Lecture Notes in Computer Science

CONFERENCE PAPER · JANUARY 2011

DOI: 10.1007/978-3-642-19893-9_2 · Source: DBLP

3 AUTHORS, INCLUDING:

Tobias Wagner
Technische Universität Dor…

51 PUBLICATIONS **415**
CITATIONS

SEE PROFILE

Heike Trautmann
University of Münster

51 PUBLICATIONS **284**
CITATIONS

SEE PROFILE

A Taxonomy of Online Stopping Criteria for Multi-Objective Evolutionary Algorithms

Tobias Wagner[1], Heike Trautmann[2], and Luis Martí[3]

[1] Institute of Machining Technology (ISF), TU Dortmund
Baroper Straße 301, 44227 Dortmund, Germany
wagner@isf.de
http://www.isf.de
[2] Department of Computational Statistics, TU Dortmund
Vogelpothsweg 87, 44227 Dortmund, Germany
trautmann@statistik.tu-dortmund.de
http://www.statistik.tu-dortmund.de
[3] Group of Applied Artificial Intelligence, Universidad Carlos III de Madrid
Av. de la Universidad Carlos III 22, 28270 Colmenarejo, Madrid, Spain
lmarti@inf.uc3m.es
http://www.giaa.inf.uc3m.es/

Abstract. The use of multi-objective evolutionary algorithms for solving black-box problems with multiple conflicting objectives has become an important research area. However, when no gradient information is available, the examination of formal convergence or optimality criteria is often impossible. Thus, sophisticated heuristic online stopping criteria (OSC) have recently become subject of intensive research. In order to establish formal guidelines for a systematic research, we present a taxonomy of OSC in this paper. We integrate the known approaches within the taxonomy and discuss them by extracting their building blocks. The formal structure of the taxonomy is used as a basis for the implementation of a comprehensive MATLAB toolbox. Both contributions, the formal taxonomy and the MATLAB implementation, provide a framework for the analysis and evaluation of existing and new OSC approaches.

Keywords: Convergence Detection, Multi-Objective Optimization, Performance Indicators, Performance Assessment, Termination Criterion.

1 Introduction

In recent years, the use of evolutionary algorithms (EAs) for solving multi-objective optimization problems has become established. The search for a set of solutions which approximates the Pareto-optimal front of a problem corresponds well to the population maintained within an EA. In particular for black-box problems where no gradient information is available, the use of biologically-inspired stochastic variation and selection operators provides a successful alternative.

However, without gradient information, the examination of formal convergence or optimality criteria, e. g., the Karush-Kuhn-Tucker conditions, is impossible. Therefore, the termination of multi-objective EA (MOEA) is often decided

R.H.C. Takahashi et al. (Eds.): EMO 2011, LNCS 6576, pp. 16–30, 2011.

based on heuristic stopping criteria, such as the maximum number of evaluations or a desired value of a performance indicator. Whereas the criteria are suitable for analytically defined benchmark problems, where the optimal indicator value is known, their applicability to real-world black-box problems is questionable. In cases where the evaluation budget or the desired indicator level is inappropriately specified, the MOEA can either waste computational resources or can be stopped although the approximation still shows a significant improvement. Consequently, heuristic stopping criteria for the online detection of the generation, where the expected improvement in the approximation quality does not justify the costs of additional evaluations, provide an important contribution to the efficiency of MOEA.

In line with these findings, research on sophisticated heuristic online stopping criteria (OSC) has obtained increasing popularity (e.g. [8,12,16,19,20]). OSC analyze the progression of single or multiple progress indicators (PI) online during the run of the MOEA. When the considered indicators seem to be converged, i. e., the expected improvement seems to be lower than a predefined threshold, the MOEA is terminated in order to avoid wasting computational resources.

Despite being proposed by different authors with different methodological background, all these criteria show structural similarities. Thus, a taxonomy of OSC is presented which is the formal contribution of this paper and makes up the basis for the implementation of the MATLAB toolbox. Based on the foundations of set-based multi-objective optimization (section 2), the special requirements for multi-objective OSC are identified. In the main section 3, a formal framework is defined (3.1), known OSC approaches are integrated within the taxonomy by structuring them into their building blocks (3.2), and a discussion of the state of the art is provided (3.3). All building blocks identified in the taxonomy are made available in a MATLAB toolbox [18] which is briefly described in section 4. By means of this toolbox, all existing and many new OSC can be designed and analyzed. The paper is summarized and conclusions are given in section 5.

2 Foundations

Without loss of generality[1], a MOP can be formally expressed as

$$\min_{\boldsymbol{x}\in\mathcal{D}} \boldsymbol{f}(\boldsymbol{x}) = \langle f_1(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x})\rangle\,, \tag{1}$$

i. e., a vector of objective functions $f_1(\boldsymbol{x}), \ldots, f_m(\boldsymbol{x})$ is jointly optimized. The feasible region $\mathcal{D} \subseteq \mathcal{X}$ of the search space $\mathcal{X}$ is denoted as decision space while the image set $\mathcal{O} \subseteq \mathbb{R}^m$ of the projection $\boldsymbol{f} : \mathcal{D} \to \mathcal{O}$ is denoted as the feasible set of the objective space $\mathbb{R}^m$.

The solution to problem (1) is the set of trade-off points jointly minimizing the objective functions. The formalism behind the joint optimization is expressed in terms of the Pareto dominance relation. A decision vector $\boldsymbol{x}$ dominates another vector $\boldsymbol{x}'$, iff $\forall i \in \{1, \ldots, m\} : f_i(\boldsymbol{x}) \leq f_i(\boldsymbol{x}')$ and $\exists i \in \{1, \ldots, m\} :$

[1] Maximization problems max $\boldsymbol{f}(\boldsymbol{x})$ can be written as min $-\boldsymbol{f}(\boldsymbol{x})$.

