Batch sequential adaptive designs for global optimization

Yao Xiao¹ · Jianhui Ning² · Zikang Xiong² · Hong Qin¹,²

Received: 23 March 2021 / Accepted: 13 January 2022 / Published online: 25 January 2022
© Korean Statistical Society 2022

Abstract
Efficient global optimization (EGO) is one of the most popular sequential adaptive design (SAD) methods for expensive black-box optimization problems. A well-recognized weakness of the original EGO in complex computer experiments is that it is serial, and hence the modern parallel computing techniques cannot be utilized to speed up the running of simulator experiments. For those multiple points EGO methods, the heavy computation and points clustering are the obstacles. In this work, a novel batch SAD method, named “Accelerated EGO”, is forwarded by using a refined sampling/importance resampling (SIR) method to search the points with large expected improvement (EI) values. The computation burden of the new method is much lighter, and the points clustering is also avoided. The efficiency of the proposed batch SAD is validated by nine classic test functions with dimension from 2 to 12. The empirical results show that the proposed algorithm indeed can parallelize original EGO, and gain much improvement compared against the other parallel EGO algorithm especially under high-dimensional case. Additionally, the new method is applied to the hyper-parameter tuning of support vector machine (SVM) and XGBoost models in machine learning. Accelerated EGO obtains comparable cross validation accuracy with other methods and the CPU time can be reduced a lot due to the parallel computation and sampling method.

Keywords Computer experiments · Efficient global optimization · Sequential adaptive designs · Expected improvement · Hyper-parameter optimization

* Jianhui Ning
jhning@mail.ccnu.edu.cn

¹ School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan 430073, Hubei, China

² School of Mathematics and Statistics, Central China Normal University, Wuhan 430079, Hubei, China

Springer
1 Introduction

With the advent of new technologies, the computer experiments have become very popular, and more and more experiments are simulated in a mathematical model. Considering the complexity of the simulator model, some emulator/surrogate models, such as the Kriging model, have been proposed to approximate the real relationship between the input factors and outputs. Since the computation of computer experiments can be very heavy, a good experimental design is important. The most popular designs for computer experiments are the space-filling designs, such as the Latin hypercube design (LHD) and Uniform design (UD). A major flaw of the fixed-run space-filling designs is that all of the experimental points are fixed before any experiments are launched, and make no use of the information gained from the surrogate model. The other class of designs for computer experiments, called sequential adaptive designs (SAD), is initialized from a small size of space-filling design, and then is updated sequentially. The augmented design points are added based on the new information/features of the surrogate model. It has been believed that the SAD can be more efficient and effective than fixed-point designs (Lam and Notz 2008; Loeppky et al. 2010; Quan 2014). Some great SAD methods have been developed for different objective oriented experiments, such as the global optimization, level surface estimation (LS), percentile estimation (PE), and global fit (GF). In this paper, we only consider the SAD for global optimization.

Efficient global optimization (EGO) is a widely used SAD method for expensive black-box global optimization problems, which starts from a space-filling design and then sequentially updates the design one point a time (Jones et al. 1998). The new point is gained by maximizing the so called ‘expected improvement (EI)’ function. According to a statement proved by Locatelli (1997), Jones (2001) claims that as long as the iterates from this method are dense, the global optimum could be reached. It can be shown that EI criterion integrates the information of predicted means and standard deviations based on surrogate model, i.e., EI represents a trade-off sampling between promising and uncertain zones (Ginsbourger et al. 2010). Due to its excellent performance on global optimization, the EI criterion has been generalized for some more complex optimization occasions, for example the multi-fidelity optimization (He et al. 2017).

To efficiently take the advantages of parallel computation facilities to accelerate the running of computer experiments, some innovative researches have been proposed. Actually, while Schonlau (1997) proposed the single point EI, multipoints EI ($q$-EI) was also suggested for batch-sequential optimization ($q$ is the number of updating points at each cycle). However targeting multiple points at the same time involves numerous numerical computation that are extremely time-consuming. Ginsbourger et al. (2010) subsequently suggested two heuristic strategies, Kriging Believer and Constant Liar, to deal with this issue. These two strategies obtain $q$ updating points via sequentially optimizing EI $q$ times per cycle with an artificially updated Kriging model. Concretely, Kriging believer is to replace the response values by the Kriging mean predictor and constant liar
is to take an arbitrary constant value $L$ set by the user (usually the minimum or maximum of all currently available observations). However, the computational cost for these algorithms is still high, and Kriging Believer may get trapped in a non-optimal region for many iterations (Ginsbourger et al. 2010). Then an explicit formula allowing a fast and accurate deterministic approximation of $q$-EI was given by Chevalier and Ginsbourger (2013), for reasonably low values of $q$ (typically, less than 10). On the other hand, Ponweiser et al. (2008); Quan (2014) suggested some algorithms to reduce the computation burden and avoid the points clustering.

In summary, these multi-points EGO might be good batch SAD for global optimization, although the cost for generating the batch of points is time consuming and some methods may lead to points clustering, which is the motivation of our work. We proposed the Accelerated EGO algorithm by embedding a refined resampling way in order to obtain the batch of points, instead of employing the optimization techniques. The key advantage of this sampling method is that the computation burden can be relieved by a lot, and hence the speed of algorithm is increased. That’s the reason of the name “Accelerated EGO”. Meanwhile the resampling process is done with a randomized quasi Monte Carlo points set, which can avert the failure resulting from points clustering.

The remainder of the paper is organized as follows: in Sect. 2 the Kriging model and original EGO are introduced briefly, and then the Accelerated EGO is described in detail in Sect. 3; Some numerical comparisons among the new algorithm, ordinary EGO, and parallel EGO are presented in Sect. 4; In Sect. 5, the application of the proposed algorithm to hyper-parameter optimization in automatic machine learning is discussed; finally the main conclusions are made in Sect. 6.

2 EGO algorithm

Let $D = [a, b] = [a_1, b_1] \times \cdots \times [a_d, b_d]$ be a rectangular experimental region in $\mathbb{R}^d$ and let $f(x)$ be the simulator model whose explicit expression is unknown or known but rather complex on the region $D$. Then there exists a black-box optimization problem for finding $x^*$ and $M$ such that

$$M = f(x^*) = \min_{x \in D} f(x).$$

(1)

$M$ is the global minimum of $f(x)$ on $D$, and $x^*$ is called a minimum point on $D$.

This black-box optimization could be viewed as computer experiments. Suppose $x_i, i = 1, \ldots, n$ are experimental points over domain $D$, and $y_i = f(x_i)$ is the output corresponding to each $x_i$. Then a surrogate model $\hat{f}(x)$ is fitted based on the input-output data and used to help finding the global optimum of the $f(x)$.

2.1 A brief introduction to Kriging

The Kriging model was proposed by the South African geologist, D.G. Krige, in his master’s thesis (Krige 1951) on analyzing mining data and was systematically
introduced to model computer experiments by Sacks et al. (1989). A comprehensive introduction can refer to Roustant et al. (2012); Santner et al. (2018); Fang et al. (2018).

The Kriging model can be expressed as

\[ y(x) = \sum_{j=1}^{L} \beta_j B_j(x) + z(x), \]  

(2)

where \( \beta = (\beta_1, \ldots, \beta_L)' \) is the unknown regression coefficients, \( \{B_j(x), j = 1, \ldots, L\} \) is a chosen basis over the experimental domain, and \( z(x) \) is a Gaussian process with mean zero, variance \( \sigma^2 \), and correlation function \( r(x_i, x_j \mid \kappa) \), \( \kappa \) denote the parameters in various correlation functions.

