Improving Robustness of Stopping Multi-objective Evolutionary Algorithms by Simultaneously Monitoring Objective and Decision Space

Md Shahriar Mahbub
Fondazione Bruno Kessler
Trento, Italy
University of Trento
Trento, Italy
mahbub@fbk.eu

Tobias Wagner
Institute of Machining
Technology (ISF)
TU Dortmund, Germany
wagner@isf.de

Luigi Crema
Fondazione Bruno Kessler
Via Sommarive 18, I-38123
Trento, Italy
crema@fbk.eu

ABSTRACT

Appropriate stopping criteria for multi-objective evolutionary algorithms (MOEA) are an important research topic due to the computational cost of function evaluations, particularly on real-world problems. Most common stopping criteria are based on a fixed budget of function evaluations or the monitoring of the objective space. In this work, we propose a stopping criterion based on monitoring both the objective and decision space of a problem. Average Hausdorff distance (AHD) and genetic diversity are used, respectively. Two-sided t-tests on the slope coefficients after regression analyses are used to detect the stagnation of the AHD and the genetic diversity. The approach is implemented for two widely used MOEAs: NSGA-II and SPEA2. It is compared to a fixed budget, the online convergence detection approach, and the individual monitoring of each space on four bi-objective and two three-objective benchmark problems. Our experimental results reveal that the combined approach achieved significantly better results than the approaches considering only one of the spaces. In particular, we find that the combined consideration runs longer and hence more robustly ensures a well-approximated Pareto front. Nevertheless, on average 29% and 17% function evaluations are saved for NSGA-II and SPEA2, respectively, compared to standard budget recommendations.

Categories and Subject Descriptors

I.2.8 [**Artificial Intelligence**]: Problem Solving, Control Methods, and Search-Heuristic methods

Keywords

Multi-objective optimization, stopping criteria, performance assessment

1. INTRODUCTION

GECCO '15, July 11 - 16, 2015, Madrid, Spain

DOI: http://dx.doi.org/10.1145/2739480.2754680

Virtually all the engineering fields have to deal with optimization problems. Most practical problems encounter multiple contradictory objectives; therefore, the problems need to be considered as multi-objective optimization problems. Multi-objective evolutionary algorithms (MOEA) have become a common technique for optimizing real-world optimization problems [15]. Unlikely to a single-objective optimization where the goal is to find one optimal solution, a MOEA generates a set of non-dominated solutions. The representation of the solutions in objective space is called Pareto front. A MOEA should find the Pareto front [2] by using minimum computational cost. Since function evaluations (FE) of most real-world optimization problems are costly, FE occupy the lion's share of total computational cost of an optimization algorithm. Therefore, finding an appropriate stopping criterion is an important task to minimize computational cost. To save wasteful function evaluations, it is necessary to spot the stagnation or convergence of an algorithm. Until now, almost all MOEAs are stopped after a certain number of function evaluations [15]. However, specifying this number for a practical problem without any prior knowledge is quite difficult. Therefore, in the last few years some techniques for detecting convergence of MOEAs have been proposed [15, 13, 1, 7, 11, 8].

Most proposed techniques are based on investigating the objective space of a problem. Our approach is based on the simultaneous monitoring of the objective and the decision space. Simultaneous stabilization of indicators on both spaces provides a more robust stopping criterion, as it is shown later that an algorithm runs longer when a decision space indicator is combined with an objective space indicator. In this context robust refers to the repeatability of the stopping decision with regard to the stochastic indicator measurements. Additionally, the proposed approach does not require any prior knowledge with regard to thresholding parameters that have to be set. The convergence is detected by analyzing the trends of the two indicators. A two-sided t-test on the slope coefficients is used to detect the stagnation of the linear trend.

The remainder of the paper is organized as follows. In section 2, we briefly describe the state of the art criteria for stopping MOEAs. Section 3 presents the details of our proposed method. Experimental results are reported and

discussed in section 4. Conclusions and future work will be described in section 5.

2. STATE OF THE ART

Deb and Jain [3] were first to propose two metrics, *convergence* and *diversity*, that can be monitored online. The stopping decision of a MOEA depends on the visual inspection of the metrics. At that time, no automatic convergence detection method was proposed. Inspired from the stopping criterion of a single objective evolutionary algorithm (EA), Rudenko and Schoenauer [11] proposed a stability measure. It is based on the density of non-dominated solutions. By studying the dynamics of NSGA-II [4], the authors experimentally showed that the algorithm converges when the maximum crowding distance [4] stagnates. The user needs to provide a threshold and the algorithm stops when the standard deviation of the maximum crowding distance falls below the threshold for a pre-defined number of generations.

