A Multicriteria Generalization of Bayesian Global Optimization

Michael Emmerich, Kaifeng Yang, André Deutz, Hao Wang, and Carlos M. Fonseca

Abstract This chapter discusses a generalization of the expected improvement used in Bayesian global optimization to the multicriteria optimization domain, where the goal is to find an approximation to the Pareto front. The expected hypervolume improvement (EHVI) measures improvement as the gain in dominated hypervolume relative to a given approximation to the Pareto front. We will review known properties of the EHVI, applications in practice and propose a new exact algorithm for computing EHVI. The new algorithm has asymptotically optimal time complexity $O(n \log n)$. This improves existing computation schemes by a factor of $n/\log n$. It shows that this measure, at least for a small number of objective functions, is as fast as other simpler measures of multicriteria expected improvement that were considered in recent years.

Introduction

In the 1970s several new ideas for global optimization were proposed. Among these the idea of *Bayesian Global Optimization* (BGO) was proposed by the Lithuanian research group Jonas Mockus and Antanas Žilinskas [14, 15, 21, 25]. It had a lasting impact on the development of both deterministic and stochastic global optimization techniques. Today variations of this idea are known under various names, such as

M. Emmerich • K. Yang • A. Deutz (✉) • H. Wang
Multiobjective Optimization and Decision Analysis (MODA) Research Group, LIACS, Faculty of Science, Leiden University, Leiden, The Netherlands
e-mail: m.t.m.emmerich@liacs.leidenuniv.nl; k.yang@liacs.leidenuniv.nl; a.h.deutz@liacs.leidenuniv.nl; h.wang@liacs.leidenuniv.nl
http://moda.liacs.nl

C.M. Fonseca
CISUC, Department of Informatics Engineering, University of Coimbra, Polo II, Pinhal de Marrocos, P-3030 290 Coimbra, Portugal
e-mail: cmfonsec@dei.uc.pt

Efficient Global Optimization [9] or *Expected Improvement Algorithm* [22]. In these techniques the goal is to find the extremum of a function $f : \mathscr{X} \rightarrow \mathbb{R}$ where $\mathscr{X}$ is a compact subspace of $\mathbb{R}^d$. BGO assumes that the objective function is the realization of a Gaussian random field. This random field can be conditioned by the knowledge of $f(\mathbf{x}^{(i)})$ at some points $\mathbf{x}^{(i)} \in \mathscr{X}, i = 1, \ldots, n$. Under this assumption, measures such as the expected improvement of a new design point are well defined, and can be used to guide search towards the global optimum.

In this chapter we describe a generalization of this approach to multicriteria optimization. It iteratively evaluates points from $\mathscr{X}$ and finds a well distributed subset of the Pareto front of a multicriteria optimization problem. The algorithm is based on a generalization of the expected improvement, which is based on the hypervolume indicator, the so-called *Expected Hypervolume Improvement* (EHVI) [3]. It has attractive theoretical properties [23], but so far its computation time was considered to be expensive. In this chapter it is shown that, for bicriteria optimization, a fast algorithm exists for computing EHVI that has only linear time complexity in the size of the intermediate approximation to the Pareto front, given that the Pareto front is given as a sorted set. It is shown that this algorithm has asymptotically optimal time complexity.

This chapter is organized as follows: section "Bayesian Global Optimization" introduces the framework of BGO. Section "Multicriteria Optimization" shows how this framework can be generalized to multicriteria optimization. Section "Expected Hypervolume Improvement" defines the EHVI, discusses some of its theoretical properties, and reviews recent applications of it. Section "Efficient Exact Computation" outlines the new, asymptotically efficient algorithm for its exact computation and proves that it has an asymptotically optimal time complexity for bicriteria problems. A numerical example is discussed in section "Numerical Example". Section "Application Notes and Further Reading" points to some recent applications and related work. Finally, section "Summary and Outlook", concludes with a summary and discusses open questions.

Bayesian Global Optimization

In BGO the goal is to solve d-dimensional global optimization problems of the type: Find $\mathbf{x}^*$ with

$$\mathbf{x}^* \in \arg \min_{\mathbf{x} \in \mathscr{X}} f(\mathbf{x}), \mathscr{X} = [\mathbf{x}_{min}, \mathbf{x}_{max}] \subset \mathbb{R}^d \qquad (1)$$

(Without loss of generality we consider minimization only.)

In order to do so, a sequence $\{\mathbf{x}^{(t)}\}_{t=1,2,\ldots}$ of points is computed such that

$$\mathbf{x}^{(t)} \in \arg \max_{\mathbf{x} \in S} (\mathrm{E}(I(\mathbf{x}) | (\mathbf{x}^{(1)}, f(\mathbf{x}^{(1)})), \ldots, (\mathbf{x}^{(t-1)}, f(\mathbf{x}^{(t-1)}))) \qquad (2)$$

Here $\mathrm{E}(I(\mathbf{x}) | (\mathbf{x}^{(1)}, f(\mathbf{x}^{(1)})), \ldots, (\mathbf{x}^{(t-1)}, f(\mathbf{x}^{(t-1)}))$ denotes the expected improvement measure that measures how promising the new point $\mathbf{x}$ is, given $t-1$ previous

evaluations of f at $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(t-1)}$. This expected improvement is an expected value of a random variable, here called $I(\mathbf{x})$ that requires further explanation.

