2nd International Conference on Engineering Optimization
September 6-9, 2010, Lisbon, Portugal

Noisy Expected Improvement and On-line Computation Time Allocation for the Optimization of Simulators with Tunable Fidelity

V. Picheny[1], D. Ginsbourger[2], Y. Richet[3]

[1]Ecole Centrale de Paris, Chatenay-Malabry, France, victor.picheny@ecp.fr
[2]Department of Mathematics and Statistics, University of Bern, Switzerland, david.ginsbourger@stat.unibe.ch
[3]Institut de Radioprotection et de Sûreté Nucléaire, Fontenay-aux-Roses, France, yann.richet@irsn.fr

Abstract
This article addresses the issue of kriging-based optimization of stochastic simulators. Many of these simulators depends 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 a rigorous treatment of heterogeneous response precisions. Secondly, we present a procedure that allocates 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 [10] and its underlying expected improvement (EI) criterion have been recognized as efficient tools for deterministic black-box optimization. However, most simulators return approximate solutions to the considered mathematical model rather than exact ones. 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 meshing density or element order. Another example is when the response stems from Monte Carlo methods: the accuracy (measured by response variance) is inversely proportional to sample size.
Such simulators are often referred to as *noisy simulators*, since they return approximate solutions that depart from the exact value by an error term. 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. The choice of this trade-off greatly impacts the efficiency of the optimization.
Using metamodels for noisy optimization has been already addressed by several authors. Many approaches consider only two fidelity levels, and the low-fidelity model is used as a helping tool to choose the high-fidelity evaluations [1, 4]. In that case, metamodels are used to estimate the difference between simulators. More integrated approaches have also been proposed, based on modifications of the EGO algorithm. Huang et al. [7] 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. Forrester et al. [3] proposed a re-interpolation technique to filter out the noise, allowing the use of the standard EGO algorithm.
This work proposes two contributions to this framework. First, we define an extension of EI based on quantiles that enables a rigorous treatment of continuous 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 describe 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. Finally, an original application in nuclear criticality safety is implemented in the Promethee workbench and applied to the Monte Carlo criticality simulator MORET5 [2].

2. Notations and concepts

We consider a single objective, unconstrained optimization problem. 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 doesn't 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.

Simulation noise can have many sources, as detailed in [3], including finite sample size (Monte-Carlo methods), and discretization error or incomplete convergence (Finite Element methods). In our context of simulations with tunable fidelity, we consider furthermore the case where for every measurement i ($1 \leq i \leq n$), the noise variance τ_i^2 is controllable and decreases monotonically with the allocated computational time t_i. 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 simulated and actual phenomena is not considered here. Following the particular case of Monte Carlo simulations, we finally make the assumption that $\varepsilon_i \sim \mathcal{N}(0, \tau_i^2)$ independently.

Kriging is a functional approximation method originally coming from geosciences, and having been popularized in machine learning (Gaussian Process paradigm [14]) and in numerous application fields. 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 unkwown 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) [11]. Additionally, assuming further that $Y|\mu$ is a Gaussian Process (GP) and μ has an improper uniform distribution over $\mathbb{R}$ leads to the convenient result that OK amounts to conditioning Y on the measurements, thus 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 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 [6], 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:

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

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

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 terms $\tau_1^2 \ldots \tau_n^2$, $\mathbf{1}_n$ is a $n \times 1$ vector of ones, and $\widehat{\mu}_n = \frac{\mathbf{1}_n^T (K_n + \Delta_n)^{-1} \widetilde{\mathbf{y}}^n}{\mathbf{1}_n^T (K_n + \Delta_n)^{-1} \mathbf{1}_n}$ is the best linear unbiased estimator 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 occurence. Specific properties of this generalization of OK include that $m_n(.)$ is not interpolating noisy measurements, that $s_n^2(.)$ doesn't 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.