$f_i(\boldsymbol{x}) \neq f_i(\boldsymbol{x}')$. The subset of $\mathcal{D}$ which contains the elements that are not dominated by any other element of $\mathcal{D}$ is denoted as the Pareto-optimal set $\mathcal{D}^*$. Its image in the objective space $\mathcal{O}$ is called the Pareto-optimal front $\mathcal{O}^*$. For continuous problems, $\mathcal{D}^*$ and $\mathcal{O}^*$ usually contain an infinite number of solutions.

MOEAs are population-based stochastic optimization algorithms. Each individual in the population $\mathcal{P}$ of the MOEA represents a candidate solution. During the optimization, the individuals are improved by means of evolutionary operators, such as mutation and crossover. The image of the non-dominated individuals in objective space is denoted as non-dominated front $\mathcal{PF}^*$. The objective vectors obtained by the individuals in $\mathcal{PF}^*$ provide a finite-size approximation of $\mathcal{O}^*$. In order to evaluate the quality of the approximation set $\mathcal{PF}_t^*$ of generation t, set performance indicators have become established [21]. The target of an OSC is to detect the generation t which provides the best possible trade-off between the approximation quality of $\mathcal{PF}_t^*$ and the required generations t.

3 Taxonomy

A brief summary of theoretical single- and multi-objective convergence detection approaches based on formal conditions has already been published [16]. In this summary, also the differences between multi- and single-objective problems is discussed. If the application of formal convergence conditions is not possible, heuristic OSC are used to detect that further improvements are unlikely, or are expected to be too small – even if no formal convergence is obtained. In this paper, we are focusing on these heuristic OSC. Thus, a formal notation of convergence of a set of points in the multi-objective context is not required. The procedure of these OSC can be separated in at least two steps:

1. The expected improvement of the MOEA is evaluated.
2. Based on this improvement and a predefined threshold, a decision about stopping the MOEA is made.

For the first step, several PIs have been proposed. A straightforward approach is the use of unary performance indicators, such as convergence metric (CM) and diversity metric (DVM) [4], maximum crowding distance (maxCD) [15], or the hypervolume (HV)[20,19,9] dominated by the current $\mathcal{PF}_t^*$ with respect to a reference point which is dominated by all individuals in $\mathcal{PF}_t^*$ [6]. Moreover, binary performance indicators, such as the ε- (Epsilon) [21], and the R2-indicator (R) [10], can be used to evaluate the improvement between different MOEA generations [20,9]. In these cases, but also for some of the unary indicators (CM and DVM), a reference set is required. Since the best set available is usually the one of the current generation $\mathcal{PF}_t^*$, a recomputation of previous PI values based on this reference set can become necessary. A clear advantage of using established performance metrics consists in the availability of formal results from the theory of performance assessment [21] which can be transferred to the PI.

Nevertheless, also specialized PIs for convergence detection have been presented. Martí et al. [11,12] proposed the mutual domination rate (MDR). MDR

contrasts how many individuals of $\mathcal{PF}_t^*$ dominate individuals of $\mathcal{PF}_{t-1}^*$ and vice versa. It is capable of measuring the progress of the optimization with almost no additional computational cost as it can be embedded in Pareto-based MOEAs and reuses their computations. Therefore it is suitable for solving large-scale or many-objective problems with large population sizes. If MDR equals 1 then the entire population of the iteration is better than its predecessor. For MDR $= 0$, no substantial progress has been achieved. MDR < 0 indicates a deterioration of the population. Bui et al. [3] introduced the dominance-based quality (DQP) of a set $\mathcal{PF}_t^*$. For each solution in $\mathcal{PF}_t^*$, the ratio of dominating individuals in the neighborhood of this solution are computed. The DQP is then defined as the average ratio over all solutions in $\mathcal{PF}_t^*$. DQP $= 0$ indicates that no improving solutions can be found in the neighborhood of the current solutions in $\mathcal{PF}_t^*$. For the estimation of the ratios, Monte Carlo sampling is performed around each solution in $\mathcal{PF}_t^*$. Thus, the DQP is only suitable if many additional evaluations of the objective function can be performed. Goel and Stander [8] proposed the consolidation ratio (CR). The CR is a dominance-based convergence metric based on an external archive of all non-dominated solutions found during the run of the MOEA. It is defined as the relative amount of the archive members in generation $t - t_{\text{mem}}$ which are still contained in the archive of the current generation t. In improving phases CR should be low whereas it asymptotically approaches one when convergence is achieved. This PI can be inefficient because the archive can become very large, in particular for many-objective problems.

Because of the non-deterministic nature of EAs, it can be of avail to have an *evidence gathering process* (EGP) that combines different PI values. This EGP can take into account the measurements of previous generations or more than one PI in order to increase the robustness of the approach. Different EGP approaches are discussed in the following while the descriptive notation in brackets is later on used in the formal framework in section 3.1. Many approaches [3,8] directly use the value of the PI computed in the current generation t for deciding if the MOEA is stopped (Direct). However, a single PI evaluation usually cannot provide enough information for a robust conclusion. A straightforward idea of aggregating different PI values is the use of descriptive statistics. In particular, the second moment, i. e., the standard deviation (STD) of the values, is used in order to evaluate the variability within the PI [15,20,19]. Martí et al. [11,12] propose the use of a simplified Kalman filter (Kalman). Due to the recursive formulation, the estimation at iteration t is based on all PI values gathered until then. Moreover, it considers the associated covariance error, i. e., the minimum possible error under linear conditions. A similar, but simpler, idea is proposed by Goel and Stander [8] which use a moving average as EGP (Moving). In other approaches, a linear regression analysis on the PI values of the last t_{mem} generations is performed [20,9] (Reg) in order to estimate the expected improvement and to filter out the stochastic noise in the PI evaluations.