In BGO one makes the assumption that the function f is the realization of a Gaussian random field $\mathbf{F}$. A Gaussian random field is an infinite set of random variables. Each random variable in F is identified by its spatial index $\mathbf{x} \in \mathbb{R}^d$. We will denote it with $\mathbf{F_x}$. It is assumed that the random variables share the same global mean value β and global variance s^2. Moreover, a correlation $\rho(\mathbf{F_u}, \mathbf{F_v})$ is defined for every pair of indices $\mathbf{u} \in \mathbb{R}^d$ and $\mathbf{v} \in \mathbb{R}^d$. This correlation depends on the relation between $\mathbf{u}$ and $\mathbf{v}$. A typical family of correlation functions is

$$\rho(\mathbf{F_u}, \mathbf{F_v}) = \exp\left(-\sum_{i=1}^{d} \theta_i |u_i - v_i|^{q_i}\right)$$

It is important that this correlation function is positive definite. It obtains the value of 1, if $\mathbf{v} = \mathbf{u}$ and gets smaller with increasing distance between $\mathbf{v}$ and $\mathbf{u}$. The parameters q_i and θ_i are either set by the user or obtained from data fitting. The parameters θ_i are positive.

The Gaussian random field can be viewed as a multivariate Gaussian distribution of infinite dimension. We can use well-known expressions for the marginal distributions of the multivariate distribution to find the conditional distribution, given that some of the realizations of one dimensional random variables are known. That is, given the prior information $\mathbf{F}_{\mathbf{x}^{(1)}} = f(\mathbf{x}^{(1)}), \ldots, \mathbf{F}_{\mathbf{x}^{(t-1)}} = f(\mathbf{x}^{(t-1)})$ we can compute the parameters μ (conditional mean) and σ^2 (conditional variance) of the conditioned random variable:

$$\mathbf{F_x} \mid \mathbf{F}_{\mathbf{x}^{(1)}} = f(\mathbf{x}^{(1)}), \ldots, \mathbf{F}_{\mathbf{x}^{(t-1)}} = f(\mathbf{x}^{(t-1)}) \tag{3}$$

As a shortcut we will denote this random variable with $\mathbf{F_x} \mid \mathbf{X}, f(\mathbf{X})$, where $\mathbf{X} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(t-1)})$ denote the indices for which we know realizations and the values of the corresponding realizations are abbreviated with

$$f(\mathbf{X}) = \left(f(\mathbf{x}^{(1)}), \ldots, f(\mathbf{x}^{(t-1)})\right)$$

The estimation of hyperparameters θ_i and q_i, $i = 1, \ldots, d$, of the correlation function, as well as the global variance and mean can be accomplished by maximum likelihood methods. For details on the computations of the parameters of the conditional distribution we refer to the specialized literature [19].

Now, the expected improvement can be defined: The improvement of a function value $y \in \mathbb{R}$ is defined as

$$I(y) = \max\{0, y_{min} - y)\} \tag{4}$$

where $y_{min} = \min\{f(\mathbf{x}^{(1)}), \ldots, f(\mathbf{x}^{(t-1)})\}$. Then the *expected improvement* is defined as

$$E(I(\mathbf{F_x}|\mathbf{X}, f(\mathbf{X}))) = \int_{y=-\infty}^{y_{min}} I(y) PDF_{\mathbf{x}|\mathbf{X}, f(\mathbf{X})}(y) dy \tag{5}$$

Here $PDF_{\mathbf{x}|\mathbf{X}, f(\mathbf{X})}$ is the probability density function of $\mathbf{F_x}|\mathbf{X}, f(\mathbf{X})$.

Multicriteria Optimization

A continuous m-dimensional multicriteria optimization problem is a problem where multiple objective functions, say $f_1 : \mathscr{X} \to \mathbb{R}, \ldots, f_m : \mathscr{X}^m \to \mathbb{R}$, are to be minimized simultaneously, $\mathscr{X} \subseteq \mathbb{R}^m$.

In the a posteriori approach to multicriteria optimization an approximation to the Pareto front of the problem is computed first. Based on this, the trade-off is analyzed and a solution is selected by the decision maker. To define a Pareto front, we introduce the Pareto dominance order $\prec$ on $\mathbb{R}^m$, with $\forall \mathbf{y}, \mathbf{z} \in \mathbb{R}^m : \mathbf{y} \prec \mathbf{z} \Leftrightarrow (\forall i \in \{1, \ldots, m\} : y_i \leq z_i)$ and $\mathbf{y} \neq \mathbf{z}$. The non-dominated subset of a multiset of vectors $Y = \{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(m)}\}$ is defined as $nd(Y) = \{\mathbf{y} \in Y \mid \nexists \mathbf{z} \in Y : \mathbf{z} \prec \mathbf{y}\}$. Given a multicriteria optimization problem, the image of $\mathscr{X}$ is defined as $\mathscr{Y} = \{\mathbf{f}(\mathbf{x}) \mid \mathbf{x} \in \mathscr{X}\}$. The Pareto front of a multicriteria optimization problem is defined as $\mathscr{Y}_{nd} := nd(\mathscr{Y})$. An important special case is bicriteria optimization, where $m = 2$.