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 [10]. 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 [9], 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 $\min(y(\mathbf{X}^n))$, 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:

$$EI_n(\mathbf{x}) := \mathbb{E}\left[\left(\min(Y(\mathbf{X}^n)) - Y(\mathbf{x})\right)^+ | A_n\right] = \mathbb{E}\left[\left(\min(\mathbf{y}^n) - Y(\mathbf{x})\right)^+ | A_n\right] \tag{3}$$

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

$$EI_n(\mathbf{x}) := \left(\min(\mathbf{y}^n) - m_n(\mathbf{x})\right) \Phi\left(\frac{\min(\mathbf{y}^n) - m_n(\mathbf{x})}{s_n(\mathbf{x})}\right) + s_n(\mathbf{x})\phi\left(\frac{\min(\mathbf{y}^n) - m_n(\mathbf{x})}{s_n(\mathbf{x})}\right), \tag{4}$$

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 used in particular in the DiceOptim package [15] 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:

$$EI_n(\mathbf{x}) = \mathbb{E}\left[\left(\underbrace{\min(Y(\mathbf{X}^n))}_{\text{unknown}} - \underbrace{Y(\mathbf{x})}_{\text{unreachable}}\right)^+ \middle| \widetilde{A_n}\right], \tag{5}$$

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. This prevents one from getting an analytical formula for EI. 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.

One temptation in order to keep a tractable criterion could be to plug in $\min(\widetilde{\mathbf{y}^n})$ in EI instead of the unkown $\min(\mathbf{y}^n)$. However, it could be greatly misleading since the noisy minimum is a biased estimate of the noiseless minimum, and it sufficies to have one highly noisy observation with a low value to deeply underestimate $\min(\mathbf{y}^n)$. A rule of thumb proposed by Vazquez et al. [16] 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, Forrester et al. [3] proposed to replace the noisy observations by the kriging best predictor $m_n(\mathbf{X}^n)$, and fit a noise-free kriging on such data, which is used for the standard EGO algorithm. A more rigorous alternative consists of estimating the EI based on Monte-Carlo simulations involving the joint distribution of $(\min(Y(\mathbf{X}^n), Y(\mathbf{x}))$ conditional on $\widetilde{A_n}$; however, such estimates are noisy and numerically costly, which makes the EI maximization challenging. The *Augmented Expected Improvement* (AEI) is another variant of EI for noisy optimization [8], with $\min(Y(\mathbf{X}^n))$ being replaced by the best predictor at the training point with smallest kriging quantile, and the EI being multiplicated by a penalization function to account for the diminishing return of additional replicates. Huang et al. generalized their AEI [7] to the case of a finitely many simulators with heterogeneous fidelity, and proposed a criterion that depends on both the next point $\mathbf{x}$ and its associated response precision. This is one of the fundamental properties of the EI criterion with tunable fidelity proposed in the next section, which additionally differs from AEI by its transparent probabilistic fundations.

4. Quantile-based EI

We now introduce a variant of EI for the case of a deterministic objective function with heterogeneously noisy measurements. 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?

Here we chose 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" whenever it has the lowest quantile, which defines a natural and tunable trade-off between performance and reliability. Let us recall that under the GP assumptions above, the β-quantile surface at step k ($k \leq n+1$) can be explicitely derived based on m_n and s_n:

$$q_n(\mathbf{x}) := \inf\{u \in \mathbb{R} : \mathbb{P}(Y(\mathbf{x}) \leq u | \widetilde{A_n}) \geq \beta\} = m_n(\mathbf{x}) + \Phi^{-1}(\beta) s_n(\mathbf{x}) \tag{6}$$

As in [8], we restrict our choice here to the already tried points, thus forbidding to end the optimization process with a point at which no measurement has ever been done. Such a restriction greatly simplifies calculations, and is compatible with good practices: who would trust a metamodel so much to propose a final candidate minimizer without any measurement at that point? So we propose to define the improvement in terms of decrease of the lowest β-quantile at the available design of experiments, between the present and the forthcoming step. Of course, like in the classical EI case, this improvement cannot be known in advance, but can be predicted based on the GP model.

