Dealing with asynchronicity in parallel Gaussian Process based global optimization
David Ginsbourger, Janis Janusevskis, Rodolphe Le Riche

To cite this version:


HAL Id: hal-00507632
https://hal.archives-ouvertes.fr/hal-00507632
Submitted on 30 Jul 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Dealing with asynchronicity in parallel Gaussian Process based global optimization *

David Ginsbourger†‡
Joint Work with Janis Janusevikis § and Rodolphe Le Riche ¶

July 30, 2010

Abstract

During the last decade, Kriging-based sequential algorithms like EGO and its variants have become reference optimization methods in computer experiments. Such algorithms rely on the iterative maximization of a sampling criterion, the expected improvement (EI), which takes advantage of Kriging conditional distributions to make an explicit trade-off between promising and uncertain search space points. We have recently worked on a multipoints EI criterion meant to simultaneously choose several points, which is useful for instance in synchronous parallel computation. Here we propose extensions of these works to asynchronous parallel optimization and focus on a variant of EI, EEI, for the case where some new evaluation(s) have to be done while the responses of previously simulations are not all known yet. In particular, different issues regarding EEI’s maximization are addressed, and a proxy strategy is proposed.

*Deliverable 2.1.1-A of the OMD² ANR project (Optimisation Multidisciplinaire Distribuée)
†Senior Assistant, Department of Mathematics and Statistics, University of Bern, Switzerland
‡Scientific collaborator, Department of Hydrogeology, University of Neuchâtel, Switzerland
§Postdoctoral Researcher, LSTI/CROCUS team, Ecole Nat. Sup. des Mines de Saint-Etienne
¶CNRS UMR 5146 & LSTI/CROCUS team, Ecole Nat. Sup. des Mines de Saint-Etienne
1 Introduction

Sampling criteria such as the Expected Improvement (EI) allow sequentially choosing evaluation points in Gaussian Process (GP) based optimization algorithms. A major strength of using EI in a sequential procedure is that it progressively integrates available observations (points and responses), and thus avoids resampling at already explored points, at least in the case of noise-free simulations.

Recent works deal with the adaptation of EI for parallel Kriging-based optimization. In particular, a multipoints version of EI, called $q$-EI (or simply EI when there is no ambiguity), has been proposed for synchronous parallel Kriging-based optimization. $q$-EI measures the joint potential of a given additional $q$-points design of experiments. So, a natural parallelization of EI algorithms for $q$ synchronous processors is to maximize $q$-EI at each iteration, with update of the Kriging metamodel (including possible re-estimation of covariance parameters) whenever new observations are assimilated. However, maximizing $q$-EI is in most cases unaffordable, all the more so that the dimension of inputs $d$ and the number of points $q$ increase. Furthermore, evaluating $q$-EI when $q \geq 3$ bases on a Monte Carlo method relying on GP conditional simulations, which would make this intermediate optimization problem noisy and potentially very time consuming. Heuristic strategies such as the Constant Liar (CL) and the Kriging Believer (KB) have been proposed in [34] to circumvent the computational difficulties associated with $q$-EI maximization, by turning it into a stepwise maximization with virtual intermediate observation values.

In this report, we focus on the case of asynchronous parallel Kriging-based optimization, i.e. where sampling decisions have to be made before all previously started simulations have produced their result. Hence we have to distinguish between already visited points, currently visited (or ”busy”) points, and candidate points for forthcoming simulations. Having to take busy points into account within GP modeling sends back to similar problems encountered in the heuristic strategies mentioned above, so that we take advantage of previous work on that topic, and revisit those heuristic strategies in a more general manner.

After some necessary definitions concerning the EI criterion with (partially or completely) enriched information, a variant of EI is proposed for the case where a processor did not give its result yet. Relying on conditional simulations, EEI proves to be a sensible criterion, but practically not straightforward to maximize. A special variant of CL with random lie is then investigated to generate a finite set of candidate EEI maximizers, and gives surprisingly excellent results on the basis of a first 1-dimensional toy example.

Some research perspectives and a rich bibliography close this report, and set a framework for additional testing and developments to be made in forthcoming project phases.
2 Injecting incomplete information in EI criteria

2.1 EI with (partially) enriched information

The regular EI is defined as follows

$$EI(x) = \mathbb{E} \left[ \left( \min(Y(x)) - Y(x) \right)^+ | A \right],$$  \hspace{1cm} (1)

where $(\cdot)^+ = \max(0, \cdot)$, and $A = \{Y(X) = y\}$ is the event summing up all currently available observation points and corresponding responses.