One way to generalize the BGO algorithm is to compute the expected improvement of the hypervolume indicator. The hypervolume indicator is the m-dimensional Lebesgue measure λ_m of the dominated subspace limited from above by some reference point $\mathbf{r}^m$. More precisely the hypervolume indicator is defined as

$$hv(Y) = \lambda_m \left(\{\mathbf{y} \in \mathbb{R}^m \mid \exists \mathbf{z} \in Y : \mathbf{z} \prec \mathbf{y} \wedge \mathbf{y} \prec \mathbf{r}\}\right) = \lambda_m \left(\bigcup_{\mathbf{y} \in Y} [\mathbf{y}, \mathbf{r}]\right) \tag{6}$$

In Fig. 1 the hypervolume indicator is illustrated for a Pareto front approximation with nine points and two objective functions ($m = 2$). Given a problem with a Pareto front bounded above by the reference point, sets that maximize the hypervolume indicator are well distributed subsets of the Pareto front [1]. This is why finding the Pareto front is sometimes recast as the problem of maximizing the hypervolume indicator over the set of all subsets of $\mathscr{X}$. We will call this problem *hypervolume maximization*.

Expected Hypervolume Improvement

For hypervolume maximization problems the generalization of the improvement function is straightforward. We generalize the best solution found up to iteration $t-1$, namely $y_{min,t-1}$, to

$$Y_{nd,t-1} = nd\left(\{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(t-1)}\}\right) \tag{7}$$
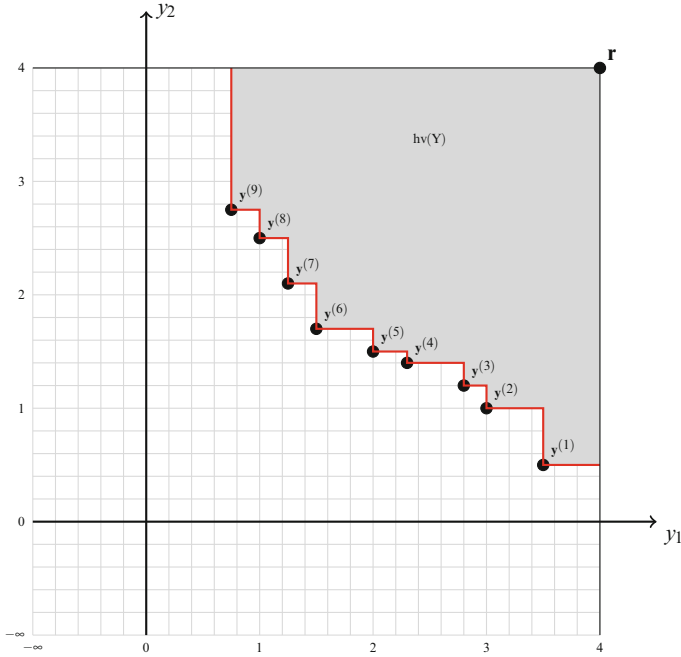
Fig. 1 Hypervolume indicator of Pareto front approximation

The improvement function is generalized by the following definition of an m-dimensional improvement:

$$I_m(\mathbf{y}, Y_{nd,t-1}, \mathbf{r}) := hv\left(Y_{nd,t-1} \cup \{\mathbf{f}(\mathbf{x})\}\right) - hv(Y_{nd,t-1}) \tag{8}$$

It is easy to show that this I_m specializes to the improvement function in one dimension, if we chose r_1 to be sufficiently large.

In order to compute the *expected* improvement in the multicriteria case, we need also to generalize the assumption on the Gaussian random field. For this, we consider one Gaussian random field per objective function and assume that there is no correlation between random variables from different random fields. For every point $\mathbf{x} \in \mathcal{X}$ we obtain an m-dimensional random variable conditioned on previous information, that is given by $\mathbf{X}$ and $\mathbf{f}(\mathbf{X}) = (\mathbf{f}(\mathbf{x}^{(1)}), \ldots, \mathbf{f}(\mathbf{x}^{(t-1)}))$.

The resulting EHVI can be denoted with

$$E\left(I_m\left((F_1(\mathbf{x}), \ldots, F_m(\mathbf{x}))\right) \mid \mathbf{X}, \mathbf{f}(\mathbf{X}), Y_{nd,t-1}, \mathbf{r}\right) =$$

$$\int_{\mathbf{y} \in \mathbb{R}^{\mathbf{m}}} I_m(\mathbf{y}, Y_{nd,t-1}, \mathbf{r}) PDF_{\mathbf{x}|\mathbf{X}, \mathbf{f}(\mathbf{X})}(\mathbf{y}) d\mathbf{y} \tag{9}$$

and it is a generalization of the single objective expected improvement, if we consider $y_{min,0} = r_1$.

Efficient Exact Computation

In this section the problem of computing the EHVI is studied and a new, efficient algorithm for bicriteria optimization will be derived. Fast algorithms for computing the EHVI are important, because in BGO a large number evaluations of the EHVI are performed in each iteration when searching for its maximizer. Although the BGO algorithm is typically used in the context of expensive function evaluations, the optimization of the EHVI can significantly contribute to the total running time of the algorithm. For instance, this was recently reported as a major drawback of using EHVI in [12], even when considering only the two dimensional case.