The parameters \( \beta, \sigma^2 \) and \( \kappa \) in the Kriging model can be estimated by maximizing the likelihood function, and the MLEs are denoted as \( \hat{\beta}, \hat{\sigma}^2 \) and \( \hat{\kappa} \) respectively. The detailed estimation procedure can be referred to Fang et al. (2006). Then under the normality assumption on \( z(x) \), the empirical best linear unbiased prediction at point \( x \) is given by

\[ \hat{y}(x) = b(x)\hat{\beta} + r(x)R^{-1}(y - B\hat{\beta}), \]  

(3)

and the uncertainty quantification of this predictor is

\[ s^2(x) = \hat{\sigma}^2 \left[ 1 - (b'(x), r'(x)) \left( \begin{array}{cc} 0 & B' \\ B & R \end{array} \right)^{-1} \left( \begin{array}{c} b(x) \\ r(x) \end{array} \right) \right], \]  

(4)

where \( y = (y_1, \ldots, y_n)' \), \( R \) is an \( n \times n \) matrix whose \((i, j)\)-element is \( r(x_i, x_j \mid \hat{\kappa}) \). In addition, \( b(x) = (B_1(x), \ldots, B_L(x)) \), \( r(x) = (r(x_1, x \mid \hat{\kappa}), \ldots, r(x_n, x \mid \hat{\kappa})) \) and

\[ B = \begin{pmatrix} B_1(x_1) & B_2(x_1) & \cdots & B_L(x_1) \\ B_1(x_2) & B_2(x_2) & \cdots & B_L(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ B_1(x_n) & B_2(x_n) & \cdots & B_L(x_n) \end{pmatrix}. \]

### 2.2 EGO algorithm

The original EGO algorithm is an efficient single point SAD method for finding global optimum. In each design stage, it involves computing the EI function for all \( x \in D \), whose analytical expression is as follows:

\[ EI(x) = (y_{min} - \hat{y}(x)) \Phi \left( \frac{y_{min} - \hat{y}(x)}{s(x)} \right) + s(x)\phi \left( \frac{y_{min} - \hat{y}(x)}{s(x)} \right), \]  

(5)

where \( y_{min} \triangleq \min_{i=1,\ldots,n} y_i \) is the current minimum response, \( \phi \) and \( \Phi \) are the density and distribution of the standard normal distribution respectively, \( \hat{y}(x) \) and \( s^2(x) \) are the prediction (3) and predicted variance (4) derived from Kriging model.
Then the point $x^{\text{new}}$ with the largest EI value in domain $D$ is selected to be the next stage experimental point. This process is continued until the stop condition is met. Usually, the stop conditions for EGO algorithm are that the computational budget, i.e., the total number of evaluations is reached; or the setup precision of the output is obtained.

The procedure of this sequential experiment for global optimization (1) can be formed as Algorithm 1.

**Algorithm 1 EGO Algorithm**

**Require:** an initial design $\mathcal{P} = \{x_i, i = 1, \cdots, n\}, j = 0$, the maximum number of iterations $j_{\max}$, the pre-specified precision $\epsilon$

1. Evaluate $y_i = f(x_i)$, set $y = (y_1, \cdots, y_n)$, and fit the Kriging model;
2. while $j < j_{\max}$ or $|M_{\text{best}} - M| < \epsilon$ do
3. Find a new point $x_j$ in the experimental domain $D$ maximizing the $EI(x)$, i.e.,

   $$x_j = \arg\max_{x \in D} EI(x);$$
4. Evaluate $f(x_j)$, and set $\mathcal{P} := \mathcal{P} \cup \{x_j\}, y := y \cup \{f(x_j)\}$;
5. Re-fit the Kriging model with the design points $\mathcal{P}$, set $j = j + 1$.
6. end while
7. The output is the best approximation $M_{\text{best}}$ of the global optimum $M$.

## 3 Accelerated EGO

### 3.1 The randomized quasi-random SIR process

As discussed in Sect. 1, the big obstacles of multi-points or parallel EGO are the heavy computation and points clustering. Here, we propose a computationally easier algorithm to parallelize EGO algorithm.

Actually, there are two key properties of the updating points in parallel EGO algorithms. Firstly, the chosen points should be the peak of EI function, or at least with large EI values. Secondly, these points should be distributed as uniformly as possible around the target area. We employ the randomized quasi-random sampling/importance resampling (RQSIR) process to target the updating points and ensure the two properties as well. In the process of RQSIR, the EI function is seen as the kernel function of a probability density function, i.e.

$$x \sim g(x) \propto EI(x),$$

and updating points are generated from $g(x)$ by RQSIR method. The process of RQSIR includes the following three steps:
1. Choosing a candidate points pool $\mathcal{Q} = \{z_i = (z_{i1}, \ldots, z_{id}), i = 1, \ldots, m\}$.

It is well known that quasi-Monte Carlo (QMC) method is an efficient method for high-dimensional integration approximation. This method uses a set of points, called QMC sequence, to approximate the integration (Niederreiter 1992). The QMC points are also called low discrepancy sequence, they are uniformly distributed in the integrated domain. There are many ways to generate them, such as the Halton and Sobol sequence, which can be directly generated by R package “randombooltox” (Christophe and Petr 2019); Or the uniform design, which can be obtained from R package “UniDOE” (Zhang et al. 2018). Of course, other space-filling designs, like Mak and Joseph (2017), also can be employed.

2. Randomizing the candidate points pool $\mathcal{Q}$ to get $\mathcal{Q}_R$:

Unlike the Monte Carlo points, the QMC points are certainly deterministic. If adopting the points in this QMC points pool directly in the third “resampling” step, only these points have the chance to be chosen. Hence we employ the Randomized QMC in this step to introduce the randomness, and keep their distribution construction invariant as well. There are several randomization methods (L’Ecuyer and Lemieux 2002). Here, we only employ the simplest way, called random shift.

• Firstly, an $s$-dimensional random vector $\Delta \sim U(D)$ is generated, and then shift the whole points set $\mathcal{Q}$ to get

$$\mathcal{Q}_* = \{z_i^* = z_i + \Delta; z_i \in \mathcal{Q}, i = 1, \ldots, m\},$$

where $D$ is the experimental domain.

• Secondly, an analogized modular operation is put on $\mathcal{Q}_*$ to get $\mathcal{Q}_R = \{\tilde{z}_i = (\tilde{z}_{i1}, \ldots, \tilde{z}_{id}), i = 1, \ldots, m\}$, i.e.

$$\tilde{z}_{ik} = \begin{cases} a_k + (z_{ik}^* - b_k), & \text{if } z_{ik}^* > b_k, \\ z_{ik}^*, & \text{if } a_k \leq z_{ik}^* \leq b_k, \\ b_k - (a_k - z_{ik}^*), & \text{if } z_{ik}^* < a_k \end{cases} \quad (6)$$

for $i = 1, \ldots, m; k = 1, \ldots, d$.

The randomize QMC process for 2-dimension domain $D$ can be visualized as shown in Fig. 1.

![Fig. 1 The process of randomized quasi-Monte Carlo. (A Sobol sequence with 200 points, the shaded region is the experimental domain.)](image-url)
3. Resampling from the randomized candidate points pool $Q_R$

At each iteration, the EI values of each $\tilde{z}_i \in Q_R, i = 1, \ldots, m$ can be calculated according to (5). Then the augmented new experimental points $P^* = \{\tilde{x}_1, \ldots, \tilde{x}_J\}$ ($J \leq m$) are sampled from the pool $Q_R$ according to the weights

$$\alpha_i = \frac{EI(\tilde{z}_i)}{\sum_{j=1}^{m} EI(\tilde{z}_j)}, \quad i = 1, \ldots, m.$$  

(7)

3.2 Accelerated EGO algorithm

All in all, the RQSIR process helps to keep the diversity of candidates, ensures that every point in the experimental domain has chance to be chosen and can bring more exploration in the neighborhood of global optimum. In addition, these updating points can keep certain distance with the existing points. For the global optimization problem (1), by nesting the RQSIR process described above, the new batch sequential adaptive design method, Accelerated EGO, can be stated as in Algorithm 2. Note that the stop conditions for Accelerated EGO algorithm are the same as EGO algorithm.