Let us now assume that an evaluation of $y$ at point $x_{\text{busy}} \in D$ (say at an EI global maximizer, assumed unique for convenience) has been started, that the result is not known yet, and that one wishes to determine the next most promising point (in terms of EI) where to perform an evaluation using a newly available processor.

Ignoring that an evaluation is going on at $x_{\text{busy}}$ may cause losing crucial information for the next evaluation(s). However, it is not possible to add the event $\{Y(x_{\text{busy}}) = y(x_{\text{busy}})\}$ to $A$ since $y(x_{\text{busy}})$ is not known yet.

**Question 1** How is it possible to inject the partial knowledge that an evaluation is going on at $x_{\text{busy}}$ in the EI criterion without knowing $y(x_{\text{busy}})$?

Before giving a probabilistic answer to that question, let us introduce some necessary additional concepts and notations.

**Definition 1** Given $X$, $y$, $A$ as above, and an additional point $x^{\text{add}} \in D$ together with its corresponding response $y^{\text{add}}$, we denote by **EI with enriched information** the function

$$x \in D \rightarrow EI(x; x^{\text{add}}, y^{\text{add}}) := \mathbb{E} \left[ \left( \min(Y(X), y^{\text{add}}) - Y(x) \right)^+ | A, Y(x^{\text{add}}) = y^{\text{add}} \right]$$  \hspace{1cm} (2)

This modification of the EI criterion basically amounts to updating the underlying conditioning event $A$ to $\{Y(X) = y, Y(x^{\text{add}}) = y(x^{\text{add}})\}$. As a consequence, the EI with enriched information is 0 at $x^{\text{add}}$ whatever the observation value $y^{\text{add}}$. This property will be useful in the sequel since we will actually often be in situations where the value of $y^{\text{add}}$ is unknown.

A probabilistic answer to question 1 is to continue treating the unknown $y(x_{\text{busy}})$ as a random variable $Y(x_{\text{busy}})$, but to insert it in on the side of conditioning events by plugging it in the previously defined EI with enriched information.

**Definition 2** Let us define the **EI with partially enriched information** as

$$EI(x; x^{\text{busy}}) := EI(x; x^{\text{busy}}, Y(x^{\text{busy}})) = \mathbb{E} \left[ \left( \min(Y(X), Y(x^{\text{busy}})) - Y(x) \right)^+ | A, Y(x^{\text{busy}}) \right]$$  \hspace{1cm} (3)
For any given fixed \( x \in D \), \( EI(x; x_{busy}) \) is hence a function of the random variable \( Y(x_{busy}) \), thus inheriting from its randomness. Indeed, the classical EI formula can be developed, and the randomness of the obtained expression appears clearly through the instances of \( Y(x_{busy}) \) in the current minimum and in the Kriging mean.

Since the function \( EI(\cdot; x_{busy}) \) is consequently a random function, it does not make sense to maximize it in order to derive the next most promising point \( x_{new} \).

**Question 2** Indeed, what would it mean to maximize a random function?

Several alternatives are possible, such as maximizing quantiles of it conditional on the current information, or basically focus on the following conditional expectation function

**Definition 3** The expectation of \( EI(\cdot; x_{busy}) \) conditional on \( A \) is denoted by

\[
EEI(\cdot; x_{busy}) := \mathbb{E}[EI(\cdot; x_{busy})|A] \tag{4}
\]

Denoting by \( f_{Y(x_{busy})|A}(\cdot) \) the probability density function (pdf) of \( Y(x_{busy}) \) conditional on \( A \), which is deduced from the standard gaussian pdf \( \phi \) by an affine change of variable,

\[
f_{Y(x_{busy})|A}(y_{busy}) = \frac{1}{s(x_{busy})} \phi \left( \frac{y_{busy} - m(x_{busy})}{s(x_{busy})} \right) \quad \text{(whenever } s(x_{busy}) \neq 0 \text{ )} \tag{5}
\]

we may in fact explicitly derive the latter conditional expectation in integral form:

\[
EEI(\cdot; x_{busy}) = \int EI(\cdot; x_{busy}, y_{busy}) f_{Y(x_{busy})|A}(y_{busy}) dy_{busy} \tag{6}
\]

Now, \( EI(\cdot; x_{busy}, y_{busy}) \) depends on \( y_{busy} \) in a quite complicated non-linear way, and there is no hope of getting an analytical expression for \( EEI \), even though \( f_{Y(x_{busy})|A}(\cdot) \) is known and rather simple. However, coming back to its probabilistic definition, \( EEI \) can be approximated by averaging out the \( EI(\cdot; x_{busy}, y_{busy}) \)'s obtained when drawing a sufficiently high number of \( y_{busy} \) candidate values generated by using the conditional law of \( Y(x_{busy})|A \).