Based on the outcome of the EGP, it can be decided whether the MOEA is stopped. Most known approaches [15,3,8] use a threshold with which the outcome is compared (Threshold). The MOEA is stopped in case the current value of the

EGP exceeds or falls below this threshold. The approaches of Martí et al. [11,12] also use confidence information (CI) based on the assumption of a normally distributed error (CInormal). The MOEA is only stopped when the estimated EGP value is below the threshold with a given probability limit. Guerrero et al. [9] do not use CIs, but ensure the quality of the regression analysis by comparing the mean squared error of the fit to a precomputed threshold (validThreshold). As an extension of this approach, Trautmann and Wagner [20,19] perform statistical tests on the outcome of the EGP. These tests are adapted to the corresponding EGP, i. e., the χ^2-test especially suited for variances (squared STD) is used with STD while a t-test is used in cases where an estimated EGP value with a normal distributed error is provided by the EGP, such as for Reg or Kalman (adaptTest). In these approaches, the p-values obtained in the tests are compared to a fixed significance level $\alpha = 0.05$. In order to further increase the robustness of the stopping decision, Wagner et al. [20] propose to wait for a repeated occurrence of the stopping condition, denoted as hits h. Moreover, the use of multiple EGPs can assist in analyzing different aspects of the PI, such as the variation (STD) and the significance of the linear trend (Reg) [20].

3.1 Formal Framework

An online stopping criterion can be formally defined as a 4-tuple,

$$
\begin{aligned}
OSC := \{&\mathcal{S}, \mathit{\Pi}(\cdot), \Upsilon(\cdot), \Phi(\cdot)\} && \text{with} \\
\mathcal{S} \;&:\; \text{data structure}, && \text{(state of the OSC)} \\
\mathit{\Pi} \;&:\; \mathcal{PF}_t^* \times \mathcal{S} \to \mathcal{S}, && \text{(progress indicator (PI) computation)} \\
\Upsilon \;&:\; \mathcal{S} \to \mathcal{S}, && \text{(evidence gathering process, EGP)} \\
\Phi \;&:\; \mathcal{S} \to \{\text{true}, \text{false}\}\,. && \text{(stopping decision)}
\end{aligned}
\tag{2}
$$

In the state $\mathcal{S}$, all information required for the computations of the EGP are stored. It necessarily includes the input data $\mathcal{M}$ for the EGP. In the following, we use $\mathcal{S}.\mathcal{M}$ in order to address the current version of $\mathcal{M}$ stored in the state $\mathcal{S}$. The state $\mathcal{S}$ can additionally contain previous Pareto front approximations or PI values, an external archive, or flags indicating whether the threshold has been reached in the last generations. These information can be changed or used in different functions of the taxonomy and are therefore exchanged via $\mathcal{S}$. The data stored in the state ensures that the OSC can make the stopping decision just based on the Pareto front approximation $\mathcal{PF}_t^*$ of the current generation.

The function $\mathit{\Pi} : \mathcal{PF}_t^* \times \mathcal{S} \to \mathcal{S}$ uses the PIs to update the input data $\mathcal{S}.\mathcal{M}$ for the EGP. This general type of function is introduced since the update can differ depending on the considered PIs, e. g., some approaches update the PI of all preceding generations based on the current generation $\mathcal{PF}_t^*$, whereas others only update the values of the last generation. Consequently, the size of $\mathcal{S}.\mathcal{M}$ can be up to $P \times t_{\text{mem}}$, where P is the number of PIs and t_{mem} is the number of preceding generations considered in the EGP. In $\mathit{\Pi}$ also all state updates required for the PI computation, such as the update of the archive and the storage of previously computed PI values, are performed. Consequently, the

input data $\mathcal{S}.\mathcal{M}$ is a necessary part of the updated state, as it would restrict the generality of the framework as sole output of Π.

The function $\Upsilon : \mathcal{S} \to \mathcal{S}$ encodes the EGP. It updates the state of the criterion based on the input data $\mathcal{S}.\mathcal{M}$ included in the current state. Usually, the EGP returns one aggregated value per PI, but also a combined analysis like in OCD [20] can be performed. In this case, the EGP value of the combined analysis is assigned to all considered PI.

The decision function $\Phi : \mathcal{S} \to \{\text{true}, \text{false}\}$ finally determines whether the current state of the criterion indicates that the expected improvement of the MOEA is below the predefined threshold ε, i. e., the MOEA should be stopped. For this decision, the EGP value, but also additional information, such as the estimation error and the degrees of freedom in the estimation of the EGP value, are usually utilized. The decision function can only return a single Boolean. If multiple EGPs are considered in parallel, also the aggregation of the corresponding decisions has to be performed in Φ.

Using these formalisms, the procedure of a generic OSC can be implemented as shown in Algorithm 1. The user has to specify the MOEA, the problem of interest and the maximum affordable number of generations $t_{\max}$, as well as PI-related data of the problem, such as a reference set and the ideal and nadir points [6]. The actual OSC is specified by the combination of the PIs, the EGPs, and the stopping decisions. For each step, also multiple functions can be provided.

After the initialization of the state in which the archive is initialized and information about the chosen PI and EGP are stored, the control parameters of the OSC are initialized. After each generation of the MOEA, $\mathcal{S}.\mathcal{M}$ and the required data structures are updated using the chosen Π_i. If there are t_{mem} measurements, the functions $\Upsilon_j(\cdot)$ are applied in order to attach the EGP value for each PI to $\mathcal{S}$. Finally, $\Phi_k(\cdot)$ can be applied to determine whether the algorithm should be stopped.