**Algorithm 2** Accelerated EGO Algorithm

**Require:** an initial design $\mathcal{P} = \{x_i, i = 1, \ldots, n\}$, a candidate points pool $\mathcal{D} = \{z_i, i = 1, \ldots, m\}$, $j = 0$, the maximum number of iterations $j_{\max}$, the pre-specified precision $\epsilon$

1. Evaluate $y_i = f(x_i)$, set $y = (y_1, \ldots, y_n)$, and fit the Kriging model;
2. **while** $j < j_{\max}$ or $|M_{\text{best}} - M| < \epsilon$ **do**
3. Find a new point $x^{\text{new}}$ in the experimental domain $D$ maximizing the $EI(x)$, i.e.,

$$x^{\text{new}} = \arg\max_{x \in D} EI(x);$$

4. Randomization: randomize $\mathcal{D} = \{z_1, \ldots, z_m\}$ to get $\mathcal{D}_R = \{\tilde{z}_1, \ldots, \tilde{z}_m\}$;
5. Resampling: draw $J$ points $\mathcal{P}^* = \{\tilde{x}_1, \ldots, \tilde{x}_J\} (J \leq m)$ from $\mathcal{D}_R$ according to the weights

$$\alpha_i = \frac{EI(\tilde{z}_i)}{\sum_{i=1}^{m} EI(\tilde{z}_i)}; i = 1, \ldots, m$$

where $EI(\tilde{z}_i), i = 1, \ldots, m$ is calculated based on current Kriging model;
6. Evaluate $f(x^{\text{new}})$ and $\{f(\mathcal{P}^*)\}$, and set $\mathcal{P} := \mathcal{P} \cup \{x^{\text{new}}\} \cup \mathcal{P}^*$, $y := y \cup \{f(x^{\text{new}})\} \cup \{f(\mathcal{P}^*)\}$;
7. Re-fit the Kriging model with the design points $\mathcal{P}$, set $j = j + 1$.
8. **end while**
9. The output is the best approximation $M_{\text{best}}$ of the global optimum $M$. 
3.3 Toy example

Here, a 2-dimensional Ackley function is taken as a demonstrative example to show the procedure of the Accelerated EGO and do the sensitivity analysis about the size of initial QMC points pool. The Ackley function

\[
f(x_1, x_2) = -20 \exp \left( -\frac{1}{5} \sqrt{\frac{1}{2} (x_1^2 + x_2^2)} \right) - \exp \left( \frac{1}{2} (\cos(2\pi x_1) + \cos(2\pi x_2)) \right) + 20 + \exp(1)
\]

with \(-2 \leq x_1, x_2 \leq 2\), has several local optima around the global minimum \(x^* = (0, 0), f(x^*) = 0\) (Surjanovic and Bingham 2013).

Firstly, the QMC points pool for RQSIR is set to be a Sobol sequence with 100 candidates and 5 points are required to update per stage, i.e, the largest EI point and 4 points sampling from the randomized QMC points set. The Accelerated EGO needs 5 cycles to find the global optimum with the stop condition being \(| M_{best} - M | < 10^{-2}\). The search process can be seen from Fig. 2. At the initial stage, a uniform design with 21 points is scattered over the experimental domain. At stage

![Fig. 2](image_url)

Fig. 2 The search process of Ackley function. (The black dots “•” are the experimental points; the blue triangle “▲” is the global minimum point; the orange rectangle “■” is the largest EI point and the red lozenge “♦” are the remaining selected points at each iterations)
1–5, the point with the largest EI value is obtained and 4 points with relative large EI values are sampled from the 100 candidates. It can be seen that the sampling points are not flocked together and bring more exploration to the subdomain global optimum located at.

Then, the sensitivity analysis about the size of QMC points pool can be done based on this Ackley function. We choose the Sobol sequence with size $m = 50, 100, 150$ respectively to be the candidates. Five and ten updating points are selected respectively per stage by the Accelerated EGO algorithm and the stop condition is set to be $| M^\text{best} - M | < 10^{-2}$. In order to average out the bias caused by some latent factors, 100 repetitions are done for each case. Table 1 lists the number of cycles required to satisfy the stop condition for each combination of pool size $m$ and the number of updating points, where it can be observed that the number of cycles are not easily impressible about the size of QMC points pool. Of course the size $m$ can not be too small and is relative to the dimension $d$ of the optimization problem. General speaking, we recommend that the pool size $m$ could be $50d \sim 100d$ or so.

### 4 Numerical comparisons

In this section, we use nine classic multi-modal test functions with different dimensions to evaluate the performance of the proposed algorithm, Accelerated EGO. The majority of these test functions are employed by other refined EGO (Huang et al. 2006; Mehdad and Kleijnen 2018; He 2019) and parallel EGO (Ginsbourger et al. 2010; Viana et al. 2013) to validate algorithms’ efficiency. As a benchmark, the performance of original EGO and the most popularly used parallel EGO—constant Liar algorithm are also included.

#### 4.1 Test functions and algorithm setting

The nine test functions are obtained from Surjanovic and Bingham (2013) and are described as follows:

1. Branin function ($d = 2$)

| Table 1 | The influence of different size of QMC points pool |
|---------|--------------------------------------------------|
| Pool size $m$ | 5-point | 10-point |
|          | Median | Mean | S.t.d | Median | Mean | S.t.d |
| 50      | 5      | 6.05 | 2.84 | 4      | 4.69 | 1.25 |
| 100     | 4      | 4.27 | 2.39 | 4      | 5.12 | 2.37 |
| 150     | 5      | 5.60 | 3.07 | 4      | 5.06 | 2.35 |

The “5-point” is denoted the accelerated EGO with 5 updating points per cycle; “10-point” is analogous.
\[ f(x, y) = \left( y - \frac{5}{4\pi^2} x^2 + \frac{5}{\pi} x - 6 \right)^2 + 10 \left( 1 - \frac{1}{8\pi} \right) \cos x + 10 \]

with \( x \in [-5, 10] \), \( y \in [0, 15] \). The global minimum \( M = 0.397887 \) are located at \( x^* = (-\pi, 12.275), (\pi, 2.275) \) and \( (9.42478, 2.475) \).

2. **Six-hump camel function** (SixCamel) \((d = 2)\)

\[ f(x, y) = 4x^2 - 2.1x^4 + x^6/3 + xy - 4y^2 + 4y^4 \]

with \(-2 \leq x \leq 2, -1 \leq y \leq 1\). It has six extreme points, among them the global minimum are \( x^* = (0.0898, -0.7126), (-0.0898, 0.7126) \) and \( M = -1.0316e \).

3. **Modified Goldstein-Price function** (GoldPrice) \((d = 2)\)

\[ f(x, y) = \frac{1}{2.427} \left[ \log((1 + (x + y + 1)^2(19 - 14x + 3x^2 - 14y + 6xy + 3y^2)) \times [30 + (2x - 3y)^2(18 - 32x + 12x^2 + 48y - 36xy + 27y^2)]) - 8.693 \right] \]

with \(-2 \leq x, y \leq 2\). The global minimum \( M = -3.129126 \) is at point \( x^* = (0, -1) \) with several local minima around it.

4. **SIN2 function** \((d = 2)\)

\[ f(x, y) = 1 + \sin^2(x) + \sin^2(y) - 0.1 \exp(-x^2 - y^2) \]

with \(-5 \leq x \leq 5, -5 \leq y \leq 5\). It has many extreme points and the global minimum is \( x^* = (0, 0), M = 0.9 \).