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

$$EI_n(\mathbf{x}^{n+1}, \tau^2_{n+1}) := \mathbb{E}\left[\left(\min_{i \leq n}(Q_n(\mathbf{x}^i)) - Q_{n+1}(\mathbf{x}^{n+1}) \right)^+ \middle| \widetilde{A_n} \right] \tag{7}$$

where the dependence on 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). As a result, the proposed quantile EI is analytically tractable, and we get by a similar calculation as in Eq. 4:

$$EI_n(\mathbf{x}^{n+1}, \tau^2_{n+1}) = \left(\min(\mathbf{q}^n) - m_{Q_{n+1}}\right) \Phi\left(\frac{\min(\mathbf{q}^n) - m_{Q_{n+1}}}{s_{Q_{n+1}}} \right) + s_{Q_{n+1}} \phi\left(\frac{\min(\mathbf{q}^n) - m_{Q_{n+1}}}{s_{Q_{n+1}}} \right) \tag{8}$$

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 proposed EI 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 EI maximization is done with respect to $\mathbf{x}$ only. An illustration of the dependence of the criterion on τ^2 is provided in section 7. β tunes the level of reliability wanted on the final result; with $\beta = 0.5$, the design points are compared based on the kriging mean predictor only, while high values of β (i.e. near to 1) penalize designs with high uncertainty, which is a more conservative approach.

A simple algorithm can be defined at this point, by choosing a unique value for τ^2 and making observations iteratively with uniform noise variances where the EI criterion is maximum. However, the choice of τ^2 is non-trivial, and such strategy hinders greatly the potential of tunable fidelity. The next section is dedicated to a method of dynamically choosing τ^2 during optimization.

5. On-line allocation of resource

For many "noisy" simulators such as those relying on Monte Carlo or on iterative solvers, the response corresponding to a given fidelity is not obtained directly but more as a limit of intermediate responses of lower fidelities. It is assumed here that the evolution of these responses can be monitored on-line, and that the noise variance is a known decreasing function of computation time

$$\tau^2 : t \in [0, +\infty[\longrightarrow \tau^2(t) \in [0, +\infty[\tag{9}$$

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 supposed to correspond to one elementary computation time $t_e \in]0, +\infty[$, so that $\forall j \in \{1, \ldots, b_i\},\ \tau_i^2[j] = \tau^2(j \times t_e)$.
We propose a procedure that, instead of fixing the computational resource (and thus, noise level) prior to the measurements, adapts it on-line by looking at the evolution of the response. The objective of such procedure is to obtain a better distribution of the computational resource. Indeed, it seems natural that more time should be allowed to designs close to the actual minimizer, and less to designs with high values of the cost function. However, knowing that a design is poor is also a valuable information, since it allows us to remove its neighbourhood from the searching region. In the following, we propose to use the quantile-based EI to measure the value of additional calculations.

5.1 On-line update of the quantile-based EI

Let us focus without loss of generality on the i^{th} measurement, made at $\mathbf{x}^i$, initially chosen with noise level $\tau_i^2[b_i] = \tau^2(T_i)$ where $T_i := b_i \times t_e$. After $j \leq b_i$ calculation steps, the current observation is $\widetilde{y}_i[j]$, with variance $\tau_i^2[j] = \tau^2(j \times t_e)$. We propose an update of the EI in order to measure the potential improvement if the measurement process is carried on until the noise level $\tau^2(T_i)$ is reached.
The EI criterion has first to take into account the current noisy measurement, which requires a straightforward update of the kriging model. If $j = 1$, the new point has to be added to the DoE and the kriging equations modified accordingly; otherwise, it only requires to replace the previous values of response and noise $\widetilde{y}_i[j-1]$ and $\tau_i^2[j-1]$ by $\widetilde{y}_i[j]$ and $\tau_i^2[j]$ in the kriging equations 1 and 2. The future noise level $\tau^2(T_i)$ also has to be modified; if not, the EI would estimate the value of a new measurement with variance $\tau^2(T_i)$, regardless of the fact that the measurement has already started.
To do so, 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 $\tau_{i,1}^2$ and $\tau_{i,2}^2$. They are equivalent to a single measurement $\widetilde{y}_{i,eq} = \left(\tau_{i,1}^{-2} + \tau_{i,2}^{-2}\right)^{-1} \left(\tau_{i,1}^{-2}\widetilde{y}_{i,1} + \tau_{i,2}^{-2}\widetilde{y}_{i,2}\right)$ with variance $\tau_{i,eq}^2$, given by:

$$\frac{1}{\tau_{i,eq}^2} := \frac{1}{\tau_{i,1}^2} + \frac{1}{\tau_{i,2}^2} \Longrightarrow \tau_{i,eq}^2 = \frac{\tau_{i,1}^2 \tau_{i,2}^2}{\tau_{i,1}^2 + \tau_{i,2}^2} \tag{10}$$