3.2 Integration of the State of the Art

In this subsection, we will present a survey of the state-of-the-art OSC in chronological publication date order. These approaches are described using the proposed formalization. A summary is provided in Table 1.

Deb and Jain: Running Metrics. Deb and Jain [4] were the first authors who proposed the investigation of performance metrics over the run of the MOEA. They used two metrics, one for evaluating the convergence and one for measuring the diversity of $\mathcal{PF}_t^*$. The convergence metric (CM) calculates the average of the smallest normalized euclidean distance from each individual in $\mathcal{PF}_t^*$ to a precomputed reference set. For the computation of the diversity metric (DVM), all objective vectors of $\mathcal{PF}_t^*$ are projected onto a hyperplane of dimension $m-1$ which is then uniformly divided into discrete grid cells. The DVM tracks the number of attained grid cells and also evaluates the distribution by assigning different scores for predefined neighborhood patterns. In order to avoid bad DVM values based on unattainable grid cells, again a reference set is used. The

Algorithm 1. Implementation of an OSC using the taxonomy definition (eq. 2)

General parameters:
- Multi-objective evolutionary algorithm of interest.
- Multi-objective problem of interest.
- $t_{\max}$, maximum number of iterations.
- $\mathcal{PI}$, set of PI functions Π_i.
- $\mathcal{EGP}$, set of EGP functions Υ_j.
- $\mathcal{SDF}$, set of stopping decision functions Φ_k, $k = \{1, \ldots, K\}$.
- Problem-based parameters (reference set, ideal and nadir points).
- Manually defined settings of control parameters (optional).

Initialize state $\mathcal{S}$.
Initialize control parameters of Π_i, Υ_j, and Φ_k.
$t = 0$.
while $t < t_{\max}$ **do**
  $t = t + 1$.
  Perform one generation of the MOEA and obtain $\mathcal{PF}_t^*$.
  for each indicator Π_i in $\mathcal{PI}$ **do**
    Update input data $\mathcal{S}.\mathcal{M}$ and PI-dependent information, $\mathcal{S} = \Pi_i(\mathcal{PF}_t^*, \mathcal{S})$.
  end for
  if $|\mathcal{S}.\mathcal{M}| = t_{\text{mem}}$ **then**
    for each EGP Υ_j in $\mathcal{EGP}$ **do**
      Update EGP value based on $\mathcal{S}.\mathcal{M}$, $\mathcal{S} = \Upsilon_j(\mathcal{S})$.
    end for
    for each stopping decision function Φ_k in $\mathcal{SDF}$ **do**
      Compute stop decision, stop(k) = $\Phi_k(\mathcal{S})$
    end for
    if $\forall k$: stop(k) = true **then**
      Stop MOEA!
      return t and $\mathcal{S}$.
    end if
  end if
end while

EGP and the final decision then rely on a visual inspection of the progression of the CM and DVM by the user. Consequently, the state $\mathcal{S}$ of this criterion contains the reference set and all values of the CM and DVM computed until the current generation.

Rudenko and Schoenauer: Stability Measure. Rudenko and Schoenauer [15] defined a stability measure for the $\mathcal{PF}_t^*$ of NSGA-II [5]. Their experimental studies showed that the stagnation of the maximum crowding distance (maxCD) within $\mathcal{PF}_t^*$ is a suitable indicator for NSGA-II convergence. Thus, the standard deviation of the last t_{mem} values of the maximum crowding distance is used as EGP (STD). For the computation, the last $t_{\text{mem}} - 1$ values of $maxCD$ are contained in the state $\mathcal{S}$. In each generation, $\mathcal{S}$ is updated using the current $maxCD$ value and STD is computed. The decision step requires a user defined threshold ε leading to an NSGA-II termination once the STD falls below this value (Threshold).

Table 1. Definition of the taxonomy functions in known approaches. The acronyms and abbreviations are introduced in section 3.

Approach	$\Pi(\mathcal{PF}^*_t, \mathcal{S})$	$\mathcal{S}$	$\Upsilon(\mathcal{S})$	$\Phi(\mathcal{S})$	Parameters (default)
Running metrics [4]	CM, DVM $\lvert\mathcal{S}.\mathcal{M}\rvert = 2 \times 1$	- $\mathcal{M}$: All CM and DVM values - Reference set	Attach current CM and DVM values to state	Visual check by decision maker	none
Stability measure [15]	maxCD $\lvert\mathcal{S}.\mathcal{M}\rvert = 1 \times 1$	- $\mathcal{M}$: maxCD of generation $(t - t_{\text{mem}})$ to t - STD	STD of $\mathcal{S}.\mathcal{M}$	Threshold	- t_{mem} (40) - ε (0.02) - hits h (1)
MBGM [11, 12]	MDR $\lvert\mathcal{S}.\mathcal{M}\rvert = 1 \times 1$	- $\mathcal{M}$: Current MDR value - $\mathcal{PF}^*_{t-1}$ - Kalman state with corresp. STD	Kalman	CInormal	- ε (0) - p-CI (97.725 %) - Kalman inertia R (0.1) - hits h (1)
OCD-Classic [20, 14, 19]	Eps, R, HV (parallel) $\lvert\mathcal{S}.\mathcal{M}\rvert = 3 \times t_{\text{mem}}$	- $\mathcal{M}$: Eps, R, and HV of generation $(t - t_{\text{mem}})$ to $(t - 1)$ - $\mathcal{PF}^*$ of generation $(t-t_{\text{mem}})$ to $(t-1)$ - Current slope β with corresp. STD - Current STD - p-values of generations t and $(t - 1)$	- STD of each PI in $\mathcal{S}.\mathcal{M}$ - Reg on $\mathcal{S}.\mathcal{M}$ (individually standardized)	- χ^2-test: STD $< \varepsilon$ - t-test: $\beta = 0$	- t_{mem} (16) - ε (0.001) - hits h (2)
DQP [3]	DQP $\lvert\mathcal{S}.\mathcal{M}\rvert = 1 \times 1$	$\mathcal{M}$: Current DQP value	Direct	Visual check by decision maker	- samples N (500) - radius r (0.05)
LSSC [9]	Eps, HV, MDR (separately) $\lvert\mathcal{S}.\mathcal{M}\rvert = 1 \times t_{\text{mem}}$	- $\mathcal{M}$: Eps, HV, or MDR of generation $(t - t_{\text{mem}})$ to $(t - 1)$ - $\mathcal{PF}^*$ of generation $(t-t_{\text{mem}})$ to $(t-1)$ - Current slope β	Reg on each PI in $\mathcal{S}.\mathcal{M}$	validThreshold	- t_{mem} (30) - ε (HV: 0.002, Epsilon: 0.0004, MDR: 0.00002) - hits h (1)
OCD-HV [19]	HV $\lvert\mathcal{S}.\mathcal{M}\rvert = 1 \times t_{\text{mem}}$	- $\mathcal{M}$: Differences between the HVs of generation $(t - t_{\text{mem}})$ to $(t - 1)$ and t - p-values of generations t and $(t - 1)$	STD of $\mathcal{S}.\mathcal{M}$	- χ^2-test: STD $< \varepsilon$	- t_{mem} (14) - ε (0.0001) - hits h (2)
CR [8]	CR $\lvert\mathcal{S}.\mathcal{M}\rvert = 1 \times 1$	- $\mathcal{M}$: CR of generation $(t - t_{\text{mem}})$ to t - Archive of non-dominated ind. - $\varepsilon_{adaptive}$ - $\text{U}^*_{t-t_{mem}}$	Direct or Moving average	CR $> \varepsilon$ or $\text{U}^*_t < \varepsilon_{adaptive}$	- t_{mem} (10) - ε (0.8) - utility ratio F (10) - hits h (1)