**Algorithm 1** Estimate \( EEI \) using MC simulations

1. for \( i \leftarrow 1, n_{sim} \) do
2. \( Y_{i,sim} \sim \mathcal{L}(Y(x_{busy})|A) \)
3. \( EI_{i}(x) = EI(x; x_{busy}, Y_{i,sim}) \)
4. end for
5. \( \hat{EEI}(x) = \frac{1}{n_{sim}} \sum_{i}^{n_{sim}} EI_{i}(x) \)
Thus, a straightforward way of getting statistical estimates of EEI$(x; x^{\text{busy}})$ (pointwise, for given $x \in D$, or of the whole function $\text{EEI}(.; x^{\text{busy}})$) is to base on Monte Carlo drawings, as in alg. 2.1. Obviously, the reliability of such a Monte Carlo approximation of EEI will depend on the invested number of drawings $n_{\text{sim}}$. To be more precise, on can state that for any fixed $x \in D$, the estimator $\widehat{\text{EEI}}(x)$ obtained above is unbiased, and has a variance given by a formula of a well-known kind:

$$\text{Var} \left( \widehat{\text{EEI}}(x) \right) = \frac{1}{n_{\text{sim}}} \int \left( \text{EI}(.; x^{\text{busy}}, y_{\text{busy}}) - \text{EEI}(x) \right)^2 f_{Y(x^{\text{busy}}), A(y_{\text{busy}})} dy_{\text{busy}} \quad (7)$$

### 2.2 Main example

Let us consider a deterministic one-dimensional function,

$$y : x \in [0, 1] \rightarrow y(x) = \frac{\sin(10x + 1)}{1 + x} + 2\cos(5x) \times x^4 \in \mathbb{R}, \quad (8)$$

and first approximate it using a Kriging model based on $y$’s observation at a three-points design $X = \{0, 0.475, 0.95\}$. The chosen Simple Kriging model is centered, has a Matérn covariance structure with regularity coefficient $\nu = \frac{3}{2}$, unit variance, and range parameter $\theta = \frac{0.5}{\sqrt{3}}$. The objective function, the Kriging mean predictor, and the corresponding Expected Improvement function are represented on figure 2.1. This very simple example will enable us to get a first illustration of the previously discussed notions, and to put the finger on some related problems, in a surprisingly rather general way (despite the low-dimensional framework). Let us now detail the example. First, we consider the Kriging predictor and the EI function of figure 2.1 both in blue. Assume that a first processor is vacant, and that one wishes to use it to perform a fourth simulation. It is then natural to perform this additional simulation at the point maximizing the current version of EI, that is to say at the point 0.698 magnified by an orange dashed line on figure 2.1.

Let us now pretend that the evaluation of $y$ at the latter point is getting started at a currently available processor, say proc$_1$, and set $x^{\text{busy}} = 0.698$. Assume further that a second processor, proc$_2$, is becoming available while proc$_1$ is still busy with the computation of $y(x^{\text{busy}})$. The main issue here is obviously to choose $x^{\text{new}}$, the point at which to evaluate $y$ with proc$_2$, in such a way that the all available information is optimally used.

Computing the EEI criterion of eq. 6 by means of a procedure such as alg. 2.1 will lead us to makes conditional simulations of $Y(x^{\text{new}})$ knowing the past observations. Before averaging out a whole collection of EI curves obtained by this means and maximizing the obtained EEI estimate, let us first focus on a single conditional simulation and detail the underlying mechanism. Figure 2.2 illustrates the effect of enriching EI with a simulated observation. The latter is represented by a green point on the upper graph, as well as the new (virtual) Kriging mean predictor. In the same color, the curve on the lower graph stands for the corresponding EI with enriched information. As mentioned earlier, this updated EI is null
Figure 2.1: A first Kriging model, with maximization of the Expected Improvement.
at $x^{busy}$. Furthermore, since the simulated value of $Y(x^{new})$ is particularly low, and hence appealing in our minimization context, EI is now pointing to values of $x$ close to $x^{busy}$ (note the EI values of smaller magnitude than at the previous step). However, choosing the global maximizer of this EI for $x^{new}$ does not completely make sense. Indeed, the simulated value of $Y(x^{new})$ is arbitrary, and another simulated value would maybe lead to discard this point and go for a candidate $x^{new}$ in a different zone of the input space.