A simplified notation will be used in the following. It focuses only on the elements that are relevant for the EHVI computation.

Symbol	Type	Description
μ	$\mathbb{R}^m$	Mean values of predictive distribution
σ	$(\mathbb{R}_0^+)^m$	Standard deviations of predictive distribution
Y	$(\mathbb{R}^m)^n$	Sequence of mutually non-dominated points (Pareto front approximation in $t-1$)
$\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(n)}$	$\mathbb{R}^m$	The vectors in Y
$\mathbf{r}$	$\mathbb{R}^m$	Reference point

For computing integrals of the expected improvement it is useful to define the function Δ. For a given vector of objective function values $\mathbf{y} \in \mathbb{R}^m$, $\Delta(\mathbf{y}, Y, \mathbf{r})$ is the subset of the vectors in $\mathbb{R}^m$ which are exclusively dominated by a vector $\mathbf{y}$ and not by elements in Y and that dominate the reference point, in symbols

$$\Delta(\mathbf{y}, Y, \mathbf{r}) = \lambda_m \{\mathbf{z} \in \mathbb{R} \mid \mathbf{y} \prec \mathbf{z} \text{ and } \mathbf{z} \prec \mathbf{r} \text{ and } \not\exists \mathbf{q} \in Y : \mathbf{q} \prec \mathbf{z}\} \qquad (10)$$

In order to simplify notation, we will write $\Delta(\mathbf{y})$ whenever $Y, \mathbf{r}$ are given by the context.

Based on this, we can now concisely (re-)define the EHVI function as

$$\text{EHVI}(\mu, \sigma, Y, \mathbf{r}) = \int_{y_1=-\infty}^{\infty} \cdots \int_{y_m=-\infty}^{\infty} \lambda_m(\Delta(\mathbf{y})) \text{PDF}_{\mu,\sigma}(\mathbf{y}) dy_1 \ldots dy_m \qquad (11)$$

Example 1. An illustration of the EHVI is displayed in Fig. 2. The light gray area is the dominated subspace of $Y = \{\mathbf{y}^{(1)} = (3,1)^\top, \mathbf{y}^{(2)} = (2,1.5)^\top, \mathbf{y}^{(3)} = (1,2.5)^\top\}$ cut by the reference point $r = (4,4)^\top$. The bivariate Gaussian distribution has the parameters $\mu_1 = 2$, $\mu_2 = 1.5$, $\sigma_1 = 0.7$, and $\sigma_2 = 0.6$. The PDF of the bivariate Gaussian distribution is indicated as a 3-D plot. Here $\mathbf{y}$ is a sample from this distribution, and the area of improvement relative to Y is indicated by the dark shaded area. The variable y_1 stands for the f_1 value and y_2 for the f_2 value.
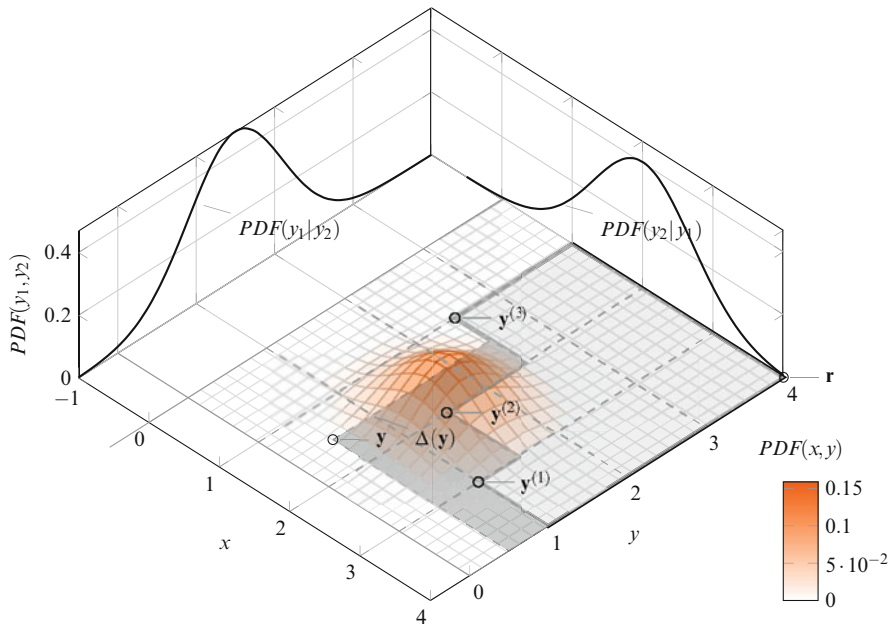
Fig. 2 Expected hypervolume improvement in 2-D (cf. Example 1)

State of the Art

To compute the EHVI (9) Monte Carlo integration is suggested in [3, 4]. Exact algorithms for computing EHVI for $m = 2$ are derived in [5] and for $m > 2$ in [2]. A different algorithm is described in [7].

Fast algorithms have been proposed in [2] and even faster algorithms for $m = 2, 3$ in [8]. So far the best known bounds for the time complexity of exact computations are $O(n^2)$ for $m = 2$, and $O(n^3)$ for $m = 3$. It is notable that the number of transcendental function evaluations scales only linearly in n in the algorithm presented in [8]. A lower bound of $\Omega(n \log n)$ is provided for unsorted Y. However, it makes sense to assume that Y is sorted in the first coordinate. In that case, as will be shown, a lower bound of $\Omega(n)$ still holds. None of the algorithms found so far for EHVI reach these lower bounds. In this paper we will present an algorithm for $m = 2$ that does so.