5. **Hartmann function** with \( d = 3 \) (Hartmann3) and \( d = 6 \) (Hartmann6)

\[ f(x) = -\sum_{i=1}^{4} \alpha_i \exp \left( -\sum_{j=1}^{d} A_{ij}(x_j - P_{ij})^2 \right) \]

where \( 0 \leq x_i \leq 1, i = 1, 2, \ldots, d \) and \( \alpha = (1.0, 1.2, 3.0, 3.2)^T \). The elements of matrices \( A, P \) and the global minimum are shown in Table 2.

| Functions       | Parameters                                              | Table 2 Global minimum and settings of Hartmann function |
|-----------------|---------------------------------------------------------|--------------------------------------------------------|
| Hartmann3       | \( x^* = (0.1146, 0.5556, 0.8525) \) \  \( M = -3.86278 \) | \( A = \begin{pmatrix} 3.0 & 10 & 30 \\ 0.1 & 10 & 35 \\ 3.0 & 10 & 30 \\ 0.1 & 10 & 35 \end{pmatrix}, \quad P = 10^{-4} \begin{pmatrix} 3689 & 1170 & 2673 \\ 4699 & 4387 & 7470 \\ 1091 & 8732 & 5547 \\ 381 & 5743 & 8828 \end{pmatrix} \) |
| Hartmann6       | \( x^* = (0.2017, 0.1500, 0.4769, 0.2753, 0.3117, 0.6573) \) \  \( M = -3.32237 \) | \( A = \begin{pmatrix} 10 & 3 & 17 & 3.5 & 1.7 & 8 \\ 0.05 & 10 & 17 & 0.1 & 8 & 14 \\ 3 & 3.5 & 1.7 & 10 & 17 & 8 \\ 17 & 8 & 0.05 & 10 & 0.1 & 14 \end{pmatrix}, \quad P = 10^{-4} \begin{pmatrix} 1312 & 1696 & 5569 & 124 & 8283 & 5886 \\ 2329 & 4135 & 8307 & 3736 & 1004 & 9991 \\ 2348 & 1451 & 3522 & 2883 & 3047 & 6650 \\ 4047 & 8828 & 8732 & 5743 & 1091 & 381 \end{pmatrix} \) |
6. Ackley function with \( d = 10 \) (Ackley10)

\[
f(x) = -a \exp \left(-b \sqrt{\frac{1}{d} \sum_{i=1}^{d} x_i^2} \right) - \exp \left(\frac{1}{d} \sum_{i=1}^{d} \cos(c x_i) \right) + a + \exp(1)
\]

where \( a = 20, \ b = 0.2, \ c = 2\pi, \ -5.12 \leq x_i \leq 5.12, \ i = 1, \ldots, d \). And the global minimum is \( M = 0 \) at \( x^* = (0, \ldots, 0) \).

7. Levy function with \( d = 10 \) (Levy10)

\[
f(x) = \sin^2(\pi w_1) + \sum_{i=1}^{d-1} (w_i - 1)^2 \times \left[1 + 10 \sin^2(\pi w_i + 1)\right] + (w_d - 1)^2 [1 + \sin^2(2\pi w_d)]
\]

where \( w_i = 1 + \frac{x_i - 1}{4}, \ i = 1, \ldots, d \). And \(-10 \leq x_i \leq 10\), the global optimum is \( M = 0 \) at \( x^* = (0, \ldots, 0) \).

8. **Trid function** with \( d = 12 \) (Trid12)

\[
f(x) = \sum_{i=1}^{d} (x_i - 1)^2 - \sum_{i=2}^{d} x_i x_{i-1}
\]

where \(-d^2 \leq x_i \leq d^2, \ i = 1, \ldots, d\). And the global optimum is \( M = -d(d + 4)(d - 1)/6 \) at \( x_i = i(d + 1 - i), \ i = 1, \ldots, d \).

As for algorithm setting, the number of initial experimental points (uniform design) for the three algorithms and the size of QMC points pool (Sobol sequence) in each test functions are list in Table 3. The number of initial experimental points is selected according to the “10 thumb” principle (Jones et al. 1998; Loeppky et al. 2009) and the size of candidate pool is determined by our empirical experience 50 \( \sim \) 100 time of the corresponding dimension. As Ginsbourger et al. (2010) shows, Constant Liar algorithm with the constant setting being the minimum of the response values can get the best results. Therefore, in this paper, constant Liar algorithm with this setting is included, denoted by “CL(min)” as Ginsbourger et al.

| Test function | Initial design | Sobol pool | Test Function | Initial design | Sobol pool |
|---------------|----------------|------------|---------------|----------------|------------|
| Branin        | 21             | 100        | Ackley10      | 100            | 750        |
| SixCamel      | 21             | 100        | Levy10        | 100            | 750        |
| GoldPrice     | 21             | 100        | Trid12        | 120            | 1000       |
| SINE2         | 21             | 100        |               |                |            |
| Hartmann3     | 35             | 150        |               |                |            |
| Hartmann6     | 65             | 300        |               |                |            |
(2010) does. And all the experiments run on a PC consisting of 3.60 GHz Intel(R) Core(TM) i7-7700 processor and 16-GB memory.

### 4.2 Low-dimensional case

For the test functions with \( d = 2 \) to 6, the comparing metric is the number of stages which the sequential design require to satisfy the stop condition \( \left| M^{\text{best}} - M \right| < \epsilon \). The number of stages can measure the ability to parallelize original EGO. In addition, the number of evaluations is equal to \( q \) (the number of updating points per stages) times the number of stages, so only the number of stages is counted. In order to eliminate the influence of other inessential factors, all the experiments are repeated 100 times under this low-dimensional case. The results (the mean, standard deviation and median of the number of stages) are listed in Tables 4 and 5 shows the CPU time three algorithms need to find the optimum in the form of “mean (standard deviation)”.

The results listed in Table 4 indicate that the resampling method is efficient to generate updating points on these test functions no matter how many updating points are needed per stage. For example, EGO algorithm needs 13.89 stages averagely to satisfy the error for Branin function, while on an average, the 12-point Accelerated

**Table 4** Optimization comparisons under low-dimensional case

| Test function | \( \epsilon \) | Number of iterations | EGO | CL (min) | Accelerated EGO |
|---------------|---------------|----------------------|-----|----------|----------------|
|               |               |                      |     | 4-point  | 8-point  | 12-point | 4-point  | 8-point  | 12-point |
| Branin        | \( 10^{-2} \) | Mean                 | 13.89 | 3.95 | 3.46 | 3 | 4.04 | 2.89 | 2.45 |
|               |               | S.t.d                | 1.48 | 1.17 | 0.50 | 0.00 | 1.59 | 0.76 | 0.50 |
|               |               | Median               | 13   | 3    | 3    | 3    | 4    | 3    | 2    |
| SixCamel      | \( 10^{-3} \) | Mean                 | 9.58 | 3.60 | 2.71 | 2.60 | 4.61 | 3.50 | 2.78 |
|               |               | S.t.d                | 1.13 | 0.83 | 0.55 | 0.49 | 1.43 | 0.85 | 0.59 |
|               |               | Median               | 10   | 4    | 3    | 3    | 4    | 3    | 3    |
| GoldPrice     | \( 10^{-2} \) | Mean                 | 56.70 | 21.70 | 18.98 | 15.78 | 20.32 | 17.84 | 13.84 |
|               |               | S.t.d                | 31.27 | 11.48 | 10.65 | 10.84 | 13.17 | 10.73 | 8.55 |
|               |               | Median               | 57   | 22   | 18.5 | 13.5 | 18.5 | 15   | 11   |
| SIN2          | \( 10^{-2} \) | Mean                 | 29.85 | 8.45 | 5.00 | 3.96 | 8.68 | 5.33 | 4.01 |
|               |               | S.t.d                | 6.59 | 1.51 | 0.73 | 0.62 | 2.22 | 1.04 | 0.88 |
|               |               | Median               | 30.5 | 8.5  | 5    | 4    | 9    | 5    | 4    |
| Hartmann3     | \( 10^{-4} \) | Mean                 | 14.34 | 5.24 | 5.00 | 4.20 | 5.94 | 5.78 | 5.18 |
|               |               | S.t.d                | 3.42 | 0.65 | 3.73 | 1.95 | 1.94 | 1.66 | 1.18 |
|               |               | Median               | 13   | 5    | 4    | 4    | 5.5  | 5    | 4    |
| Hartmann6     | \( 10^{-1} \) | Mean                 | 21.7 | 6.04 | 4.83 | 4.06 | 5.62 | 4.91 | 4.48 |
|               |               | S.t.d                | 1.83 | 0.47 | 0.38 | 0.34 | 2.27 | 2.77 | 2.12 |
|               |               | Median               | 15   | 6    | 5    | 4    | 5    | 4    | 3    |