Martí et al. [8] proposed a stopping method called MGBM criterion. The authors proposed an indicator named *mutual dominance rate (MDR)* that is basically a measurement of how many non-dominated solutions of one generation dominate the non-dominated solutions of the successive generation. They used a simplified Kalman filer to gather evidences about when to stop. A MOEA is stopped when the a-posteriori estimation of MDR falls below a pre-defined threshold.

Goel and Stander [7] use an external achieve to propose a new indicator, named *consolidation ratio(CR)*. The archive keeps non-dominated solutions and in each generation, the archive is updated. CR is the ratio of the number of solutions of the previous generation that are still present in archive with regard to the size of the archive. Two different stopping criteria are proposed: *fixed threshold approach* and *utility-function based approach*. In the fixed threshold approach, a MOEA is stopped when CR falls below the pre-defined threshold. For the other approach, the utility of evolving extra generation is calculated and the algorithm terminates when the utility falls below the utility threshold.

Bui et al. [1] suggest a stability measure named *dominance-based quality (DQ)* which is based on the dominance relation of a solution with its neighbouring solutions. The basic idea is that the number of dominating neighbouring solutions will decrease over time as a solution moves towards the Pareto front. To measure DQ of a solution, the authors use additional function evaluations and a Monte Carlo simulation approach. The authors do not provide any suggestions about the convergence value of DQ. However, it is obvious that $DQ = 0$ is a powerful stopping criterion and with the help of visual inspection, the stagnation of an algorithm can be identified.

Trautmann et al. [13] and Wagner et al. [15] apply a statistical approach to solve the problem. In [15], the authors propose two different statistical tests on three performance indicators (PI) (i.e., hypervolume, R2 and additive epsilon) to detect convergence. One of the tests is a *one-sided* χ^2*-variance test* for measuring the significance of the variances of PIs compared to pre-defined threshold. Another proposed test is the *two-sided t-test* for detecting stagnation by analyzing slope coefficients of different PI trends. A MOEA is stopped when any of the two tests detects convergence (i.e., p-value lower/above the critical level depending on the test).

Roche et al. propose a stopping criterion based on the loss of population diversity in the decision space [10]. The authors propose a formula to determine the time, when the diversity falls below a threshold. The authors also propose a statistical procedure to confidently stop an EA. However, without prior knowledge about the decision space of a problem, it is very difficult to set the threshold. The method is validated only on single objective problems. In contrast, as we are dealing with MOEAs (population consists of a set of solutions covering a Pareto front, not converging towards a single solution), the population always has certain diversity. Hence, in our context, it is more important to detect the moment in time, when the solutions covering the decision space are no longer moved. Therefore, it is preferable to detect stagnation of the diversity value instead of the value falling below a certain threshold.

3. PROPOSED METHODOLOGY

Most of the studies we have discussed above require an objective-space based indicator which has to fall below an user-defined threshold to stop. Additionally, some recent studies also address the stagnation of multiple objective-space based indicators using statistical tests. In contrast to these approaches, our proposed method is based on the simultaneous monitoring of two metrics; one in objective space and another one in decision space. To get a better stopping/convergence method, our primary assumption is that indicators on these two spaces (i.e., objective and decision) should stabilize concurrently. A stagnation of indicators in objective space (e.g., hypervolume [19], epsilon [20] and so on) does not mean that the algorithm could not improve later. The individuals of a population may contain enough diversity to generate better children in later generations. Therefore, by concurrently monitoring the two spaces, a more robust stopping criterion could be developed. *Average Hausdorff distance* [12] and *diversity* [16] metrics are used to monitor the objective and decision space, respectively. In the field of MOEA, diversity usually refers to population diversity in the objective space; however in this paper, we use the term *diversity* to refer to decision space diversity.

Average Hausdorff distance.