Next an algorithm is outlined that reaches the lower bound time complexity of $\Omega(n \log n)$. We thereby prove that the time complexity of EHVI is $\Theta(n \log n)$. However, this complexity stems from the complexity that is inherent to sorting Y by the first coordinate.

To keep Y sorted in the first coordinate requires an effort of amortized time complexity $O(\log n)$ per iteration. It makes therefore sense to assume a sorted Y. For this case we can show that the time complexity is $\Theta(n)$. To do so, we will first establish a lower bound of $\Omega(n)$ for this case:

Lemma 1. *The computational time complexity of computing the EHVI for a set Y that is sorted by the first coordinate is bounded from below by $\Omega(n)$.*

Proof. An adversary argument can be used to prove this statement. The algorithm has to "look at" all n points. If one point is not used, it could be moved by an adversary and this move will not be noticed by the algorithm; a move of any single point can, in general, change the EHVI. $\square$

Efficient Algorithm

For $m = 2$ the expected improvement can be computed in linear time, given that Y is already sorted by the first coordinate. Next, a formula will be derived that consists of $n + 1$ integrals, each of which can be solved in constant time.

The starting point of the derivation is to partition the objective space into $n + 1$ disjoint rectangular stripes $S_1, \ldots, S_{n+1}$, as indicated in Fig. 3 (left). In order to define the stripes formally, augment Y with two sentinels: $\mathbf{y}^{(0)} = (r_1, -\infty)$ and $\mathbf{y}^{(n+1)} = (-\infty, r_2)$. The stripes are now defined by

$$S_i = \left(\begin{pmatrix} y_1^{(i)} \\ -\infty \end{pmatrix}, \begin{pmatrix} y_1^{(i-1)} \\ y_2^{(i)} \end{pmatrix} \right), i = 1, \ldots, n+1$$

We can now express the improvement of a point $\mathbf{y} \in \mathbb{R}^2$ by

$$I_2(\mathbf{y}, Y, \mathbf{r}) = \sum_{i=1}^{n+1} \lambda_2[S_i \cap \Delta(\mathbf{y})] \tag{12}$$
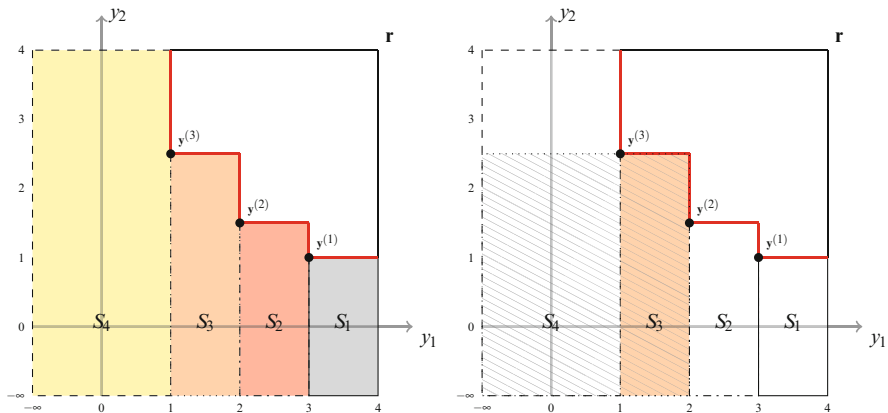
Fig. 3 *Left*: partitioning of the integration region into stripes. *Right*: new partitioning of the reduced integration region after first iteration of the algorithm

This gives rise to the compact integral for the original EHVI, $\mathbf{y} = (y_1, y_2)$:

$$\text{EHVI}(\mu, \sigma, Y, \mathbf{r}) = \int_{y_1=-\infty}^{\infty} \int_{y_2=-\infty}^{\infty} \sum_{i=1}^{n+1} \lambda_2[S_i \cap \Delta(y_1, y_2)] \cdot PDF_{\mu,\sigma}(\mathbf{y}) d\mathbf{y} \qquad (13)$$

It is observed that the intersection of S_i with $\Delta(y_1, y_2)$ is non-empty if and only if $\mathbf{y} = (y_1, y_2)$ dominates the upper right corner of S_i. In other words, if and only if $\mathbf{y}$ is located in the rectangle with lower left corner $(-\infty, -\infty)$ and upper right corner $(y_1^{(i-1)}, y_2^{(i)})$. See Fig. 3 (right) for an illustration. Therefore

$$\text{EHVI}(\mu, \sigma, Y, \mathbf{r}) = \sum_{i=1}^{n+1} \int_{y_1=-\infty}^{y_1^{(i-1)}} \int_{y_2=-\infty}^{y_2^{(i)}} \lambda_2[S_i \cap \Delta(y_1, y_2)] \cdot PDF_{\mu,\sigma}(\mathbf{y}) d\mathbf{y} \qquad (14)$$

In (14) also the summation is done after integration. This is allowed, because integration is a linear mapping.