The “4-point” means that the number of updating points per iteration for CL (min) and accelerated EGO is four, others are analogous.
EGO needs only 2.45 iterations. For SIN2 function, the comparison is 29.85–4.01, and is 12.7–4.48 for Hartmann6 function. As the number of updating points increases, the mean, standard deviation and median reduce therewith. For example, when the number of updating points increases from 4 to 12, the mean of the number of stages is from 5.94 to 5.18, the standard deviation is 1.94–1.48 and the median is 5.5—4 for Hartmann3 function. These show that the Accelerated EGO becomes more efficient and robust when more updating points are selected per stage. However, the improvement of accelerated EGO is not strictly linear. That is to say, the reduction of iterations is not strictly in proportion to the number of updating points at each stage. For example, 4-point accelerated EGO algorithm speeds up \( \approx \frac{29.85}{8.68} \) times for SIN2 function, which should be 4 times for a linear speedup algorithm. Since the speedup of accelerated EGO algorithm is below linear, the number of evaluations that this algorithm requires to reach the error would be greater than that of ordinary EGO. But with the ability of parallel computation, this could not be a serious problem.

The comparisons between accelerated EGO and CL(min) algorithms under low-dimensional case are also shown in Table 4. It can be observed that Accelerated EGO performs similarly to CL (min) algorithm from the number of stages aspect for most test functions. For example, the number of stages of Accelerated EGO is 0.18 more than CL (min) averagely and the median of stages is the same to be 3 for SixCamel function with 12 updating points. And for Hartmann3 function, when 8 points are selected at each stage, the mean of stages is 4.83–4.91 for CL (min) and Accelerated EGO algorithms. However, it seems that Accelerated EGO has more advantages than CL (min) algorithm on test functions with complex surface. For example, GoldPrice function is thought to be hard to optimize with a sharp region of global minimum. The Accelerated EGO performs much better on this test function no matter how many updating points are selected. The comparisons of average number of stages are 21.70–20.32, 18.98–17.84 and 15.78–13.84e for CL (min) and Accelerated EGO algorithms with 4, 8 and 12 updating points per stage.

Another significant superiority, the CPU time needed to reach the given error of test functions, can be obtained for Accelerated EGO comparing with CL (min) algorithm. The CPU time is in the form of “mean (standard deviation)”

| Test functions | CL (min) | Accelerated EGO |
|----------------|----------|-----------------|
|                | 4-point  | 8-point | 12-point | 4-point  | 8-point | 12-point |
| Branin          | 5.9 (2.6)| 10.2 (2.5)| 12.3 (0.5)| 2.3 (1.1)| 1.6 (0.6)| 1.2 (0.4) |
| SixCamel       | 9.5 (2.2)| 12.2 (2.8)| 19.0 (4.4)| 3.6 (1.3)| 2.8 (0.8)| 2.3 (0.6) |
| GoldPrice      | 44.2 (26)| 95.1 (61)| 103.8 (92)| 14.9 (11)| 19.1 (14)| 15.8 (16) |
| SIN2           | 26.7 (5.8)| 30.12 (5.7)| 33.8 (7.1)| 8.0 (2.4)| 5.2 (1.6)| 3.9 (1.1) |
| Hartmann3      | 20.2 (4.1)| 35.0 (29.7)| 42.3 (21.8)| 10.1 (4.1)| 9.5 (2.9)| 8.1 (2.4) |
| Hartmann6      | 68.7 (12)| 98.9 (17)| 106.7 (15)| 33.8 (23)| 34.0 (32)| 38.5 (57) |
is similar for CL (min) and accelerated EGO algorithm, the most noticeable issue is that the complexity of targeting the multiple updating points. As Table 5 shows, accelerated EGO is a faster method to optimize all the test functions no matter how many updating points are used at each stage. For example, the CPU time is 3.6 s for 4-point Accelerated EGO to satisfy the stop condition of SixCamel function, while the CL (min) algorithm runs 9.5 s averagely. This advantage can be more significant when more updating points are selected per stage. For example, when 4 updating points are added per stage for SIN2 function, the CPU time of CL (min) is averagely 3.3 (≈ 26.7/8.0) times as much as Accelerated EGO, while for 12-point situation, the ratio is 8.6 (≈ 33.8/3.9). This phenomenon indicates that with Accelerated EGO algorithm, the computational burden for targeting multiple points can be eased a lot.

4.3 High-dimensional case

In this subsection, the Ackley10, Levy10 and Trid12 test functions are studied. Under the high-dimensional case, it is hard to reach the given error for the test function with limited evaluations. Thus, the stop condition is set to be the number of experimental runs in total. And the comparing metric is the minimum value the three algorithms can obtain with the given number of evaluations.

For Ackley10 and Levy10 test functions, the total runs is set to be 250. Subtracting the runs of initial design (a uniform design with 100 points), 150 updating points are available. For Trid test function, the number of updating points is 150, and with a 120-point initial design the total runs are 270. EGO algorithm sequentially run these 150 updating points one by one for these three test functions. Parallel algorithms, CL (min) and accelerated EGO, run 10 and 15 updating points respectively per stage. All the simulations in this case are repeated 10 times. The results (the minimum value the algorithms can get and the CPU time) are recorded in Table 6 in the form of “mean (standard deviation)”.

Under high-dimensional case, Accelerated EGO algorithm outperforms CL (min) significantly. From the minimum value the algorithms can obtain aspect, Accelerated EGO can reach smaller response value than CL (min) algorithm for all three

| Test functions | EGO | CL (min) | Accelerated EGO |
|----------------|-----|---------|-----------------|
|                |     | 10-point | 15-point        | 10-point | 15-point |
| Ackley10       | $M_{\text{best}}$ | 0.89 (0.8) | 1.23 (0.7) | 1.54 (0.8) | 0.89 (0.4) | 0.97 (0.2) |
|                | Time | 4793 (698) | 3574 (473) | 4057 (1254) | 412 (82) | 230 (22) |
| Levy10         | $M_{\text{best}}$ | 2.86 (1.8) | 3.77 (1.1) | 6.80 (3.4) | 3.63 (1.3) | 5.54 (1.5) |
|                | Time | 5970 (776) | 5688 (1239) | 5099 (264) | 564 (168) | 365 (37) |
| Trid12         | $M_{\text{best}}$ | −347 (2) | −245(54) | −230(38) | −251(17) | −243(51) |
|                | Time | 3042 (215) | 5512 (197) | 24,116 (2105) | 399 (77) | 306 (111) |

The total number of updating points is 150 for every case. The results are in the form of “mean (standard deviation)’. The “min” is the the minimum response value the algorithms can reach.
high-dimensional test functions with the same number of total runs and updating points per stage. For example, for Ackley10 function, the minimum value 10-point Accelerated EGO obtains is 0.89 on average, while the data of 10-point CL (min) is 1.23. And the comparison is $-251 \text{ to } -245$ for Trid12 function. From CPU time’s point of view, much less CPU time is required for Accelerated EGO compared to CL (min) algorithm. When parallel computation is available, reduction for the time of iterations is vital, especially under high-dimensional case. For example, with total 150 updating points for Levy10 function, 10-point CL (min) algorithm costs 5688 s (nearly 1.5 h) averagely to do the simulations, while Accelerated EGO only requires 564 s (nearly 10 min). Especially when the dimension and the number of updating points increase, this phenomenon becomes more highlighted. For example, for Trid function, the 15-point CL(min) algorithm run 24,116 seconds (nearly 7 h) unexpectedly while Accelerated EGO needs only 306 s (nearly 5 min) on average. These results show that accelerated EGO is actually a computationally tractable method to parallelize EGO, especially under high-dimensional case.