It seems therefore necessary to propagate the uncertainty concerning the new EI function by integrating it over the possible scenarios concerning the value of $Y(x^{new})$, or at least to circumvent the problems associated to this unknown value and its variability by using some adapted heuristic. Here we propose at first to use brute force: simulating 100 values of $Y(x^{new})$ conditional on the past, calculate the 100 associated EI with enriched information,
Figure 2.3: 100 conditional simulations at $x^{\text{busy}}$ and the corresponding EEI estimate.

and take the arithmetic mean of them as proxy for EEI.

The lower graph of figure [2.3] above represents the estimate of EEI obtained by averaging out the 100 EI's with enriched information obtained by conditional simulation of $Y(x^{\text{new}})$ (maximum reached at $x \approx 0.347$). The corresponding values are materialized by colored points on the upper graph of figure 2.3. The 100 EI curves are printed on figure 2.4 with colors matching the one previously used on figure 2.3 for the simulated responses at $x^{\text{new}}$.

A natural surrogate for EEI maximization is to maximize its Monte Carlo estimate $\hat{\text{EEI}}$, such as represented on figure 2.3 (lower graph, see maximization result above). However, $n_{\text{sim}}$ has to be chosen sufficiently large in order to guarantee a good approximation of EEI, which may cause high computational costs. Furthermore, based on figure 2.4 above, one can imagine that a small set of carefully selected EI's might lead the user to a simplified
Figure 2.4: 100 simulations of the Expected Improvement with enriched information, based on the simulated $Y(x^{new})$ of figure 2.3.

description of this family of curves, yielding an efficient proxy for EEI maximization. Let us now explore these questions in some more detail.
3 Maximization of the EEI and related problems

3.1 Direct approach: maximizing the mean of a random function

3.1.1 A bit of theory and bibliography

Maximizing the previously introduced EEI criterion is in fact a special case of a classic stochastic optimization problem, i.e. the maximization of an expectation function such as

\[ x \in D \rightarrow \mathbb{E}[Q(x, \zeta)], \]  

(9)

where \( Q \) is a function depending on \( x \) and on the scalar- or vector-valued random variable \( \zeta \), and the distribution of \( \zeta \) is assumed known or at least reproducible by simulation.

Such problems are commonly solved in practice thanks to approximations of \( \mathbb{E}[Q(x, \zeta)] \) by finite averages of the form \( \frac{1}{N} \sum_{i=1}^{N} Q(x, \zeta^{(i)}) \), where \( N \) and the \( \zeta^{(i)} \) are chosen in order to get a good trade-off between the sample size (\( N \), wished small) and a good approximation of the distribution of \( \zeta \) by a uniform discrete law over the \( N \)-sample \( \{\zeta^{(1)}, \ldots, \zeta^{(N)}\} \).

One natural idea is to generate \( \{\zeta^{(1)}, \ldots, \zeta^{(N)}\} \) by Monte Carlo, i.e. to draw by simulation a \( N \)-sample of variables independently following the distribution of \( \zeta \). This is what Shapiro calls SAA (Sample Average Approximation). To quote him,

"The term sample average approximation method was coined in [47], although this approach was used long before that paper under various names. Statistical properties of the SAA method are discussed in [74] and complexity of two and multi-stage stochastic programming in [75]. Rates of convergence of Monte Carlo and Quasi-Monte Carlo estimates of the expected values are discussed in [58]. Numerical experiments with the SAA method are reported in [50, 51, 83], for example."

Such an approach makes sense, and without going into difficult detail concerning the different kinds of stochastic convergence, the law of large numbers ensures (under specific conditions) that the Sample Average Approximation will resemble the expectation objective function \( \mathbb{E}[Q(x, \zeta)] \) always more as \( N \) grows, and that so will do their respective maximizers. See e.g. [81] for a discussion on sufficient conditions for the sequence of proxy maximizers to converge to the actual maximizer in a suitable way.