Martí et al.: MGBM Criterion. Martí, García, Berlanga, and Molina [11,12] proposed the MGBM criterion (according to the authors' last names), which combines the mutual domination rate (MDR) with a simplified Kalman filter that is used as EGP. The function Π considers $\mathcal{PF}^*_{t-1}$ and $\mathcal{PF}^*_t$ and applies the MDR indicator to update $\mathcal{S}.\mathcal{M}$. Thus, the Pareto front of the previous generation has to be stored in the state $\mathcal{S}$. The EGP function Υ applies the Kalman filter and updates the Kalman state and the corresponding estimated error in $\mathcal{S}$. The decision function Φ is realized by stopping the MOEA when the confidence interval of the a-posteriori estimation completely falls below the prespecified threshold ε.

Wagner et al.: Online Convergence Detection (OCD). In the Online Convergence Detection [20] approach, the established performance measures HV, R2- and additive ε-indicator are used as PIs. The function Π updates all t_{mem} PI values stored in $\mathcal{S}.\mathcal{M}$ using the current generation $\mathcal{PF}^*_t$ as reference set. Consequently, the sets $\mathcal{PF}^*_{t-t_{\text{mem}}}$ to $\mathcal{PF}^*_{t-1}$ have to be additionally stored in the state $\mathcal{S}$. In Υ, the variance of the values in $\mathcal{S}.\mathcal{M}$ is computed for each PI. Moreover, a least-squares fit of a linear model with slope parameter β is performed based on the individually standardized values in $\mathcal{S}.\mathcal{M}$. In Φ, the variance is then compared to a threshold variance ε by means of the one-sided χ^2-variance test with H_0: $\text{VAR}(\mathcal{S}.\mathcal{M}) \geq \varepsilon$ and a p-value is looked up. By testing the hypothesis H_0: $\beta = 0$ by means of a t-test, a second p-value is obtained. For these tests, the variance obtained by STD, β, and its standard error have to be stored in the state. The same holds for the resulting p-values. The MOEA is stopped when the p-values of two consecutive generations are below the critical level $\alpha = 0.05$ for one of the variance tests (the null hypothesis H_0 is rejected) or above $\alpha = 0.05$ for the regression test (H_0 is accepted). Consequently, the p-values of the preceding generations have to be stored in $\mathcal{S}$.

In [19] a reduced variant of the OCD approach for indicator-based MOEA was introduced. This approach was illustrated for the HV indicator and the SMS-EMOA [1] (OCD-HV). Since the HV is a unary indicator, only the absolute HV values have to be stored. The previous $\mathcal{PF}^*_t$ can be neglected. For better compliance with the other PI, the differences to the value of the current set $\mathcal{PF}^*_t$ are stored in $\mathcal{S}.\mathcal{M}$ in order to minimize the PI. In case the internally optimized performance indicator monotonically increases, as for the SMS-EMOA and the HV, OCD should only consider this PI. The regression test can be neglected. Consequently, the complexity of OCD is reduced by concentrating on the variance test for one specific PI.

Bui et al.: Dominance-Based Quality of $\mathcal{P}$ (DQP). Bui et al. [3] introduce a dominance-based stability measure which approximately evaluates the local optimality of a solution (DQP). The DQP is the only PI that requires many additional evaluations of the objective function for estimating the ratio of dominating solutions in the neighborhood of a solution. A Monte Carlo simulation with 500 evaluations per solution in $\mathcal{PF}^*_t$ was used. Consequently, the DQP is a very expensive, but powerful measure. No additional state informations or EGPs are

required. No clear guidelines for stopping the MOEA are provided. Instead, a visual analysis of the convergence behavior and possible stagnation phases is performed. However, a clear stopping criterion would be $DQP = 0$, as this would be the case when no local improvements are possible. In fact DQP is closely related to the gradient of a solution in single-objective optimization. In line with this observation, the authors also use DQP as measure for guiding a local search [3].