Comparing with EGO algorithm, it validates again that Accelerated EGO is not a strictly linear speedup algorithm. Under high-dimensional case, EGO can reach a smaller response value than accelerated EGO for most situations with the same number of total runs. For example, the minimum value of Levy10 function is 2.86 averagely for EGO algorithm, while the 15-point accelerated EGO gets 5.54. Only in one situation, Accelerated EGO obtains the same minimum value 0.98 averagely with EGO algorithm when 10 updating points is selected per stage for Ackley10 function. Of course, when multiple points are updated per stage, the CPU time of Accelerated EGO is much less than EGO with the same number of total runs. For example, for Trid12 function EGO algorithm runs 3042 s to evaluate 150 updating points, while 15-point Accelerated iterates 10 times with 306 s. But CL(min) does not have this advantage due to its hard computation for targeting the multi-points under high-dimensional case. Therefore, Accelerated EGO would be a better choice if the the parallel computation can be conducted.

As discussed in previous sections, the searching speed of the Accelerated EGO is much faster than other methods. Then we would like to know whether can the saved time be used to find better results. That is, if the computation time of Accelerated EGO is extended a little bit, whether can a much smaller response value be observed. Actually, some additional simulations give the positive answer. For example, for Ackley10 function, if we run the 10-point Accelerated EGO 5 iterations more (i.e., 20 iterations in total), the minimum response value is 0.63 averagely, which is much smaller than the value 0.89 of EGO algorithm. And the CPU time of Accelerated EGO increases slightly being 448 s averagely. Similar results can be seen for other simulation cases.

### 5 Application in hyper-parameter optimization

Recently, the task of choosing the best set of hyper-parameters for a machine learning model automatically, i.e., a set of hyper-parameters that yields the best performance of the predictor on the available data set, is routinely facing with data
scientists (Hutter et al. 2019). This problem can be treated as a black-box optimization situation for finding the global optimum of an expensive function which is only vaguely specified and has many local optima.

5.1 The procedure of hyper-parameter optimization

The visual representation for the process of machine learning with hyper-parameter optimization is shown in Fig. 3. The data set is divided into two part—training data and testing data—according to certain criterion. Then the hyper-parameter optimization approaches are taken to select the best hyper-parameter combination for machine learning model based on the training data. As long as the optimal hyper-parameters setting is determined, the best machine learning model can be trained with the training data. Hence the test data can be predicted with the best machine learning model.

The k-fold cross-validation is used to measure the estimates of average accuracy (or generalization error) in the training data for each hyper-parameter combination. The measure metrics could be the accuracy of prediction or classification, the value of AUC, etc. The k-fold cross-validation is the most common performance assessment method to measure the performance for machine learning model. That is the training data is randomly split into \( k \) mutually independent subsets with approximately equal sizes, then the \( k - 1 \) subsets are employed to train the machine learning model and the model is tested on the remaining one subset. This procedure is repeated \( k \) times and the estimate of average accuracy is obtained by averaging the test accuracies over \( k \) trials.

In this section, we mainly focus on methods discussed above (EGO, CL and accelerated EGO) to do hyper-parameters optimization. The result of the \( k \)-cross-validation corresponding to each hyper-parameter combination is regarded as the response value to fit the Kriging model. And then the best hyper-parameter trial can be obtained by these surrogate-based optimization methods.

5.2 Support vector machine (SVM)

The SVM with linear and nonlinear kernels is one of the most promising learning models for classification (Cristianini and John 2000). Selecting optimal
hyper-parameters (including the regulation parameter $C$ and the Gaussian kernel parameter $\gamma$) to lead to a better generalization performances has become a major challenge at hand. As shown in Fig. 4, we plot the 5-fold cross validation average accuracy for three data sets, “german.n”, “svmguide2” and “glass”, provided by Chang and Lin (2011) in a 3-dimensional surface, where $x$-axis and $y$-axis are $\log_2 \gamma$ and $\log_2 C$, respectively. The $z$-axis is the 5-fold cross-validation average accuracy. It is obvious that there are many local optima and it is hard to obtain the global optimum.

In this subsection, the EGO, CL(min) and our proposed Accelerated EGO algorithms are applied to hyper-parameter optimization for SVM with different data sets: heart, german.n, australian, svmguide2 and glass. These data sets are provided by Chang and Lin (2011), and the details (classes, samples, features, etc.) of five data sets are shown in Table 7. As a benchmark, the grid search method is also conducted. Firstly, each data set is divided randomly into two parts: the training data set with 75% samples and testing data set with the remaining samples.

For grid search method, 441 hyper-parameter combinations are searched and the hyper-parameter pair with the largest 5-fold cross-validation average accuracy among them is chosen. Then this selected hyper-parameter pair is employed to “train” the SVM model with all training data and the predicting accuracy of testing data could be obtained.

For EGO, CL (min) and Accelerated EGO algorithms, a uniform design with 21 points is set to be the initial design and 5 updating points are selected at each iteration for CL (min) and accelerated EGO. We run EGO algorithm 20 iterations,
constant Liar and Accelerated EGO 4 iterations, i.e., all algorithms are tuning 41 hyperparameters including the initial experimental points. For Accelerated EGO, the initial QMC point pool is the sobol sequence with size 100. In addition, 5-fold cross validation is used to evaluate the performance of SVM model.

In the process of hyper-parameter optimization, in order to realize the ‘real’ parallelization of constant Liar and accelerated EGO algorithms, the R packages “foreach” (Ooi et al. 2019b) and “doParallel” (Ooi et al. 2019a) are employed to do the parallel computing. Specifically speaking, at each iteration, the evaluations of the 5 updating experimental points are simultaneously done on 5 different cores of the computer respectively, i.e., the 5-fold cross-validation average accuracy corresponding to the 5 updating hyper-parameter combinations are obtained simultaneously.

Table 8 demonstrates the comparison among the results of grid search, EGO, CL (min) and the proposed Accelerated EGO methods for each data set using the SVM classifier. In each methods, the largest cross validation accuracy denoted by “CV” and the prediction accuracy of the testing data set denoted by “predict” are listed. In addition, the CPU time (in seconds) charged for the execution of user instructions of the hyper-parameter optimization process are also recorded. For grid search 10 replications are run due to its high computation cost, and others run 100 replications. The results are shown in the form of “mean (standard deviation)”.