Now, to update the EI, we make the assumption that it is equivalent to carry the measurement process until the noise level $\tau_i^2[b_i]$ is reached, or to make a new measurement with noise variance:

$$\tau_i^2[j \to b_i] := \frac{\tau_i^2[j]\tau_i^2[b_i]}{\tau_i^2[j] - \tau_i^2[b_i]} = \frac{\tau^2(j \times t_e)\tau^2(T_i)}{\tau^2(j \times t_e) - \tau^2(T_i)} =: \tau^2(j \times t_e \to T_i) \tag{11}$$

Such hypothesis implies that the increment between the current and final values $\widetilde{y}_i[j]$ and $\widetilde{y}_i[b_i]$ is a realization of a variable following the law $\mathcal{N}(0, \tau_i^2[j \to b_i])$ independently of $\widetilde{Y}_i[j]$, which is exact in Monte Carlo, and may be acceptable in partially converged iterative solvers, provided that the frequence of sampling is low compared to the oscillations of convergence around the actual solution.
The updated EI indicates the expected quantile improvement *if the observation variance is reduced to* $\tau^2(T_i)$. Such quantity tends by construction to decrease when computation time is added, since (1) the kriging uncertainty reduces at the observation point and (2) it decreases when $\tau^2(j \times t_e \to T_i)$ increases. However, if the measurement converges to a good (small) value, EI can increase temporarily. Inversely, if the measurement converges to a high value, EI decreases faster. Hence, we can define a ("point switching") stopping criterion for resource allocation based on EI. If the EI 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 EI is less than a proportion of the initial EI value (that is, the value of EI when starting the measurement process at that point), for instance 50%.

5.2 Optimization with finite computational budget

In many practical applications, the total computational budget is bounded, and prescribed by industrial constraints such as time and power limitations. In the case of Monte Carlo-based simulators, this computational budget can be defined in terms of sample size. i.e. total number of drawings. Taking into account this limitation may modify the optimization strategy: depending on the total budget, the optimization process may be more or less exploratory, with more or less noisy measurements. Integrating

this limitation in the algorithm would result in improved efficiency and prevent the user from having to make manual trade-offs between accuracy and rapidity. Mokus [12] followed by Ginsbourger et al. [5] demonstrated the relevance of such approach for the deterministic case.
In fact, it is possible to take into account such computational boundedness in the proposed algorithm. Indeed, the future noise level τ^2_{n+1}, which is a parameter of the EI criterion, will stand here for the finite resource. Given a computational budget T_{n+1}, the smallest noise variance achievable for a new measurement is $\tau^2(T_{n+1})$, assuming that all the remaining budget will be attributed to this measurement. In the course of the optimization process steps $j \leq b_{n+1}$, the remaining budget decreases, so $\tau^2\left(j \times t_e \rightarrow T_{n+1}\right)$ increases. By setting $\tau^2_{n+1} = \tau^2(T_{n+1})$, the new experiment is chosen knowing that even if all the remaining budget is attributed to the same observation, the noise variance will not decrease below a certain value. Consequently, the EI will behave differently at the beginning and the end of the optimization. When the budget is high, EI will be highest in unexplored regions, since it is where accurate measurements are likely to be most efficient. At the end of the optimization, when the remaining time is small, the EI will be small in those regions since that 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 EI will be highest close to the current best point.