Anyway, even under conditions where nice asymptotic properties might be guaranteed, the major question to answer in practice is the number of Monte Carlo simulations \( N \) needed to ensure that \( \arg \min \frac{1}{N} \sum_{i=1}^{N} Q(x, \zeta^{(i)}) \) is a reasonable approximation of \( \arg \min \mathbb{E}[Q(x, \zeta)] \). A theoretical study of this would certainly be very demanding, and is somehow out of scope here. However, we will come back in the next section to our previous example, and propose some illustrations related to these convergence questions.
3.1.2 Back to the main example

Before focusing on the estimation of EEI and on the maximization of the obtained estimate, let us briefly study the variability of the maximum and maximizer of EI with enriched information when $y_{busy}$ randomly varies following the distribution defined by density $\int$. The $y_{busy}$ values used here are the same as on figures 2.3 and 2.4. The variability shown by the upper left histogram of fig. 3.1 illustrates the fact that the value of $y_{busy}$ significantly influences the EI with enriched information and its maximum. Indeed, a very low value of $y_{busy}$ changes the current minimum, and makes it harder to exceed. On the other hand, a very high value of $y_{busy}$ does not affect the current minimum but may locally attract the Kriging predictor, and prevents from any improvement in the close neighbourhood of $x_{busy}$. However, depending on the rigidity of the covariance kernel used, such high $y_{busy}$ values might cause some low overshooting and could magnify the value of EI in farther areas. Finally, medium values of $y_{busy}$, in particular slightly above the current minimum, are susceptible of keeping the EI maximum at the same order of magnitude as earlier.

Figure 3.1: Left: histograms of the minima and minimizers of the 100 EI with simulated enriched information represented on figure 2.4. Right: joint distribution.

The lower left histogram of fig. 3.1 shows how the maximizer of the EI with enriched information behaves when $y_{busy}$ varies. We clearly distinguish three modes. The first one, around 0.35, roughly matches with the left local maximum of EI at the previous step. As can be seen on the left graph of fig. 3.2, this mode corresponds to relatively high values of $y_{busy}$. The second mode, around 0.6, gathers a smaller number of points with medium $y_{busy}$ values. A clear linear trend appears in the corresponding cluster of points on figure 3.2.
higher values of $y_{\text{busy}}$ tendentially matching with minimizers on the left, while smaller $y_{\text{busy}}$ seem pushing the maximizer of the enriched EI closer to $x_{\text{busy}}$. Note that a reverse trend is similarly visible on the first cluster, which may presumably be imputed to some kind of rigidity effect, as evoked earlier: a very high $y_{\text{busy}}$ value would drastically lower EI in the close neighbourhood of $x_{\text{busy}}$, but the associated curvature in the Kriging predictor would also create a steeper hill, moving the maximizer of the enriched $x_{\text{busy}}$ to the right of the left cluster. Finally, the right cluster corresponds to the lowest simulated values of $y_{\text{busy}}$, with a similar slope to the one of the first cluster. A moderately small value of $y_{\text{busy}}$ causes a slight decrease of the current minimum response, and sends the maximizer of the enriched EI to a zone with high variance, the right of the third cluster. As $y_{\text{busy}}$ decreases further, however, the influence of this very appealing value at $x_{\text{busy}}$ reaches the third cluster, progressively moving the maximizer of the enriched EI to the left of the cluster.

![Figure 3.2: Left: Dependence between the location of the maximizer of the enriched EI and $y_{\text{busy}}$. Right: maximum of the enriched EI as a function of $y_{\text{busy}}$.](image)

The right graph of fig. 3.2 represents the evolution of the maximum of the enriched EI as $y_{\text{busy}}$ increases. Contrarily to the previous graph, this scatter plot seems to stem from a continuous function, which makes sense from a mathematical point of view. Indeed, the enriched EI is a continuous function of $y_{\text{busy}}$, and hence absolutely continuous over any compact set. One can show that the maximum of EI with enriched information is consequently a continuous function of $y_{\text{busy}}$. As could already be felt on the basis of the previous graphs, only the range of medium values of $y_{\text{busy}}$ breaks the monotonicity of this function. Indeed, for very small values of $y_{\text{busy}}$, there isn’t much room for improvement. Then, the maximum increases until $y_{\text{busy}}$ reaches the actual current minimum (red vertical
line on the right graph of fig. 3.2. At this stage, the maximum of the enriched EI starts decreasing with increasing $y_{busy}$: higher $y_{busy}$ now do not change anything to the minimum of reference, but do pull the response surface up in the neighbourhood of the maximizer of the enriched EI, hence affecting its maximum negatively through the increasing Kriging mean predictor. At a certain point, however, the maximizer of the enriched EI jumps to the left cluster, and the associated maximum starts again to increase with $y_{busy}$. This last effect seems due to the previously evocated rigidity artifact, a higher observation value at $x_{busy}$ causing deeper valleys in the surrounding clusters.