Guerrero et al.: Least Squares Stopping Criterion (LSSC). LSSC [9] can be seen as an approach to integrate both EGP of OCD into a single EGP and to also simplify the PI computation and the stopping decision. Therefore, only one PI is considered and the variance-based EGP and the statistical tests for the stopping decision are omitted. Still, a regression analysis of the PI is performed as EGP and the PI values of the last t_{mem} generations are updated using the current generation as reference set. Thus, the last t_{mem} Pareto front approximations have to be stored in the state $\mathcal{S}$ in order to update $\mathcal{S}.\mathcal{M}$. In contrast, the PIs are not standardized allowing the estimation of the expected improvement by means of the slope β. If β falls below the predefined threshold ε, the MOEA is stopped. In order to prevent a loss of robustness by omitting the statistical tests, a threshold for a goodness-of-fit test based on the regression residuals is computed via the Chebyshev inequality. Only if the model is valid, the estimated slope is compared to ε. Consequently, the analyses performed in OCD and LSSC differ. Whereas LSSC directly tries to detect whether the expected improvement falls below the allowed threshold ε, OCD tests the significance of the linear trend whereas the magnitude of the expected improvement is evaluated via the variance of $\mathcal{S}.\mathcal{M}$.

Goel and Stander: Non-dominance based convergence metric. Goel and Stander [8] use a dominance-based PI based on an external archive of non-dominated solutions which is updated in each generation. The current archive is stored in $\mathcal{S}$ and is used to determine the CR. The authors provide empirical evidence for the robustness of the CR, so that no EGP is applied (Direct). The stopping decision is made by comparing the CR with a predefined threshold ε (Threshold).

In addition, an utility-based approach is proposed. The utility is defined as the difference in the CR between the generations t and $t - t_{\text{mem}}$. In order to increase the robustness of the approach, a moving average $U_t^* = (U_t + U_{t-t_{\text{mem}}})/2$ is used as EGP (Moving). The MOEA is stopped when the utility falls below an adaptively computed threshold $\varepsilon_{\text{adaptive}}$. Moreover, a minimum CR of $\text{CR}_{\text{min}} = 0.5$ has to be reached in order to avoid a premature stopping due to perturbances in early generations. The adaptive threshold $\varepsilon_{\text{adaptive}}$ is defined as the fraction $\text{CR}_{\text{init}}/(F \cdot t_{\text{init}})$ of the initial utility U_{init}, which corresponds to the first CR value CR_{init} exceeding 0.5 and the corresponding generation t_{init}. F is a user parameter that specifies which ratio of the averaged initial utility $\text{CR}_{\text{init}}/t_{\text{init}}$ is at least acceptable. For this version, also $\varepsilon_{\text{adaptive}}$ and $\text{U}^*_{t-t_{\text{mem}}}$ have to be stored in $\mathcal{S}$.

3.3 Discussion

Basically, the existing PIs can be classified with respect to their optimization goal. One class is formed by the PIs based on analyzing the dominance relation between the current population (or archive) and a previous one, e. g., MDR and CR. Other approaches provide information about the distribution (maxCD, DVM) or local optimality of the solutions (DQP). Only a few of the PI try to combine some of these goals, e. g., HV, R, Epsilon, and CM, each with different trade-offs.

The dominance-based PI the convergence of the population to be formally assessed. The probability of improving the diversity and distribution and therewith the quality of the discrete approximation of O^* is not specifically addressed. The improvements in these PI will therefore reduce much faster. Moreover, the magnitude of the improvement generated by a new non-dominated solution is not considered. This information would be important in order to evaluate an expected improvement. As shown in the last years [17], the dominance relation has only a weak explanatory power for many-objective problems.

The dominance-based PI usually reuse the information provided by the selection method of the MOEA. Thus, they do not require expensive additional calculations. PIs like CM, R, and HV have to be additionally computed in each MOEA generation, where especially the dominated hypervolume has a complexity which increases exponentially with the objective space dimension. Bui et al. [3] even perform additional evaluations for convergence detection. In general, the use of additional computational time or evaluations should be kept below the effort of the alternative option of just allowing the MOEA to precede for an affordable number of additional generations.

In addition, reference and nadir points, as well as reference sets, can be required for some PIs, e. g., the reference set for the CM and DVM, the ideal and nadir point for R2, and the reference point for HV. In contrast to mathematical test cases, this information is usually not existing for practical applications. Strategies to obtain this data have to be derived which could comprise preliminary algorithm runs, random sampling, or evaluations on a grid covering the whole search space. Based on approximations of the objective boundaries, the normalization of the PI to unit intervals is possible – an approach that is often recommended [4,21]. However, even the normalization can lead to scalarization effects which make the specification of thresholds difficult [19]. For the dominance-based indicators, usually relative amounts are calculated, e.g., $-1 \leq MDR \leq 1$ or $0 \leq CR \leq 1$, which facilitate the definition of adequate threshold values. Nevertheless, the only reasonable threshold for these approaches is $\varepsilon = 0$ based on the above considerations.

Some methods do not use a distinct EGP. They rely on a single evaluation of the considered PI. Due to the stochastic nature of MOEAs, it is obvious that those approaches will not be as robust as alternative ones using an EGP gathering PIs over a time window. Moreover, the EGP-based approaches are usually flexible with respect to the kind of integrated PI. By means of a suitable PI, the performance aspects (e. g., convergence, distribution, spread) which are the

most important for the optimization task at hand can be considered in the OSC. In this context, also the considered MOEA has an important role. Mathematical convergence can only be expected if the corresponding MOEA is based on this PI, e. g., the SMS-EMOA in combination with the HV [2]. Furthermore, most OSC are designed for separately using a single PI. As performance of a MOEA has different aspects [4,21], it should be analyzed if the usage of PIs covering these aspects of the approximation quality could support an efficient OSC decision.