Classical Hausdorff distance is a widely used metric to measure the distance between two sets. It is the maximum distance among all the distances form every point of one set to the closest point in the other set. Therefore, the original version of Hausdorff distance is not a suitable metric for measuring the convergence of multi-objective optimization algorithms [12]. The main reason is that a set with an outlier is largely penalized compared to a set without any outlier when measuring distance from a reference set. That is not compatible with the stochastic nature of MOEAs. The situation with outliers can be improved by averaging the distance between the set and reference set. Therefore, average Hausdorff distance (AHD) is proposed by Schutze et al. in [12] by exploiting generational distance (GD) and inverted generational distance (IGD), as GD and IGD both inherently measure average distance between two sets. Let $A = \{a_1, a_2 \ldots, a_n\}$ and $B = \{b_1, b_2, \ldots, b_m\}$ are two finite sets. GD is defined as follows:

$$GD(A,B) = \frac{1}{n}\left(\sum_{i=1}^{n} dis(a_i,B)^p\right)^{\frac{1}{p}} \quad (1)$$

Where $dis(a_i,B) = \inf_{b_i \in B}\|a_i - b_i\|$ and $\|.\|$ refers to a Euclidean norm. The indicator suffers the problem that a large candidate set having many similar solutions may reduce the GD value. Therefore, it becomes difficult to compare candidate sets with different numbers of solutions (e.g., larger and smaller candidate sets). To avoid this problem, the authors of [12] propose a slightly modified version of GD by taking the power mean of average distances.

$$GD(A,B)_p = \left(\frac{1}{n}\sum_{i=1}^{n} dis(a_i,B)^p\right)^{\frac{1}{p}} \quad (2)$$

Now, the larger candidate set does not always have good GD value. Moreover, the penalizing effect of outliers can be controlled though p. The same approach is applied on IGD [12].

$$IGD(A,B)_p = \left(\frac{1}{m}\sum_{i=1}^{m} dis(b_i,A)^p\right)^{\frac{1}{p}} \quad (3)$$

Finally, the average Hausdorff distance is defined as follows:

$$AHD_p(A,B) = max(GD_p(A,B), IGD_p(A,B)) \quad (4)$$

For small p values, the set with outliers are less penalized than the classical Hausdorff distance. Larger p values have more penalizing effects. Moreover, when $p = \infty$, the AHD is deduced to Hausdorff distance.

Diversity.

There are many possible techniques to measure the genetic diversity of a population. Two popular methods are Hamming distances among all pairs of chromosomes and Shannon entropy on gene frequencies [16]. The authors of [16] argue that the underlying mechanisms of all diversity measurements are similar. Fundamentally, population diversity is a measurement of how individuals of a population are different to each other.

Considering P is a population of n individuals and $\{x_{i,1}, x_{i,2}, \ldots, x_{i,l}\}$ are l chromosomes of an arbitrary i$^{\text{th}}$ individual. The diversity of the k$^{\text{th}}$ chromosome is based on all possible distances between the chromosomes of the individuals. It can be written in the following way [16]:

$$Div_k^2(P) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} dis(x_{i,k}, x_{j,k}) \quad (5)$$

Summing up over all the chromosomes, the equation (5) becomes

$$Div^2(P) = \frac{1}{2}\sum_{k=1}^{l}\sum_{i=1}^{n}\sum_{j=1}^{n} dis(x_{i,k}, x_{j,k}) \quad (6)$$

By using some algebraic manipulation, the diversity can be written in the following form:

$$Div^2(P) = n^2\sum_{k=1}^{l}\left(\overline{x_k^2} - \overline{x_k}^2\right) \quad (7)$$

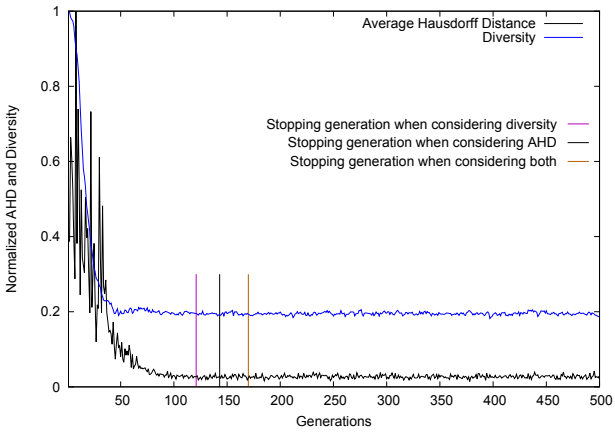

Figure 1: Normalized AHD and diversity are represented by black and blue trends. Three vertical lines with three different colors indicate the stopping generations for three different approaches.