The grid search can get the best cross validation accuracy for most data sets due to its exhausted computation, however other methods can be little short of grid search for their less computation and the nearly same accuracy. Comparing among EGO, CL (min) and accelerated EGO, EGO can get the best cross validation accuracy for all the data set except “german.n”. For this “german.n” data set, the cross

| Data set   | Grid CV | Grid predict | Grid time | EGO CV | EGO predict | EGO time |
|------------|---------|--------------|-----------|--------|-------------|----------|
| Heart      | 0.853 (0.007) | 0.835 (0.015) | 52.82 (1.29) | 0.853 (0.594) | 0.843 (0.014) | 14.64 (2.22) |
| Australian | 0.879 (0.002) | 0.849 (0.007) | 139.16 (0.96) | 0.877 (0.486) | 0.847 (0.007) | 15.29 (1.53) |
| German.n   | 0.770 (0.007) | 0.759 (0.009) | 1003.59 (34.84) | 0.769 (0.430) | 0.756 (0.006) | 30.12 (13.33) |
| Svmguide2  | 0.850 (0.007) | 0.829 (0.013) | 82.28 (2.41) | 0.847 (0.705) | 0.831 (0.012) | 18.67 (3.09) |
| Glass      | 0.704 (0.017) | 0.733 (0.020) | 46.33 (2.12) | 0.699 (1.413) | 0.729 (0.029) | 14.26 (2.76) |

| Data set   | CL (min) CV | CL (min) predict | CL (min) time | Accelerated EGO CV | Accelerated EGO predict | Accelerated EGO time |
|------------|-------------|------------------|---------------|---------------------|-------------------------|----------------------|
| Heart      | 0.851 (0.558) | 0.845 (0.019) | 11.71 (2.12) | 0.853 (0.553) | 0.841 (0.018) | 3.20 (0.56) |
| Australian | 0.872 (0.433) | 0.841 (0.010) | 11.66 (1.68) | 0.872 (0.506) | 0.844 (0.008) | 2.76 (0.48) |
| German.n   | 0.769 (0.382) | 0.757 (0.007) | 11.21 (2.27) | 0.771 (0.422) | 0.757 (0.008) | 2.64 (0.58) |
| Svmguide2  | 0.842 (0.721) | 0.828 (0.013) | 17.04 (2.71) | 0.844 (0.670) | 0.826 (0.014) | 3.66 (0.69) |
| Glass      | 0.695 (1.393) | 0.733 (0.029) | 12.86 (1.54) | 0.698 (1.268) | 0.736 (0.028) | 3.47 (0.76) |

The results are in the form of “mean (standard deviation)”. The “CV” means the cross-validation average accuracy. The “predict” is the predicting accuracy of the testing data set.
validation accuracy of Accelerated EGO is 0.771 averagely which is the largest among these three algorithms. And accelerated EGO performs better than CL (min) from the cross validation accuracy aspect for most data set, for example, for “heart”, the cross validation accuracy of Accelerated EGO is 0.002 (= 0.853 – 0.851) more than CL (min) averagely.

It can be observed that, thanks to the parallel computation provided by R packages, the CPU time for the hyper-parameter optimization process of CL (min) and Accelerated EGO are less than EGO algorithm. For accelerated EGO, much significant improvement can be observed. For example, for “australian” data set, the CPU time of EGO is 5.5 (∼ 15.29 / 2.76) times as much as Accelerated EGO, and the CL(min) gets little reduction of the CPU time with 11.66 s since it may cost a lot to find the multiple updating points per cycle. That is to say, accelerated EGO indeed speeds up and lessens the computation burden greatly for targeting the multiple updating points per iteration.

In addition, if we conduct additional experiments for this situation, much better results can be recorded. If 8 more iterations is added for Accelerated EGO algorithm, i.e., we run this algorithm 12 iterations (60 updating points totally), for “australian”, the CPU time of the hyper-parameter tuning process is 9.48 s on average which is still less than that (15.29 s) of EGO. And the cross validation accuracy of Accelerated EGO is 0.879 averagely, which is the same as the result of exhausted grid search and is larger than that (0.877) of EGO. Similar results can be observed for “svmguide2” and “glass” data set.

5.3 XGBoost

Extreme gradient boosting (XGBoost) model is one of the popular modern machine learning models. However, it usually involves many hyper-parameters required to be set. In this subsection, the black-box optimization methods are employed to tune the hyper-parameters for this model.

A data set “winequality” about the evaluation of wine quality from UCI machine learning database is selected. The data set is consisted of 6497 samples and 11 features. For XGBoost, the hyper-parameters are usually divided into general parameters, Booster parameters and task parameters. More details can be referred to Chen and Guestrin (2016). The seven hyper-parameters required to tune is listed in Table 9. Note that the tuning space of “eta” and “gamma” is transformed into their log values.

A uniform design with 70 points is selected as the initial design for EGO, CL (min) and Accelerated EGO algorithms. The total experimental points are set to 120. For accelerated EGO algorithms, the candidate point set is a Sobol sequence with 750 points, and 10 experimental runs are updated at each iteration. At the same time, because random search are more efficient than grid search in high dimensional space, random search with 120 hyper-parameter combinations is used as competitor.

First, the data with a “quality” score greater than 6 is marked as “1” and the rest as “0”. In the meanwhile, the training set and testing set are randomly divided in a ratio of 7 : 3. The average AUC value of 5-fold cross-validation is taken as
the evaluation standard of generalization ability for each hyper-parameter. AUC value is a commonly used index to evaluate the performance of a machine learning binary classifier. It represents the area under the ROC curve. Then, hyper-parameter optimization methods (including random search, EGO, CL (min) and accelerated EGO algorithms) are used to select the optimal hyper-parameter combination. The optimal hyper-parameter combination obtained by each optimization method is shown in Table 10. Finally, XGBoost model is trained with all the training set data under the optimal hyper-parameters, and the test set is predicted, its AUC value and classification accuracy are calculated. It can be seen that Accelerated EGO performs better than others, especially the prediction ability.

Table 9  The hyper-parameters of XGBoost model

| Hyper-parameters     | Type     | Range                | Experimental domain |
|----------------------|----------|----------------------|---------------------|
| Max_depth            | Discrete | {2, 3, 4, …, 8, 9}   | {2, 3, 4, …, 8, 9}   |
| Eta                  | Continuous | [10^{-5}, 10^0]       | [-5, 0]             |
| Gamma                | Continuous | [10^{-5}, 10^0]       | [-5, 0]             |
| Subsample            | Continuous | [0.6, 0.9]            | [0.6, 0.9]          |
| Colsample_bytree     | Continuous | [0, 1]                | [0, 1]              |
| Min_child_weight     | Discrete | {1, 2, 3, …, 40}      | {1, 2, 3, …, 40}    |
| Max_delta_step       | Discrete | {1, 2, 3, …, 10}      | {1, 2, 3, …, 10}    |

Table 10  The results of XGBoost model

|                      | Random  | EGO     | CL (min) | Accelerated EGO |
|----------------------|---------|---------|----------|-----------------|
| Max_depth            | 8       | 8       | 9        | 8               |
| Eta                  | 0.0276  | 0.0138  | 0.0065   | 0.0393          |
| Gamma                | 7.3 \times 10^{-2} | 3.4 \times 10^{-5} | 4.2 \times 10^{-3} | 5.8 \times 10^{-3} |
| Subsample            | 0.79    | 0.80    | 0.87     | 0.72            |
| Colsample_bytree     | 0.83    | 0.49    | 0.37     | 0.69            |
| Min_child_weight     | 3       | 1       | 1        | 2               |
| Max_delta_step       | 7       | 5       | 10       | 2               |
| AUC (CV)             | 0.8907  | 0.8993  | 0.8958   | 0.8968          |
| AUC (predict)        | 0.7707  | 0.7670  | 0.7333   | 0.7739          |
| Accuracy (predict)   | 0.8850  | 0.8901  | 0.8785   | 0.8914          |

“Random” stands for the random search. “AUC (CV)” stands for the cross-validation AUC of training data, “AUC (predict)” and “accuracy (predict)” are the AUC and accuracy of testing data.
6 Conclusion and discussion

In this paper, a novel batch sequential adaptive design method, accelerated EGO, is proposed to parallelize EGO algorithm and to release the computation burden for targeting the multiple updating points per stage. This algorithm employs the techniques of randomized quasi-Monte Carlo and resampling to select the updating points and it is a fast way to obtain samples with good representativeness and large EI values.

The efficiency of the proposed Accelerated EGO algorithm is validated over nine test functions with dimension from 2 to 12. Summarized the empirical computations, we can get following conclusions:

1. Accelerated EGO is a batch version of ordinary EGO algorithm. The results show accelerated EGO indeed speeds up the convergence of EGO algorithm and saves the cost of stages. Further more, the empirical simulation also show that the saved time can be transformed to the better results.
2. Compared with the counterpart parallel EGO—constant Liar algorithm, much significant improvement can be observed. Under low-dimensional case, accelerated EGO performs similarly with constant Liar with less computational burden to select the multiple updating points. However, accelerated EGO can get more acceptable optimum value than constant Liar algorithm under high-dimensional case. And in this situation, the reduction of computation cost for Accelerated EGO is much highlighted.