Another important OSC design issue is concerned with the choice of the stopping decision. Statistical tests or confidence intervals lend themselves to draw robust decisions from random variables monitored over time. However, in order to choose an adequate test or distribution, some assumptions on the behavior of the considered PI are necessary. As a first approach, Mersmann et al. [13] analyze the distribution of the final HV value of different MOEAs. Among other characteristics it is shown to be unimodal in most cases. Consequently, the use of classical tests is possible, maybe based on additional transformations.

The parametrization of the OSC requires special attention as well. Parameters have to be carefully chosen in order to obtain the desired results with respect to the trade-off between runtime and approximation quality. For most approaches, no clear guidelines for setting up the required parameters are given or a visual analysis is suggested [4,3]. In contrast, Wagner and Trautmann [19] empirically derive guidelines for reasonably setting the OCD parameters t_{mem} and ε based on statistical design-of-experiment methods. The resulting parameter recommendations can be found in Table 1. For reasonable comparisons between the OSC, such kind of studies should also be performed for the other OSC. Furthermore, the problems and possibilities resulting from a combination of the methods with respect to the proposed PI, EGP, and stopping decisions should be a matter of future research. In this context, an analysis of the compatibility of the PI, EGP, and decision criteria would be of special interest.

4 MATLAB Toolbox for Online Stopping Criteria

In the previous subsection, many open questions in the field of OSC are discussed. However, all choices of test problems and MOEAs for the analysis of OSC put a subjective bias to the results. In order to assist researchers in analyzing these questions, a MATLAB toolbox based on the OSC taxonomy was implemented [18]. Thus, the framework allows the application of the OSC to the test problems and favorite MOEAs of the user. Based on the framework, an interested user can analyze and tune the OSC on his specific setup.

The framework follows the pseudocode provided in Algorithm 1. It allows the arbitrary combination of building blocks which can be used to design an adapted OSC for the specific task at hand. Consequently, the analysis of the compatibility of different subfunctions can directly be performed. Within the framework, the abbreviations of section 3 are used to address the corresponding subfunctions. Accordingly, each subfunction is documented in this paper.

The control parameters of the OSC are initialized automatically using default values. This procedure prevents the user from searching for default values of the parameters before using the framework and also encourages researchers to perform parameter studies before proposing an OSC. Nevertheless, experienced users have the opportunity to specify some of the parameters on their own using a options structure.

In order to allow arbitrary combinations of Π, Υ, and Φ, as proposed in Algorithm 1, some additional features are integrated within the MATLAB framework:

- The use of different stopping decisions in parallel is possible.
- It is possible to combine the stopping decisions for different PI and EGP by more than the already proposed rule: all PI for at least one EGP [20]. Further possibilities are: all, any, all EGP for at least one PI, and a majority voting.
- The standard deviation of STD-EGP is calculated using bootstrapping [7].

The choice of allowing multiple stopping decisions in parallel is motivated by the different amounts of information provided by the different EGP. The CI- and t-test-based approaches require EGP that also provide error estimates. By combining these methods with the threshold decision, which will always stop when the CI- or test-based approaches would, also these EGP can be handled. Thus, if some information is missing, e. g., the standard deviation after applying the Direct EGP, the adaptTest- or CInormal-EGP are ignored and only Threshold decision is used. This enhancement makes the framework more flexible for new conceptually different stopping decisions.

By means of the formalization through the taxonomy, the interfaces for the framework are clearly defined. Researchers in the field of OSC can easily integrate their methods by structuring their OSC following the taxonomy. Then each subfunction is implemented within the framework, and a benchmark with all state-of-the-art OSC can directly be performed. As a side effect, a systematic integration of new OSC into the state of the art is promoted.

5 Conclusion and Outlook

In this paper, a comprehensive overview of sophisticated heuristic online stopping criteria (OSC) for EA-based multi-objective optimization is provided. The approaches are integrated into a taxonomy by splitting them into their building blocks which cover the different steps to be performed when applying an OSC. The presented taxonomy allows comparisons of OSC approaches to be systematically performed. The analysis of the strengths and weaknesses of a specific OSC can be broken down to the responsible subfunctions, e. g., the methods can be classified by the kind of PI used, the complexity of the EGP, and the integration of statistical techniques in the decision making. Concluding, OSC methods relying on an EGP with respect to PIs gathered from preceding generations are likely to be more robust, but computationally expensive. The additional use of statistical techniques can further increase robustness, but needs to be adapted

to the data of the EGP. In contrast, complex and expensive PI like DQP may not require a sophisticated EGP or stopping decision.

The parametrization of an individual OSC is not an easy task and strongly influences its performance. Unfortunately, sufficient and comprehensive guidelines for the required parameter settings are only presented for a small subset of the OSC strategies. Moreover, the recommended thresholds for the specific PI are different, making a fair comparison almost impossible. In order to simplify a systematic comparison, a MATLAB toolbox [18] was implemented. This toolbox is structured according to the building blocks of the presented taxonomy and all approaches discussed in this paper were integrated. By means of this toolbox, the expert can evaluate the approaches – and also combinations of them – on his problem and can then choose the OSC which provides the best performance with regard to his objectives.

A systematic evaluation and comparison of all presented approaches will be the main focus of our future research. This includes a parameter tuning, as well as the combination of algorithmic concepts. To accomplish this, a systematic performance assessment of OSC has to be proposed and discussed.