Where $\overline{x_k} = \frac{1}{n}\sum_{i=1}^{n} x_{i,k}$ and $\overline{x_k^2} = \frac{1}{n}\sum_{i=1}^{n} x_{i,k}^2$. Finally, omitting constants and applying a monotone transformation, diversity is defined as follows [16]:

$$Div(P) = \frac{1}{l}\sqrt{\sum_{k=1}^{l}\left(\overline{x_k^2} - \overline{x_k}^2\right)} \quad (8)$$

Figure 1 illustrates the values of two metrics on the ZDT1 [17] problem when solved by NSGA-II. In the figure, the values of AHD and diversity are normalized to show both metrics in the same plot. The figure illustrates that after some time the metrics stabilize. There are still some random fluctuations, but no significant increasing or decreasing trends can be found.

The reason for the remaining perturbations can be explained by diversity controlling mechanism. Most classical MOEAs, such as NSGA-II [4] and SPEA2 [18], have a mechanism to maintain diversity in the objective space (e.g., crowding distance, archive truncation). The perturbations reflect that some individuals are moved from a crowded region to a less crowded region by means of the diversity controlling mechanism. Figure 2 illustrates the Pareto fronts of the 140$^{\text{th}}$ and 141$^{\text{th}}$ generation. The region covered by a rectangle illustrates the behaviour discussed before. Within the zoomed view in the upper part of the figure, one can see the re-positioning of an individual in the 140$^{\text{th}}$ generation marked by '+', which is marked by '□' in the 141$^{\text{th}}$ generation. Therefore, if the trend (figure 1) stabilizes for a longer time, we conclude that significant improvements are unlikely.

To detect stability, a regression analysis is performed. More specifically, a two-sided t-test is carried out to check the significance of the decreasing linear trend [15]. Using the test, it is possible to measure the slope (β) of the indicator values using a least-squares method [15]. Moreover, a statistical hypothesis test (i.e., t-test) is carried out to determine the significance of β. Since we want to detect the disappearance of the trend (i.e., $\beta = 0$), the statement for the hypothesis test can be expressed as:

$$H_0 : \beta = 0 \text{ vs } H_1 : \beta \neq 0 \quad (9)$$

Consequently, we try to detect the generation where the null hypothesis cannot be rejected anymore. Figure 1 shows the

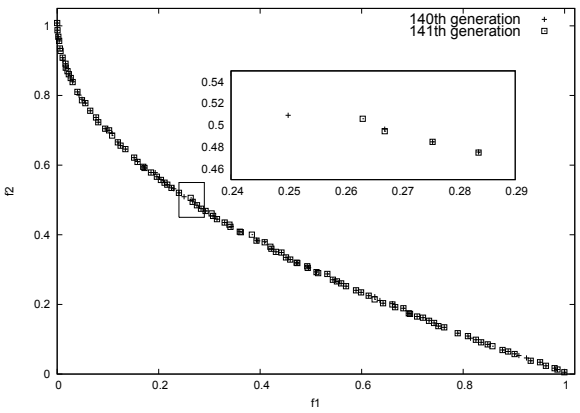

Figure 2: Pareto fronts extracted by NSGA-II for two consecutive generations.

stopping generations (three vertical lines) when the test is applied individually and concurrently on different metrics (i.e., AHD and diversity). It is clear from preliminary results that each metric stabilizes at different time. Moreover, by considering two metrics simultaneously, the algorithm stops later. Hence, it is possible to detect convergence more robustly by monitoring the objective and decision space concurrently.

Algorithm 1 presents the details of the proposed method. There are four intuitive parameters that need to be specified. The user has to provide a significance level (α) for the statistical test. Mainly two values are found: 0.05 (standard) and 0.1 (conservative) [15]. A user needs to specify the number of previous generations ($nGenLT$), for which AHD and *Diversity* values are considered in order to estimate the slope of the linear regression model. The next parameter ($nGenUnCh$) is the number of preceding generations for which no significant improvement can be obtained (i.e., H_0 cannot be rejected, p-value$>\alpha$). Finally, the last parameter ($MaxGen$) is the number of maximum allowable generations. AHD and *Diversity* are calculated for each generation (step 9 and 10). After $nGenLT$ generations, p-values for AHD and diversity are calculated (step 12 and 13). To calculate p-values for corresponding metrics the previous $nGenLT$ metric values are considered. A MOEA will be stopped when either the algorithm already evolved for $MaxGen$ generations or the p-values for both metrics (AHD and *Diversity*) are larger than α for the previous $nGenUnCh$ generations (step 15).