6. Algorithm overview

This section summarizes the optimization algorithm we propose, which features all the elements proposed in section 4 and 5. The principal of the following algorithm is to choose the point with highest quantile expected improvement given the whole remaining budget, to store the corresponding EI value as reference, and to invest new elementary measurement at this point until the EI with updated data falls under a given proportion $\gamma \in]0,1[$ of the reference EI value. The operation of choosing the most promizing point is then started again, and so on until the total computational budget has been spent. Note that the final number of measurements is not determined beforehand but adapts automatically to the budget and resource distribution. The complete algorithm is presented in pseudo-code form in table 1.
The total computational budget T needs to be defined before optimization, and discretized in incremental steps $\{t_e, 2 \times t_e, \ldots, b \times t_e\}$, where $b = \frac{T}{t_e}$. Smaller steps (i.e. a smaller t_e) result in increased precision, but requires more Kriging updates and EI maximization, which can become computationally intensive. A prescribed fraction T_0 of this budget is allocated to build an 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 30 to 50% of the total budget on a space-filling DoE (for instance, an LHS design) with uniform observation variances is a reasonable option.

Table 1: Quantile EI algorithm with on-line resource allocation

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

while $T_n > 0$
- Choose new design point $\mathbf{x}^{n+1}$ that maximizes $EI_n\left(., \tau^2(T_n)\right)$
- 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** $EI_{n+1}\left(\mathbf{x}^{n+1}, \tau^2(t \rightarrow T_n)\right) > \gamma EI_n\left(\mathbf{x}^{n+1}, \tau^2(T_n)\right)$
  - Generate $\widetilde{y}_{n+1}[j+1]$ by adding one time increment
  - Update Kriging model with: $\widetilde{y}_{n+1} = \widetilde{y}_{n+1}[j+1]$, $\tau^2_{n+1} = \tau^2(t)$
  - Set $T_{n+1} = T_{n+1} - t_e$, $j = j + 1$, and $t = t + t_e$
- **end while**
- Set $n = n + 1$

end while

- Choose final design based on the Kriging quantile

Here, the variance level depends only on computational time; however, the algorithm writes similarly if the variance is design-dependent, by replacing $\tau^2(t)$ by $\tau^2(t; \mathbf{x})$. Also, a minimum achievable noise can be set by the user or the simulator itself. 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 EI expressions above.

7. Experiments

Two examples are proposed in this section: a one-dimensional analytical example, for illustrative purpose, and a three-dimensional nuclear safety problem.

7.1 One-dimensional example

First, we study a one-dimensional problem, with objective function defined over $[0, 1]$ by:

$$y(x) = \frac{1}{2}\left(\frac{sin(20x)}{1+x} + 3x^3 cos(5x) + 10(x-0.5)^2 - 0.6\right) \tag{12}$$

The noise is here inversely proportional to computational time, and independent of $\mathbf{x}$: $\tau^2(t) = \frac{0.1}{t}$.
First, we represent the quantile EI criterion (with $\beta = 0.9$) for the initial DoE and kriging model, for three different τ^2 values : 1, 0.1 and 0.01. The initial DoE consists of four 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 true function, kriging model, and EI are shown on Figure 1. We can see that the choice of the future noise level has a great influence on the criterion. With small noise variance, the quantile-EI behaves like the classical EI, with highest values in regions with high uncertainty and low mean predictions. With higher noise variances, the criterion becomes very conservative since is it high only in the vicinity of existing measurements. Indeed, a very noisy future observation can only have a very limited influence on the kriging model, so adding such observation in an uncertain region is insufficient to lower enough the quantile to have a high quantile EI.

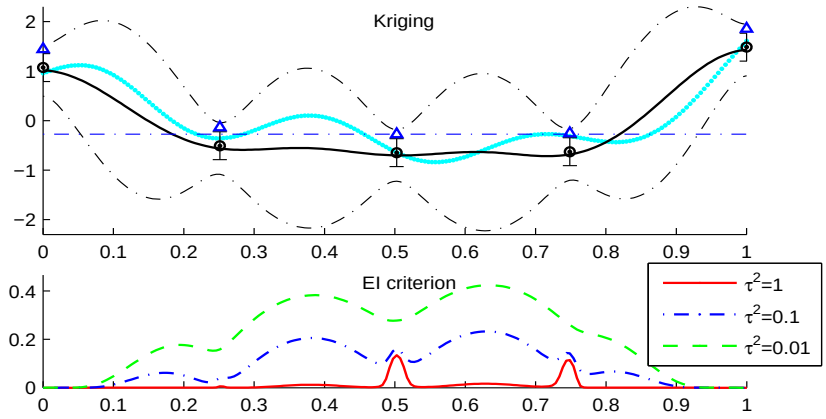

Figure 1: quantile-EI for three different future noise levels.