Acknowledgments

This paper is based on investigations of the collaborative research centers SFB/TR TRR 30 and SFB 823, which are kindly supported by the Deutsche Forschungsgemeinschaft (DFG). L. Martí acknowledges support from projects CICYT TIN2008-06742-C02-02/TSI, CICYT TEC2008-06732-C02-02/TEC, SINPROB, CAM CONTEXTS S2009/TIC-1485 and DPS2008-07029-C02-02.

References

1. Beume, N., Naujoks, B., Emmerich, M.: SMS-EMOA: Multiobjective selection based on dominated hypervolume. European Journal of Operational Research 181(3), 1653–1669 (2007)
2. Beume, N., Laumanns, M., Rudolph, G.: Convergence rates of (1+1) evolutionary multiobjective optimization algorithms. In: Schaefer, R., Cotta, C., Kołodziej, J., Rudolph, G. (eds.) PPSN XI. LNCS, vol. 6238, pp. 597–606. Springer, Heidelberg (2010)
3. Bui, L.T., Wesolkowski, S., Bender, A., Abbass, H.A., Barlow, M.: A dominance-based stability measure for multi-objective evolutionary algorithms. In: Tyrrell, A., et al. (eds.) Proc. Int'l. Congress on Evolutionary Computation (CEC 2009), pp. 749–756. IEEE Press, Piscataway (2009)
4. Deb, K., Jain, S.: Running performance metrics for evolutionary multi-objective optimization. In: Simulated Evolution and Learning (SEAL), pp. 13–20 (2002)
5. Deb, K., Pratap, A., Agarwal, S.: A fast and elitist multi-objective genetic algorithm: NSGA-II. IEEE Trans. on Evolutionary Computation 6(8) (2002)
6. Deb, K., Miettinen, K., Chaudhuri, S.: Toward an estimation of nadir objective vector using a hybrid of evolutionary and local search approaches. Trans. Evol. Comp. 14, 821–841 (2010)

7. Efron, B.: Bootstrap methods: Another look at the jackknife. Annals of Statistics 7(1), 1–26 (1979)
8. Goel, T., Stander, N.: A non-dominance-based online stopping criterion for multi–objective evolutionary algorithms. International Journal for Numerical Methods in Engineering (Online access) (2010) doi: 10.1002/nme.2909
9. Guerrero, J.L., Martí, L., García, J., Berlanga, A., Molina, J.M.: Introducing a robust and efficient stopping criterion for MOEAs. In: Fogel, G., Ishibuchi, H. (eds.) Proc. Int'l. Congress on Evolutionary Computation (CEC 2010), pp. 1–8. IEEE Press, Piscataway (2010)
10. Hansen, M.P., Jaszkiewicz, A.: Evaluating the quality of approximations to the non-dominated set. Tech. Rep. IMM-REP-1998-7, Institute of Mathematical Modelling, Technical University of Denmark (1998)
11. Martí, L., García, J., Berlanga, A., Molina, J.M.: A cumulative evidential stopping criterion for multiobjective optimization evolutionary algorithms. In: Thierens, D., et al. (eds.) Proc. of the 9th Annual Conference on Genetic and Evolutionary Computation (GECCO 2007), p. 911. ACM Press, New York (2007)
12. Martí, L., García, J., Berlanga, A., Molina, J.M.: An approach to stopping criteria for multi-objective optimization evolutionary algorithms: The MGBM criterion. In: Tyrrell, A., et al. (eds.) Proc. Int'l. Congress on Evolutionary Computation (CEC 2009), pp. 1263–1270. IEEE Press, Piscataway (2009)
13. Mersmann, O., Trautmann, H., Naujoks, B., Weihs, C.: On the distribution of EMOA hypervolumes. In: Blum, C., Battiti, R. (eds.) LION 4. LNCS, vol. 6073, pp. 333–337. Springer, Heidelberg (2010)
14. Naujoks, B., Trautmann, H.: Online convergence detection for multiobjective aerodynamic applications. In: Tyrrell, A., et al. (eds.) Proc. Int'l. Congress on Evolutionary Computation (CEC 2009), pp. 332–339. IEEE press, Piscataway (2009)
15. Rudenko, O., Schoenauer, M.: A steady performance stopping criterion for pareto-based evolutionary algorithms. In: The 6th International Multi-Objective Programming and Goal Programming Conference, Hammamet, Tunisia (2004)
16. Trautmann, H., Wagner, T., Preuss, M., Mehnen, J.: Statistical methods for convergence detection of multiobjective evolutionary algorithms. Evolutionary Computation Journal, Special Issue: Twelve Years of EC Research in Dortmund 17(4), 493–509 (2009)
17. Wagner, T., Beume, N., Naujoks, B.: Pareto-, aggregation-, and indicator-based methods in many-objective optimization. In: Obayashi, S., et al. (eds.) EMO 2007. LNCS, vol. 4403, pp. 742–756. Springer, Heidelberg (2007)
18. Wagner, T., Martí, L.: Taxonomy-based matlab framework for online stopping criteria (2010), `http://www.giaa.inf.uc3m.es/miembros/lmarti/stopping`
19. Wagner, T., Trautmann, H.: Online convergence detection for evolutionary multi-objective algorithms revisited. In: Fogel, G., Ishibuchi, H. (eds.) Proc. Int'l. Congress on Evolutionary Computation (CEC 2010), pp. 3554–3561. IEEE press, Piscataway (2010)
20. Wagner, T., Trautmann, H., Naujoks, B.: OCD: Online convergence detection for evolutionary multi-objective algorithms based on statistical testing. In: Ehrgott, M., et al. (eds.) EMO 2009. LNCS, vol. 5467, pp. 198–215. Springer, Heidelberg (2009)
21. Zitzler, E., Thiele, L., Laumanns, M., Fonseca, C., Fonseca, V.: Performance assessment of multiobjective optimizers: An analysis and review. IEEE Transactions on Evolutionary Computation 8(2), 117–132 (2003)