We now continue with the maximization of the estimated EEI, and with an investigation concerning the number of $y_{busy}$ simulations needed here to achieve a good result.

Figure 3.3: Evolution of EEI maxima and maximizers with progressively growing sample

Fig. 3.3 shows how the maximum and the global maximizer of the estimated EEI evolves, when the sample size of $y_{busy}$ values used for the estimation of EEI varies. Here, only the 100 values previously simulated are used, and the increasing number of $y_{busy}$ values means that nested subsets of that 100 values are progressively considered (the first one, the two
first ones, and so on . . .). It is quite impressive to observe how fast the maximizer of the estimated EEI converges to the actual EEI maximizer. Here, after an unstable phase during the 8 first iterations (averaging with less or equal than the 8 first simulated $y_{bus}$ values), the maximizer of the estimated EEI stabilizes to a unique point. Meanwhile, the estimated EEI maximum seems to converge with a classical rate of convergence of $n_{\text{sim}}^{-\frac{1}{3}}$.

In order to see in what measure this early stability is due to the particular sample drawn in the previous example, let us generate other samples at random following the same distribution and observe the variability of the new estimated EEI maxima and maximizers. We follow for each simulated sample the same nesting scheme as above.

Figure 3.4: Boxplots retracing the evolution of EEI maxima and maximizers with progressively growing sample (1 to 20 here), when the sample is drawn 100 times at random.
Despite the particularly good results obtained in the previous example (stability for a sample of size greater than 8), the boxplots of fig. 3.4 are here to signify that this first attempt was rather lucky, and that significantly larger samples were needed in order to guarantee a successful approximate EEI maximization with high probability.

Figure 3.5: Left: boxplots retracing the evolution of EEI maxima and maximizers with progressively growing sample (1 to 100 here, 100 sample realizations). Right: proportion of approximate EEI maximizers in the first cluster in function of the sample size.

Let us finally see what happens if we replicate 100 times the complete previous experiment, i.e. every time with 100 simulated $y_{busy}$ values. The results are presented on fig. 3.5. On the left graph, the boxplots confirm again that we had been lucky in the previous experiment: here, we need to invest about 100 $y_{busy}$ simulations to get rid of approximate EEI maximizers landing in the right cluster (upper points in lower graph). As we can see on the right graph of figure 3.5, investing 20 conditional simulations leads to the good cluster (i.e. near to the actual EEI maximizer) in about 80% of the cases. This leaves 20% of misleading approximate optimization results, which a reasonable user may choose not to accept. Always on the same graph, increasing the sample size to 50 leads to a probability of about 95% to get an approximate optimizer in the good cluster, while the 100% probability level seems to be reached for a sample of 100 simulated $y_{busy}$ values. Working with an average of so many EI functions can be computationally very demanding, so that reliable and efficient methods involving a limited number of $y_{busy}$ values might be appealing. We address this issue in the following section, where encouraging results are obtained in the modest present framework case of a unique busy processor.
3.2 Towards reducing to a limited number of scenarii

3.2.1 Theory: per un pugno di quantili . . .

In the last section, we proposed to approximate EEI by averaging several EI criteria with enriched information, where the unknown response \( y_{busy} \) was randomly simulated following the "Kriging conditional distribution" of \( Y(x_{busy})|A \).

Here we depart from this Monte Carlo approach by deterministically choosing a representative subset of \( y_{busy} \) values, and restricting the EEI maximization to the maximizers of the different EI enriched with the latter \( y_{busy} \) values.