In addition, EGO, constant Liar and accelerated EGO algorithms are applied to hyper-parameter optimization for SVM and XGBoost models. The results show with the real utility of parallel computational technique, accelerated EGO is more efficient than EGO and releases the computational burden to target the multiple updating points per iteration dramatically.

Supplementary information Some additional numerical results are presented in the Supplementary Materials. The codes for this paper are provided in the Github website: https://github.com/zuelxy/Accelerated_EGO.

Appendix

The constant Liar (CL), one of the parallel EGO algorithms, is a sequential strategy for targeting the multiple points at each cycle, in which the surrogate model is updated (still without hyper-parameter re-estimation) at each iteration with a value \( L \in \mathbb{R} \) exogenously fixed by the user, called a “lie”. Three values, \( \min\{y\} \), \( \max\{y\} \) and \( \text{mean}\{y\} \) were considered in Ginsbourger et al. (2010). In our simulations, the lie “L” is taken as \( \min\{y\} \).
Algorithm 3 Constant Liar Algorithm

Require: an initial design $\mathcal{D} = \{x_i, i = 1, \ldots, n\}$, a fixed number “L”, $j = 0$, the maximum number of iterations $J_{\text{max}}$, the pre-specified precision $\varepsilon$

1: Evaluate $y_i = f(x_i)$, set $y = (y_1, \ldots, y_n)$, and fit the Kriging model;
2: while $j < J_{\text{max}}$ or $|M^{\text{best}} - M| < \varepsilon$ do
3:     for $k$ in $1 : q$ do
4:         $x^{n+k} = \arg \max_{x \in \mathcal{D}} EI(x)$
5:         $\mathcal{D} := \mathcal{D} \cup \{x^{n+k}\}$
6:         $y := y \cup \{L\}$
7:     end for
8:     Evaluate $\mathcal{D}$ and re-fit the Kriging model with the design points $\mathcal{D}$, set $j = j + 1$;
9: end while
10: The output is the best approximation $M^{\text{best}}$ of the global optimum $M$.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s42952-022-00161-9.

Acknowledgements This study is partially supported by the National Natural Science Foundation of China (Nos. 11571133, 11871237 and 12001540).

References

Chang, C. C., & Lin, C. J. (2011). Libsvm: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 27(2), 1–27. http://www.csie.ntu.edu.tw/~cjlin/libsvm
Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 785–794). New York: ACM.
Chevalier, C., & Ginsbourger, D. (2013). Fast computation of the multi-points expected improvement with applications in batch selection. In International conference on learning and intelligent optimization (pp. 59–69). New York: Springer.
Christophe, D., & Petr, S. (2019), randtoolbox: Generating and Testing Random Numbers. R package version 1.30.0. https://CRAN.R-project.org/package=randtoolbox
Cristianini, N., & John, S. T. (2000). An introduction to support vector machines and other kernel-based learning methods. Cambridge: Cambridge University Press.
Fang, K. T., Li, R. Z., & Sudjianto, A. (2006). Design and modeling for computer experiments. London: Chapman & Hall/CRC.
Fang, K. T., Liu, M. Q., Qin, H., et al. (2018). Theory and application of uniform experimental designs. New York: Springer.
Ginsbourger, D., Riche, R. L., & Carraro, L., et al. (2010). Kriging is well-suited to parallelize optimization. In L. M. Hiot, Y. S. Ong, & Y. Tenne (Eds.), Computational intelligence in expensive optimization problems (pp. 131–162). Berlin: Springer.
He, X. (2019). Sliced rotated sphere packing designs. Technometrics, 61(1), 66–76.
He, X., Tuo, R., & Wu, C. F. J. (2017). Optimization of multi-fidelity computer experiments via the EQIE criterion. Technometrics, 59(1), 58–68.
Huang, D., Allen, T., Notz, W., et al. (2006). Global optimization of stochastic black-box systems via sequential kriging meta-models. Journal of Global Optimization, 32, 441–466.
Hutter, F., Kotthoff, L., & Vanschoren, J. (2019). Automated machine learning—methods, systems, challenges. New York: Springer.

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

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

Krige, D. G. (1951). A statistical approach to some mine valuations and allied problems at the witwatersrand. Master’s thesis, University of Witwatersrand.

Lam, C., & Notz, W. (2008). Sequential adaptive designs in computer experiments for response surface model fit. *Statistics and Applications, 6*(2), 207–233.

L’Ecuyer, P., & Lemieux, C. (2002). Recent advances in randomized quasi-monte carlo methods. In M. Dror, P. L’Ecuyer, & F. Szidarovszky (Eds.), *Modeling uncertainty: An examination of stochastic theory, methods, and applications* (pp. 419–474). Boston: Kluwer Academic.

Locatelli, M. (1997). Bayesian algorithms for one-dimensional global optimization. *Journal of Global Optimization, 10*, 57–76.

Loopeky, J. L., Moore, L. M., & Williams, B. J. (2010). Batch sequential designs for computer experiments. *Journal of Statistical Planning and Inference, 140*(6), 1452–1464.

Loopeky, J. L., Sacks, J., & Welch, W. J. (2009). Choosing the sample size of a computer experiment: A practical guide. *Technometrics, 51*(4), 366–376.

Mak, S., & Joseph, V. R. (2017). Support points. *Annals of Statistics, 46*(6), 2562–2592.

Mehdad, E., & Kleijnen, J. P. (2018). Efficient global optimisation for black-box simulation via sequential intrinsic kriging. *Journal of the Operational Research Society, 69*(11), 1725–37.

Niederreiter, H. (1992). *Random number generation and quasi-Monte Carlo methods*. Philadelphia: Society for Industrial and Applied Mathematics.

Ooi, H., & Microsoft, Weston, S. (2019b). foreach: Provides foreach looping construct. R package version 1.4.7. https://CRAN.R-project.org/package=foreach

Ooi, H., Corporation, M., & Weston, S., et al. (2019a). doParallel: Foreach parallel adaptor for the ‘treparallel’ Package. R package version 1.0.15. https://CRAN.R-project.org/package=doParallel

Ponweiser, W., Wagner, T., & Vincze, M. (2008). Clustered multiple generalized expected improvement: A novel infill sampling criterion for surrogate models. *IEEE Congress on Evolutionary Computation, 1*, 3514–3521.

Quan, A. (2014). *Batch sequencing methods for computer experiments*. Doctor of Philosophy. Ohio State University, Ohio.

Roustant, O., Ginsbourger, D., & Deville, Y. (2012). DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by kriging-based metamodeling and optimization. *Journal of Statistical Software, 51*(1), 1–55.

Sacks, J., Welch, W. J., Mitchell, T. J., et al. (1989). Design and analysis of computer experiments (with discussion). *Statistical Science, 4*(4), 409–435.

Santner, T., Williams, B., & Notz, W. (2018). *The design and analysis of computer experiments* (2nd ed.). New York: Springer.

Schonlau, M. (1997). *Computer experiments and global optimization*. PhD thesis, University of Waterloo, Waterloo.

Surjanovic, S., & Bingham, D. (2013). Virtual library of simulation experiments: Test functions and datasets. Retrieved August 30, 2021, from http://www.sfu.ca/~ssurjanov

Viana, F. A. C., Haftka, R. T., & Watson, L. T. (2013). Efficient global optimization algorithm assisted by multiple surrogate techniques. *Journal of Global Optimization, 56*(2), 669–689.

Zhang, A., Li, H., & Quan, S., et al. (2018). UniDOE: Uniform Design of Experiments. R package version 1.0.2. https://CRAN.R-project.org/package=UniDOE

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.