process conditional on $\widetilde{A_n}$. We finally get:

$$\mathbb{E}[Q_{n+1}(\mathbf{x})|\widetilde{A_n}] = \sum_{j=1}^{n} \lambda_{n+1,j}(\mathbf{x})\tilde{y_i} + \lambda_{n+1,n+1}(\mathbf{x})m_n(\mathbf{x}) + \Phi^{-1}(\beta)s_{n+1}(\mathbf{x}), \tag{27}$$

$$Cov[Q_{n+1}(\mathbf{x}), Q_{n+1}(\mathbf{x}')|\widetilde{A_n}] = \lambda_{n+1,n+1}(\mathbf{x})\lambda_{n+1,n+1}(\mathbf{x}')\left(s^2_n(\mathbf{x}^{n+1}) + \tau^2_{n+1}\right), \tag{28}$$

and the values used in the quantile Expected Improvement (equation 15) are:

$$m_{Q_{n+1}} = \mathbb{E}[Q_{n+1}(\mathbf{x}^{n+1})|\widetilde{A_n}] \tag{29}$$

$$s^2_{Q_{n+1}} = Var\left[Q_{n+1}(\mathbf{x}^{n+1})|\widetilde{A_n}\right] = \left(\lambda_{n+1,n+1}(\mathbf{x}^{n+1})\right)^2\left(s^2_n(\mathbf{x}^{n+1}) + \tau^2_{n+1}\right) \tag{30}$$

APPENDIX B: DERIVATION OF THE EQUIVALENT MEASUREMENT FOR TWO NOISY OBSERVATIONS AT THE SAME POINT

Let $\tilde{Y}_1 = \tilde{Y}_{\mathbf{x}_0} + \varepsilon_1$ and $\tilde{Y}_2 = \tilde{Y}_{\mathbf{x}_0} + \varepsilon_2$ be two measurements at $\mathbf{x}_0$, where ε_1 and ε_2 are two independent centered Gaussian variables independent of Y with respective variances τ_1^2 and τ_2^2. We will now show that conditioning the process Y on $\tilde{Y}_1$ and $\tilde{Y}_2$ is equivalent to conditioning it on a so-called equivalent measurement

$$\widetilde{Y}_{eq} = \frac{\tau_1^{-2}\widetilde{Y}_1 + \tau_2^{-2}\widetilde{Y}_2}{\tau_1^{-2} + \tau_2^{-2}} \tag{31}$$

It is straightforward that the equivalent observation is of the form $\tilde{Y}_{eq} = \tilde{Y}_{\mathbf{x}_0} + \varepsilon_{eq}$, where ε_{eq} is a centered gaussian too with variance $\tau_{eq}^2 = \frac{\tau_1^2\tau_2^2}{\tau_1^2+\tau_2^2}$.

Let us first assume that $(Y_x)_{x\in D}$ is a gaussian process with known mean $\mu(\mathbf{x})$ and covariance kernel $c(\mathbf{x}, \mathbf{x}')$ (both non necessarily stationary), and that a single observation of Y at $\mathbf{x}_0$ was observed with noise variance τ_{eq}^2. Classical conditioning (or simple kriging) equations for gaussian vectors give:

$$E\left[Y(\mathbf{x})|\tilde{Y}_{eq}\right] = \mu(\mathbf{x}) + \frac{c(\mathbf{x}_0,\mathbf{x})}{c(\mathbf{x}_0,\mathbf{x}_0)+\tau_{eq}^2}(\tilde{Y}_{eq} - \mu(\mathbf{x})) \tag{32}$$

$$var\left[Y(\mathbf{x})|\tilde{Y}_{eq}\right] = c(\mathbf{x},\mathbf{x}) - \frac{c(\mathbf{x}_0,\mathbf{x})^2}{c(\mathbf{x}_0,\mathbf{x}_0)+\tau_{eq}^2} \tag{33}$$