The approach proposed here to get a subset of \( y_{busy} \) values representing well the Kriging conditional distribution is to select quantiles of it. We focus on quantiles of level \( \alpha \), \( \alpha \) varying regularly between 0.05 and 0.95 with a finesse depending on the chosen number of subset points. Denoting \( n_{quant} \) this number of points, the proposed protocol reads:

1. Calculate \( n_{quant} \) quantiles of \( Y(x_{busy})|A \)
2. For \( y_{busy} \) equal to each of these quantiles, maximize the enriched EI
3. For each of the obtained maximizers, estimate EEI (possibly with the same \( n_{quant} \) \( y_{busy} \) values as earlier)
4. Choose the enriched EI maximizer with highest estimated EEI

Such a technique should a priori not be expected to work well. In fact, in a general stochastic optimization framework, nothing guarantees that the maximum of an expected function will be reached at the maximizer of some of the underlying realizations. It is even easy to build counter-examples: consider for instance the piecewise linear function \( f_1 \) over \([-1, 1]\) with slope 2 over \([-1, 0]\) and slope 1 over \([-1, 0]\), \( f_2 \) similarly defined with slopes \(-1\) and \(-2\), and \( f \) being a random function equal to \( f_1 \) or \( f_2 \) with probability one half. It is then clear that \( f_1 \) takes its global maximum (1) at 1, \( f_2 \) takes its global maximum (1) at \(-1\), but none of these points is a maximizer of the expectation of \( f \). Indeed, the expectation value of \( f \) at these two points is \(-1\), while the latter takes the value 0 (its maximum) at 0, which is not a maximum of \( f_1 \) nor \( f_2 \). Note that a variant of this counter-example with positive functions may be constructed by a simple vertical translation.

Coming back to our EI functions with enriched information, things are going much better than in the counter-example above. Indeed, it is quite unlikely that a point being an EI maximizer for some value of \( y_{busy} \) becomes dramatically bad for another \( y_{busy} \) value. Furthermore, it is natural to suspect that a point maximizing EI with a high EI value for some \( y_{busy} \) is a good one for EEI. If however the high potential of a point would be due to an artifact and rely on the precise \( y_{busy} \) value considered in the enriched EI, then we could expect that the fact of averaging enriched EI values over several \( y_{busy} \) candidates
would smooth this outlier out and cancel the domination of this point. Even if we are not convinced that this method should work in every situation, the example below shows quite encouraging results. Before coming to any general conclusions however, the latter have to be questioned and complemented by further investigations.

### 3.2.2 Back to the main example

Let us first choose \( n_{\text{quant}} = 10 \) and detail review the protocol above in the framework of our main example. We recall that the EEI maximizer obtained by Monte Carlo with a 100-sample over a 200-elements grid was 0.3467337. Now, we replace the Monte-Carlo sample by the 10 quantiles of \( Y(x^{\text{busy}}) | A \) with respective levels 0.05, 0.1, ..., 0.9, 0.95:

```r
> quantiles
[1] -1.52060808 -1.11769068 -0.87799880 -0.68650068 -0.51454523 -0.34811045
[7] -0.17615500 0.01534313 0.25503501 0.65795240
```

Maximizing the 10 obtained enriched EI criteria, we get the following candidates:

```r
> EI_maximizers
[1] 0.7487437 0.7688442 0.7788945 0.7939698 0.5929648 0.5728643 0.3467337
[8] 0.3517588 0.3567839 0.3618090
```

We notice that the seventh is exactly the previous EEI maximizer, and that the three following candidates are in its immediate neighbourhood. These values are plotted against the corresponding enriched EI values on the left graph of fig. 3.6. We can observe that the three clusters are well represented, even if the associated enriched EI estimates do not respect the order observed in fig 3.2. This is due to the approximation of these enriched EI on the basis of 10 \( y_{\text{busy}} \) values only.

We now restrict our attention to the 10 previously derived maximizers, and compute EEI estimates obtained at these points, enriched EI averages over the same 10 \( y_{\text{busy}} \) values. The right graph of fig 3.2 represents the 10 enriched EI values obtained for each one of the 10 points. The four last points seem to clearly dominate the others in terms of median EI, which would not be necessarily the case for other metrics penalizing uncertainty. Here we give the estimated EEI values obtained by the "quantile approximation":

```r
> EEI_values
[1] 0.03858103 0.04777052 0.05104971 0.05436474 0.05516403 0.05399162
[7] 0.07446641 0.07434650 0.07404384 0.07355171
```