Then, an optimization is performed with a total computational budget of $T = 100$, starting from the DoE described above. T 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. Figure 2 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 Application to a 3D 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 a heterogeneous uranium sphere partially moderated by water. The safety of this system is measured by the criticality coefficient (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 criticality coefficient depends on the composition of fissile materials, and on operational parameters such as water (as a moderator for the kinetic energy of neutrons), structure materials, fissile geometry, and heterogeneity characteristics. For a given set of parameters, the value of k_{eff} can be evaluated using

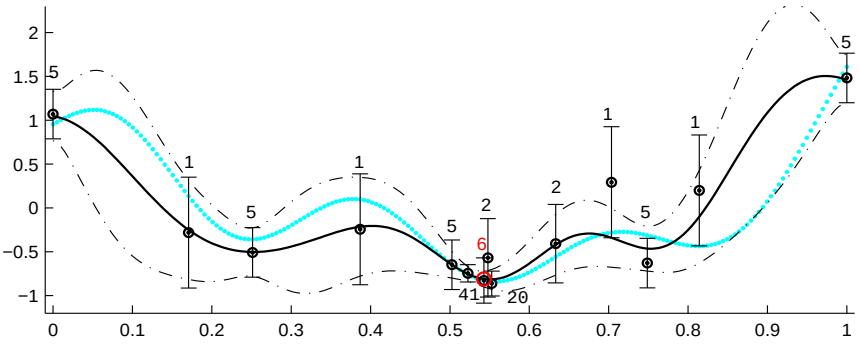

Figure 2: Observations and kriging after optimization. The best measurement point is circled in red.

the MORET stochastic simulator [2, 13], which is based on Markov Chain Monte-Carlo 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):

$$\max_{\mathbf{x}\in D} \quad k_{\text{eff}}(\mathbf{x}) \leq 1.0 - \text{margin} \tag{13}$$

If this relation is not ensured, then D is reduced and safety checked again. 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. In this article, we focus on the maximization of k_{eff} with respect to three parameters (the other possible inputs being fixed to their nominal values):

- m_{mf}, the whole fissile mass of the system, with range $[11, 20]$ kg,
- α, the ratio of moderated mass of fissile, with range $[0.05, 0.5]$ (after a preliminary rescaling)
- m_{mod}, the moderator mass, with range $[0.22, 1.3]$ kg.

Hence: $\mathbf{x} = (m_{\text{mf}}, \alpha, m_{\text{mod}})$, 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). The variance of the k_{eff} estimate is 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 $10 \times 10 \times 10$ grid, and for each new measurement the EI 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 1.69×10^{-3} (one time step) and 2.38×10^{-3} (fifty time steps). The k_{eff} range is approximately $[0.8, 1.0]$, so with one time step, the measurement 95% confidence interval length is $4 \times 0.0169 = 0.068$, which is 33% of the response range. With fifty time steps, the length is 5% of the range. The total computational budget considered here is $T = 200$, which corresponds to only four points of highest accuracy. The initial DoE consists of 50 points randomly chosen on the grid, with one time step used for each measurement (so 25% of the budget is allocated to the initial DoE).

After the optimization, 11 measurements have been added and one measurement of the initial DoE has been refined. Four new measurements used one time step, two used 50, the others used intermediate values. The best design (with lowest kriging quantile) has second lowest measurement, and used 50 time steps. Its kriging standard deviation is 1.6×10^{-3}, which is less than 1% of the response range.