Details of the Constant Time Integration

$$\text{EHVI}(\mu, \sigma, Y, \mathbf{r}) = \sum_{i=1}^{n+1} \int_{y_1=-\infty}^{y_1^{(i-1)}} \int_{y_2=-\infty}^{y_2^{(i)}} \lambda_2[S_i \cap \Delta(y_1, y_2)] \cdot PDF_{\mu,\sigma}(\mathbf{y}) d\mathbf{y} \qquad (15)$$

$$= \sum_{i=1}^{n+1} \int_{y_1=-\infty}^{y_1^{(i)}} \int_{y_2=-\infty}^{y_2^{(i)}} \lambda_2[S_i \cap \Delta(y_1, y_2)] \cdot PDF_{\mu,\sigma}(\mathbf{y}) d\mathbf{y} +$$

$$\sum_{i=1}^{n+1} \int_{y_1=y^{(i)}}^{y_1^{(i-1)}} \int_{y_2=-\infty}^{y_2^{(i)}} \lambda_2[S_i \cap \Delta(y_1, y_2)] \cdot PDF_{\mu,\sigma}(\mathbf{y}) d\mathbf{y}.$$

$$(16)$$

Recall the definition of the standard Gaussian PDF and CDF: $\phi(x) = \frac{1}{\sqrt{2\pi}}\exp(-x^2/2)$, $\Phi(x) = \frac{1}{2}(1 + \text{erf}(\sqrt{2}))$, and a function Ψ that was defined in [8] as follows:

$$\Psi(a, b, \mu, \sigma) = \int_{-\infty}^{b} (a - z) \frac{1}{\sigma} \phi\left(\frac{z - \mu}{\sigma}\right) dz.$$

Moreover it can be shown that

$$\Psi(a,b,\mu,\sigma) = \int_{-\infty}^{b} (a-z)\frac{1}{\sigma}\phi\left(\frac{z-\mu}{\sigma}\right)dz \quad = \sigma\phi\left(\frac{b-\mu}{\sigma}\right) + (a-\mu)\Phi\left(\frac{b-\mu}{\sigma}\right).$$

Then the first summand of (16) can be written as follows:

$$= \sum_{i=1}^{n+1} \int_{y_1=-\infty}^{y_1^{(i)}} \int_{y_2=-\infty}^{y_2^{(i)}} \lambda_2[S_i \cap \Delta(y_1,y_2)] \cdot PDF_{\mu,\sigma}(\mathbf{y})d\mathbf{y},$$

$$= \sum_{i=1}^{n+1} \int_{y_1=-\infty}^{y_1^{(i)}} (y_1^{(i-1)} - y_1^{(i)}) \cdot PDF_{\mu_1,\sigma_1}(y_1)dy_1 \int_{y_2=-\infty}^{y_2^{(i)}} (y_2^{(i)} - y_2) \cdot PDF_{\mu_2,\sigma_2}(y_2)dy_2,$$

$$= \sum_{i=1}^{n+1} (y_1^{(i-1)} - y_1^{(i)}) \int_{y_1=-\infty}^{y_1^{(i)}} PDF_{\mu_1,\sigma_1}(y_1)dy_1 \int_{y_2=-\infty}^{y_2^{(i)}} (y_2^{(i)} - y_2) \cdot PDF_{\mu_2,\sigma_2}(y_2)dy_2,$$

$$= \sum_{i=1}^{n+1} (y_1^{(i-1)} - y_1^{(i)}) \cdot \Phi\left(\frac{y_1^{(i)} - \mu_1}{\sigma_1}\right) \cdot \Psi(y_2^{(i)}, y_2^{(i)}, \mu_2, \sigma_2). \tag{17}$$

And the second summand of (16) can be written as follows:

$$= \sum_{i=1}^{n+1} \int_{y_1=y^{(i)}}^{y_1^{(i-1)}} \int_{y_2=-\infty}^{y_2^{(i)}} \lambda_2[S_i \cap \Delta(y_1,y_2)] \cdot PDF_{\mu,\sigma}(\mathbf{y})d\mathbf{y},$$

$$= \sum_{i=1}^{n+1} \int_{y_1=y^{(i)}}^{y_1^{(i-1)}} (y_1^{(i-1)} - y_1) \cdot PDF_{\mu_1,\sigma_1}(y_1)dy_1 \cdot \int_{y_2=-\infty}^{y_2^{(i)}} (y^{(i)} - y_2) \cdot PDF_{\mu_2,\sigma_2}(y_2)dy_2,$$

$$= \sum_{i=1}^{n+1} \left(\Psi(y_1^{(i-1)}, y_1^{(i-1)}, \mu_1, \sigma_1) - \Psi(y_1^{(i-1)}, y_1^{(i)}, \mu_1, \sigma_1)\right) \cdot \Psi(y_2^{(i)}, y_2^{(i)}, \mu_2, \sigma_2). \tag{18}$$

The C++ and MATLAB source-code for computing the EHVI is made available under http://moda.liacs.nl or on request by the authors. The code has been compared to results of Monte Carlo integration and earlier implementations of the exact EHVI.

Numerical Example

The behavior of the BGO based on the EHVI will be illustrated by a single numerical experiment.