And the point maximizing this rough EEI estimate among the 10 candidate points is the 7th, i.e. 0.3467337. This excellent result might seem relatively surprising, in particular its precision. But one may not forget that the search (enriched EI maximization) has
Figure 3.6: Left: maximizers and maxima of the enriched EI criteria associated with the 10 quantile-based \(y_{\text{busy}}\) values mentioned above. Right: enriched EI

been restricted in all cases to 200 points. Furthermore, this good finding might be highly depending on the number of quantiles, and even be non monotonic in \(n_{\text{quant}}\), which could lead to serious practical issues in the \textit{a priori} choice of \(n_{\text{quant}}\). So let us go one step deeper, and investigate the influence of \(n_{\text{quant}}\) on the approximate EEI maximizer obtained by the proposed method. Here below are the results obtained by repeating the quantile approximation and optimization method for \(n_{\text{quant}}\) varying from 1 to 30:

\[>	ext{maximizers}
\begin{align*}
[1] & 0.7487437 0.3618090 0.3618090 0.3467337 0.3517588 0.3517588 0.3467337 \\
[8] & 0.3517588 0.3467337 0.3467337 0.3467337 0.3467337 0.3467337 0.3467337 \\
[15] & 0.3467337 0.3467337 0.3467337 0.3467337 0.3467337 0.3467337 0.3467337 \\
[22] & 0.3467337 0.3467337 0.3467337 0.3467337 0.3467337 0.3467337 0.3467337 \\
[29] & 0.3467337 0.3467337
\end{align*}
\]

We observe that after a first point (0.7487437) in the third cluster, all followers belong to the first cluster, and are close neighbours of 0.3467337. Starting from \(n_{\text{quant}} = 9\), they are even all equal to 0.3467337. Compared to the variability observed on fig. 3.4 for the case of Monte Carlo EEI estimation, the "quantile method" provides a deterministic result, free of any variability, and with very good performances on this example in terms of locating the EEI correctly based on a small number of scenarios. Indeed, 2 or 3 quantile values are necessary to locate the EEI maximizer in the right zone ("first cluster"), while \(n_{\text{quant}} \geq 9\) guarantees the same result as found by Monte Carlo with a 100-elements sample (search
restricted to a 200-points grid). Despite the impressively good results obtained on this example, one has also to notice and remember that $n_{\text{quant}} = 1$ does lead to a sub-optimal cluster, which means in essence that plugging in the Kriging mean prediction (i.e. the KB strategy) as a proxy to the unknown $y_{\text{busy}}$ response is clearly not the best thing to do.

4 Conclusion and research perspectives

A variant of the EI criterion, EEI, has been proposed for the case of asynchronous Kriging-based parallel optimization. Here we have restricted our attention to the canonical situation where one processor is free while another one is still busy with the evaluation of the objective function at a candidate point $x_{\text{busy}}$. The unknown response at $x_{\text{busy}}$ can be simulated conditional on the available information, hence providing a means to estimate EEI.

Estimating EEI by conditional simulations amounts to averaging several enriched EI criteria obtained with the simulated response values at $x_{\text{busy}}$, i.e. applying several times a Liar strategy with a random lie generated by Monte Carlo. Maximizing the estimated EEI may then become quite computationally demanding, all the more so as the number of simulations gets large when it comes to approximating the actual EEI well. Here we relate that to a classical stochastic optimization problem, referred to as Sample Average Approximation in the literature, and propose to simplify it by summarizing the conditional distribution of the unknown response(s) by a small set of representative scenarios.

The idea of reducing to a few scenarios is not new, and what is proposed here is just a first draft of what could be developed for EEI maximization. In our main example, summarizing the unknown response distribution by a few quantiles has given sensible results, and a good picture of the full Monte Carlo results (with 100 drawings) with a lower computational cost (10 evaluations). Moreover, restricting EEI maximization to the maximizers of the enriched EI criteria obtained by the "quantile method" has delivered the same results as by maximizing the 100-drawings Monte Carlo EEI estimate over a 200-points grid. This is a quite unexpected outcome, to be urgently investigated with further test configurations. Apart from a bigger scale testing, and the generalization to multiple busy and vacant processors, supplementary research could be done on several aspects including:

- a fine study of the impact of $y_{\text{busy}}$ on the enriched EI and its maximizer(s),
- applying curve classification to the EI functions obtained with many $y_{\text{busy}}$ values,
- distinguishing regions of low EI variation, in particular in higher dimensions.

The final report will include further tests and theoretical considerations related to the questions above. A further study of several articles listed in the following bibliography —and with the help of [82]— is expected to shed some more light on existing tricks in Monte Carlo methods and scenarios selection for stochastic optimization.
References