Alternatively, in the case of two observations, we have, following the kriging equations:

$$\begin{aligned} E\left[Y(\mathbf{x})|\tilde{Y}_1,\tilde{Y}_2\right] &= \mu(\mathbf{x}) + [c(\mathbf{x},\mathbf{x}_0) \quad c(\mathbf{x},\mathbf{x}_0)] \begin{bmatrix} c(\mathbf{x}_0,\mathbf{x}_0)+\tau_1^2 & c(\mathbf{x}_0,\mathbf{x}_0) \\ c(\mathbf{x}_0,\mathbf{x}_0) & c(\mathbf{x}_0,\mathbf{x}_0)+\tau_2^2 \end{bmatrix}^{-1} \begin{bmatrix} \tilde{Y}_1 - \mu(\mathbf{x}) \\ \tilde{Y}_2 - \mu(\mathbf{x}) \end{bmatrix} \\ &= \mu(\mathbf{x}) + \frac{c(\mathbf{x},\mathbf{x}_0)}{\tau_1^2\tau_2^2+\tau_1^2+\tau_2^2} [1 \quad 1] \begin{bmatrix} c(\mathbf{x}_0,\mathbf{x}_0)+\tau_2^2 & -c(\mathbf{x}_0,\mathbf{x}_0) \\ -c(\mathbf{x}_0,\mathbf{x}_0) & c(\mathbf{x}_0,\mathbf{x}_0)+\tau_1^2 \end{bmatrix} \begin{bmatrix} \tilde{Y}_1 - \mu(\mathbf{x}) \\ \tilde{Y}_2 - \mu(\mathbf{x}) \end{bmatrix} \end{aligned}$$

which after simplifications, gives:

$$E\left[Y(\mathbf{x})|\tilde{Y}_1,\tilde{Y}_2\right] = \frac{c(\mathbf{x}_0,\mathbf{x})}{c(\mathbf{x}_0,\mathbf{x}_0)+\tau_{eq}^2}\left(\tilde{Y}_{eq} - \mu(\mathbf{x})\right) = E\left[Y(\mathbf{x})|\tilde{Y}_{eq}\right] \tag{34}$$

Similarly, for the kriging variance:

$$var\left[Y(\mathbf{x})|\tilde{Y}_1,\tilde{Y}_2\right] = c(\mathbf{x},\mathbf{x}) - [c(\mathbf{x},\mathbf{x}_0) \quad c(\mathbf{x},\mathbf{x}_0)] \begin{bmatrix} c(\mathbf{x}_0,\mathbf{x}_0)+\tau_1^2 & c(\mathbf{x}_0,\mathbf{x}) \\ c(\mathbf{x}_0,\mathbf{x}_0) & c(\mathbf{x}_0,\mathbf{x}_0)+\tau_2^2 \end{bmatrix}^{-1} \begin{bmatrix} c(\mathbf{x},\mathbf{x}_0) \\ c(\mathbf{x},\mathbf{x}_0) \end{bmatrix}$$

which returns, after simplification:

$$var\left[Y(\mathbf{x})|\tilde{Y}_1,\tilde{Y}_2\right] = c(\mathbf{x},\mathbf{x}) - \frac{c(\mathbf{x}_0,\mathbf{x})^2}{c(\mathbf{x}_0,\mathbf{x}_0)+\tau_{eq}^2} = var\left[Y(\mathbf{x})|\tilde{Y}_{eq}\right] \tag{35}$$

Conditioning the process Y on the two noisy measurements is hence equivalent to conditioning on the equivalent measurement since such a Gaussian conditional distribution is entirely defined by its two first moments. Now, showing that the equivalence holds in cases where Y was already observed at a design $\mathbf{X}^n$ previous to the two noisy measurements is just a question of applying the previous property conditionally on $Y(\mathbf{X}^n)$. Replacing $\mu(\mathbf{x})$ by $E\left[Y(\mathbf{x})|\widetilde{A}_n)\right] = m_n(\mathbf{x})$ and $c(\mathbf{x},\mathbf{x}')$ by $cov\left[Y(\mathbf{x}),Y(\mathbf{x}')|\widetilde{A}_n)\right] = k(\mathbf{x},\mathbf{x}') - \mathbf{k}_n(\mathbf{x})^T(K_n+\Delta_n)^{-1}\mathbf{k}_n(\mathbf{x}') + \frac{\left(1-\mathbf{k}_n(\mathbf{x})^T(K_n+\Delta_n)^{-1}\mathbf{k}_n(\mathbf{x}')\right)}{\mathbf{1}_n^T(K_n+\Delta_n)^{-1}\mathbf{1}_n}$ delivers the desired property.