The numerical example is visualized in the plots of Fig. 4. The bicriteria optimization problem is: $f_1(\mathbf{x}) = ||\mathbf{x} - \mathbf{1}|| \to \min$, $f_2(\mathbf{x}) = ||\mathbf{x} + \mathbf{1}|| \to \min$, and $\mathbf{x} \in [-2,2] \times [-2,2] \subset \mathbb{R}^2$. The Pareto front is the line segment from $(0, 2 \cdot \sqrt{2})$ to
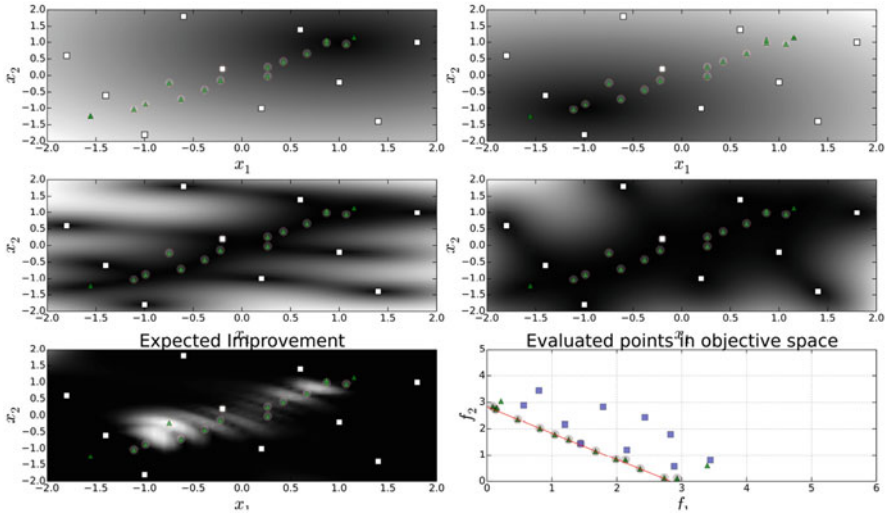
Fig. 4 Example run of multicriteria Bayesian global optimization

$(2 \cdot \sqrt{2}, 0)$, the efficient set is the line segment that connects $(-1, -1)$ and $(1, 1)$. The metamodel used is a Gaussian random field model with Gaussian correlation function $\exp(-\theta||\mathbf{x}^{(1)} - \mathbf{x}^{(2)}||^2)$, for $\mathbf{x}^{(1)} \in \mathbb{R}^m$ and $\mathbf{x}^{(2)} \in \mathbb{R}^m$. We set $\theta = 0.0001$, which was estimated by maximum likelihood method for initial sample. An initial set of 10 points was evaluated, indicated by the dark blue squares. From this starting set 15 new points were generated using the EHVI. The maximizer of the expected improvement was found using a uniform grid. In total each objective function was evaluated 25 times.

The results of the experiment are depicted in plots. In all pictures, points that have been evaluated are indicated by triangles. The points from the initial set are additionally marked by squares. Efficient points are surrounded by circles. The top row depicts the mean value of the Gaussian random field model at $\mathbf{x} \in [-2, 2] \times [-2, 2]$ for f_1 and f_2, resp. Likewise, the middle row depicts the variance of the Gaussian random field model at $\mathbf{x} \in [-2, 2] \times [-2, 2]$ for f_1 and f_2, resp. On the left-hand side of the bottom row the hypervolume-based expected improvement values after 25 iterations are shown. The final set of points in the objective space and the Pareto front approximation is seen in the plot in the lower right corner. Using only *25 evaluations of the original objective functions*, the algorithm finds a good approximation to the Pareto front.

Application Notes and Further Reading

In addition to this experiment, other applications of the EHVI have been recently reported. It was first used as selection criterion in evolutionary optimization [3] and in the context of airfoil design [4] and quantum control [18]. To our knowledge, it was used for the first time in BGO in the context of airfoil optimization in [13] and conceptually compared other multicriteria infill criteria, including proposal made in [11], [10], and in [23]. Other applications are robotics [20], biogas plant controllers [6], event detection in water quality management [24], structural design optimization [17], and tuning of machine learning tools [12]. An empirical comparison with other infill criteria is found in [16].

Summary and Outlook

This chapter described the EHVI as a multicriteria generalization of the expected improvement used in BGO. This generalization is based on the hypervolume indicator, which is a quality indicator for Pareto front approximations. It has recently served as an infill criterion in a number of BGO case studies, but was criticized for its high computational complexity. In this chapter, the time complexity of the 2-D EHVI was shown to be only $\Theta(n)$. The linear time algorithm presented in this paper improves upon previously proposed algorithms which required quadratic time complexity. It assumes a sorted Pareto front (otherwise its complexity is $O(n \log n)$), which is typically given in BGO. During a single iteration of BGO a large number of evaluations need to be performed, in order to find a minimizer based on the Gaussian random field model. Therefore the fast algorithm will be of great benefit for reducing the running time of multicriteria BGO based on EHVI.

Future research will investigate in more depth the theoretical properties of the EHVI. For the first results in this direction refer to [5], where it was shown that the 2-D EHVI is monotonic in the mean values and variance. Also it will be interesting to analyze the time complexity of EHVI for more than two objective functions.