Figure 3 presents the final sequence of measurements. The first 50 points are the inital DoE, and only the 28-th point presents a better accuracy due to further enrichment during optimization. The last 11 points are the new points added by the algorithm. The first three measurements added have a very high precision; however, the computational resource was not attributed to these points all at once, but the algorithm switched to other points and came back to enrich the measurement several times during the optimization. The five last measurements have very large noises and kriging quantiles; these designs correspond to exploration points, which measurements were stopped early when they were detected to be poor designs. On the other hand, the three best points (in terms of kriging quantile) are the ones where

most of the computational time have been allocated.

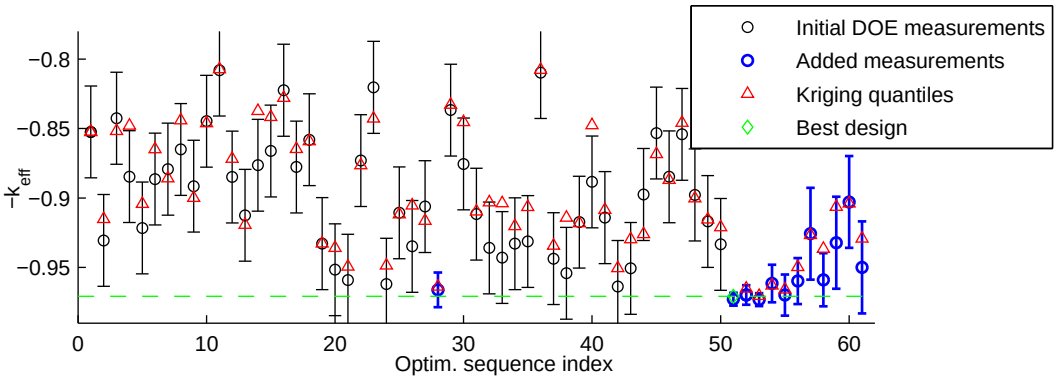

Figure 3: DoE after optimization. The vertical bars are the measurements confidence intervals $\tilde{y}_i \pm 2 \times \tau_i$.

8. Conclusion and perspectives

In this paper, we have proposed a quantile-based expected improvement for the optimization of noisy back-box simulators. This criterion allows a rigorous 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 favours a more local search when running out of computational credit. The online allocation optimization algorithm was tested on two problems: an analytical function, and an original application in nuclear criticality safety, the Monte Carlo criticality simulator MORET5. On both problems, the algorithm showed promising results, using coarse measurements for exploration and accurate measurements at best designs. Future work may include comparison of the quantile-based EI to other criteria for point selection, analysis of the effect of on-line allocation compared to a uniform allocation strategy, and a comparison of our algorithm to classical noisy optimization algorithms.

Acknowledgements

This work was partially supported by the ANR project OMD2 (Optimisation Multi-Disciplinaire Distribuée).

References

[1] N.M. Alexandrov, R.M. Lewis, CR Gumbert, LL Green, and PA Newman. Optimization with variable-fidelity models applied to wing design. AIAA paper, 841(2000):254, 2000.

[2] F. Fernex, L. Heulers, O. Jacquet, J. Miss, and Y. Richet. 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, 2005.

[3] A.I.J. Forrester, A.J. Keane, and N.W. Bressloff. Design and Analysis of" Noisy" Computer Experiments. AIAA journal, 44(10):2331, 2006.

[4] S.E. Gano, J.E. Renaud, J.D. Martin, and T.W. Simpson. Update strategies for kriging models used in variable fidelity optimization. Structural and Multidisciplinary Optimization, 32(4):287–298, 2006.

[5] D. Ginsbourger and R. Le Riche. Towards GP-based optimization with finite time horizon. http://hal.archives-ouvertes.fr/hal-00424309/en/, 2009.

[6] D. Ginsbourger, V. Picheny, O. Roustant, and Y. Richet. 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), 2008.

[7] D. Huang, T.T. Allen, W.I. Notz, and R.A. Miller. Sequential kriging optimization using multiple-fidelity evaluations. Structural and Multidisciplinary Optimization, 32(5):369–382, 2006.