References

1. Auger, A., Bader, J., Brockhoff, D., Zitzler, E.: Theory of the hypervolume indicator: optimal μ-distributions and the choice of the reference point. In: Proceedings of the Tenth ACM SIGEVO Workshop on Foundations of Genetic Algorithms, pp. 87–102. ACM, Chicago (2009)
2. Couckuyt, I., Deschrijver, D., Dhaene, T.: Fast calculation of multiobjective probability of improvement and expected improvement criteria for Pareto optimization. J. Global Optim. **60**(3), 575–594 (2014)

3. Emmerich, M.: Single-and multi-objective evolutionary design optimization assisted by Gaussian random field metamodels. Ph.D. thesis, Fachbereich Informatik, Chair of Systems Analysis, University of Dortmund (2005)
4. Emmerich, M., Giannakoglou, K.C., Naujoks, B.: Single-and multiobjective evolutionary optimization assisted by Gaussian random field metamodels. IEEE Trans. Evol. Comput. **10**(4), 421–439 (2006)
5. Emmerich, M., Deutz, A.H., Klinkenberg, J.W.: Hypervolume-based expected improvement: monotonicity properties and exact computation. In: 2011 IEEE Congress on Evolutionary Computation (CEC), pp. 2147–2154. IEEE, New Jersey (2011)
6. Gaida, D.: Dynamic real-time substrate feed optimization of anaerobic co-digestion plants. Ph.D. thesis, Leiden Institute of Advanced Computer Science (LIACS), Faculty of Science, Leiden University (2014)
7. Hupkens, I., Emmerich, M., Deutz, A.: Faster computation of expected hypervolume improvement. arXiv preprint arXiv:1408.7114 (2014)
8. Hupkens, I., Deutz, A., Yang, K., Emmerich, M.: Faster exact algorithms for computing expected hypervolume improvement. In: Evolutionary Multi-Criterion Optimization, pp. 65–79. Springer, Berlin, Heidelberg (2015)
9. Jones, D.R., Schonlau, M., Welch, W.J.: Efficient global optimization of expensive black-box functions. J. Global Optim. **13**(4), 455–492 (1998)
10. Keane, A.J.: Statistical improvement criteria for use in multiobjective design optimization. AIAA J. **44**(4), 879–891 (2006)
11. Knowles, J.: ParEGO: a hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems. IEEE Trans. Evol. Comput. **10**(1), 50–66 (2006)
12. Koch, P., Wagner, T., Emmerich, M.T., Bäck, T., Konen, W.: Efficient multi-criteria optimization on noisy machine learning problems. Appl. Soft Comput. **29**, 357–370, New Jersey (2015)
13. Łaniewski-Wołłk, Ł., Obayashi, S., Jeong, S.: Development of expected improvement for multi-objective problem. In: Proceedings of 42nd Fluid Dynamics Conference/Aerospace Numerical Simulation Symposium (2010)
14. Mockus, J.: Bayesian Approach to Global Optimization: Theory and Applications, vol. 37. Springer Science & Business Media, New York (2012)
15. Mockus, J., Tiesis, V., Žilinskas, A.: The application of Bayesian methods for seeking the extremum. In: Towards Global Optimization, vol. 2, pp. 117–129. North-Holland, Amsterdam (1978)
16. Shimoyama, K., Sato, K., Jeong, S., Obayashi, S.: Comparison of the criteria for updating Kriging response surface models in multi-objective optimization. In: 2012 IEEE Congress on Evolutionary Computation (CEC), pp. 1–8. IEEE, New Jersey (2012)
17. Shimoyama, K., Sato, K., Jeong, S., Obayashi, S.: Updating Kriging surrogate models based on the hypervolume indicator in multi-objective optimization. J. Mech. Des. **135**(9), 094503 (2013)
18. Shir, O.M., Emmerich, M., Bäck, T., Vrakking, M.J.: The application of evolutionary multi-criteria optimization to dynamic molecular alignment. In: IEEE Congress on Evolutionary Computation, 2007, CEC 2007, pp. 4108–4115. IEEE, New Jersey (2007)
19. Stein, M.L.: Interpolation of Spatial Data: Some Theory for Kriging. Springer Science & Business Media, New York (2012)
20. Tesch, M., Schneider, J., Choset, H.: Adapting control policies for expensive systems to changing environments. In: 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 357–364. IEEE, New Jersey (2011)
21. Törn, A., Žilinskas, A.: Global Optimization. Springer, New York (1989)
22. Vazquez, E., Bect, J.: Convergence properties of the expected improvement algorithm with fixed mean and covariance functions. J. Stat. Plan. Inference **140**(11), 3088–3095 (2010)

23. Wagner, T., Emmerich, M., Deutz, A., Ponweiser, W.: On expected-improvement criteria for model-based multi-objective optimization. In: Parallel Problem Solving from Nature. PPSN XI, pp. 718–727. Springer, Berlin, Heidelberg (2010)
24. Zaefferer, M., Bartz-Beielstein, T., Naujoks, B., Wagner, T., Emmerich, M.: A case study on multi-criteria optimization of an event detection software under limited budgets. In: Evolutionary Multi-Criterion Optimization, pp. 756–770. Springer, Berlin, Heidelberg (2013)
25. Žilinskas, A., Mockus, J.: On one Bayesian method of search of the minimum. Avtomatika i Vychislitel'naya Teknika **4**, 42–44 (1972)