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

[9] D.R. Jones. A taxonomy of global optimization methods based on response surfaces. Journal of Global Optimization, 21(4):345–383, 2001.

[10] D.R. Jones, M. Schonlau, and W.J. Welch. Efficient global optimization of expensive black-box functions. Journal of Global Optimization, 13(4):455–492, 1998.

[11] G. Matheron. Le krigeage universel. Cahiers du centre de morphologie mathématique, 1, 1969.

[12] J. Mockus. Bayesian Approach to Global Optimization. Kluwer academic publishers, 1988.

[13] Promethee project. IRSN, "Grid computing for numerical engineering". http://promethee.irsn.org.

[14] C.E. Rasmussen and C.K.I. Williams. Gaussian processes for machine learning. Springer, 2006.

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

[16] E. Vazquez, J. Villemonteix, M. Sidorkiewicz, and É. Walter. Global optimization based on noisy evaluations: an empirical study of two statistical approaches. In Journal of Physics: Conference Series, volume 135, page 012100, 2008.

A1. 1-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. 1 and 2. Writing

$$M_{n+1}(\mathbf{x}) = \left(\sum_{j=1}^{n} \lambda_{n+1,j}(\mathbf{x})\widetilde{Y}_j\right) + \lambda_{n+1,n+1}(\mathbf{x})(Y(\mathbf{x}^{n+1}) + \varepsilon_{n+1}), \text{ where} \tag{14}$$

$$(\lambda_{n+1,.}(\mathbf{x})) := \left(\mathbf{k}_{n+1}(\mathbf{x})^T + \frac{(1 - \mathbf{k}_{n+1}(\mathbf{x})^T(K_{n+1} + \Delta_{n+1})^{-1}\mathbf{1}_{n+1})}{\mathbf{1}^T_{n+1}(K_{n+1} + \Delta_{n+1})^{-1}\mathbf{1}_{n+1}}\mathbf{1}^T_{n+1}\right)(K_{n+1} + \Delta_{n+1})^{-1}, \tag{15}$$

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

$$\mathbb{E}[M_{n+1}(\mathbf{x})|\widetilde{A_n}] = \sum_{j=1}^{n} \lambda_{n+1,j}(\mathbf{x})\tilde{y}_i + \lambda_{n+1,n+1}(\mathbf{x})m_n(\mathbf{x}) \text{ and} \tag{16}$$

$$\text{Cov}[M_{n+1}(\mathbf{x}), M_{n+1}(\mathbf{x}')|\widetilde{A_n}] = \lambda_{n+1,n+1}(\mathbf{x})\lambda_{n+1,n+1}(\mathbf{x}')(s^2_n(\mathbf{x}^{n+1}) + \tau^2_{n+1}). \tag{17}$$

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:

$$\mathbb{E}[Q_{n+1}(\mathbf{x})|\widetilde{A_n}] = \sum_{j=1}^{n} \lambda_{n+1,j}(\mathbf{x})\tilde{y}_i + \lambda_{n+1,n+1}(\mathbf{x})m_n(\mathbf{x}) + \Phi^{-1}(\beta)s_{n+1}(\mathbf{x}), \tag{18}$$

$$Cov[Q_{n+1}(\mathbf{x}), Q_{n+1}(\mathbf{x}')|\widetilde{A_n}] = \lambda_{n+1,n+1}(\mathbf{x})\lambda_{n+1,n+1}(\mathbf{x}')\left(s^2_n(\mathbf{x}^{n+1}) + \tau^2_{n+1}\right), \tag{19}$$

and the values used in the quantile Expected Improvement (equation 8) are:

$$m_{Q_{n+1}} = \mathbb{E}[Q_{n+1}(\mathbf{x}^{n+1})|\widetilde{A_n}] \tag{20}$$

$$s^2_{Q_{n+1}} = Var\left[Q_{n+1}(\mathbf{x}^{n+1})|\widetilde{A_n}\right] = \left(\lambda_{n+1,n+1}(\mathbf{x}^{n+1})\right)^2\left(s^2_n(\mathbf{x}^{n+1}) + \tau^2_{n+1}\right) \tag{21}$$