4. EXPERIMENTS AND RESULTS

A number of experiments has been conducted to analyze the performance of the proposed MOEA stopping criterion. For the experiments, we have selected two widely used MOEAs (i.e., NSGA-II [4] and SPEA2 [18]) and six benchmark problems (*ZDT1, ZDT2, ZDT3, ZDT4, DTLZ2, DTLZ5*) [17, 5] providing different characteristics of the Pareto fronts. All the problems of the ZDT family are bi-objective, whereas DTLZ2 and DTLZ5 are three-objectives problems. Our concurrent approach will be compared with an individual monitoring of each space. Additionally, it will be compared with a state-of-the-art stopping criterion, called *online convergence detection (OCD)* [15]. To run OCD, the framework proposed in [14] is used. Finally, a comparison with standard budget recommendations will be presented.

4.1 Experimental settings

The proposed methodology is implemented in jMetal [6], a multi-objective meta-heuristic framework developed in JAVA. The framework contains a large set of implemented meta-heuristics including NSGA-II and SPEA2, as well as many benchmark problems. Table 1 shows the standard parameter settings for our experiments. Moreover, we have used simulated binary crossover [2], polynomial mutation [2] and binary tournament selection [2] for both of the algorithms. Additionally, we set $nGenLT = 30$, $nGenUnCH = 10$ and $\alpha = 0.05$ for the experiments. $p = 2$ is used to calculate AHD. For each problem, the algorithms independently run *30* times. Moreover, for each run, we let the algorithms evolve *500* generations. The reason of letting the algorithms to run more than standard $MaxGen$ (table 2) is that we want to investigate when exactly the proposed stopping method is activated. Finally, we have used the following parameters in the variance test of OCD for the three indicators: $\epsilon_{HV} = 1e^{-4}$, $\epsilon_R = 1e^{-6}$ and $\epsilon_{Epsilon} = 2e^{-4}$.

Table 1: Parameter settings for NSGA-II and SPEA2

	NSGA-II	SPEA2
Population size	100	100
Archive Size	–	100
Crossover probability	0.9	0.9
Mutation Probability	$1/l$	$1/l$
Distribution Index	20	20

4.2 Evaluation metrics

To evaluate our proposed method, we have defined four metrics/indicators. The first indicator is the difference between the hypervolumes (HV) [19] achieved by the different approaches ($HV_i : i \in \{AHD, Div, AHD{+}Div, OCD\}$) and the HV achieved after the standard $MaxGen$ (i.e., $HV_i - HV_{std}$), we call the difference HV_d. In the same manner, differences between the epsilon [20] indicator values ($eps_d = eps_{std} - eps_i : i \in \{AHD, Div, AHD{+}Div, OCD\}$) are computed. The Pareto reference fronts provided by the jMetal framework [1] are used to calculate the epsilon indicator. Negative HV_d values corresponds to a smaller HV achieved when the algorithm is stopped by the approach compared to the standard $MaxGen$. In contrast, negative eps_d values mean the epsilon indicator is larger when the algorithm is stopped by the approach than the one standard $MaxGen$. Therefore, HV_d and eps_d values close to 0 indicate a good approximation of the Pareto front. The standard stopping generations for the different problems can be found in table 2. Additionally, we want to measure the percentage of the HV [9] of true Pareto front obtained by an algorithm at the time the algorithm is stopped by the combined approach ($AHD + Div$). The calculation is done by taking the ratio between the HV achieved by the algorithm and the HV of the true Pareto front (i.e., $HV_{per} = \frac{HV_{AHD+Div}}{HV_{tPF}}$). We again use the Pareto-reference fronts provided by jMetal framework. We will also report the average number of function evaluations (*NoSFE*) saved for each problem. The *NoSFE* will be calculated by

[1]http://jmetal.sourceforge.net/problems.html

Algorithm 1 Algorithm for detecting stopping generation

Require:

α ▷ Significance level for statistical test
$nGenLT$ ▷ Number of previous generations considered for calculating slope of linear regression model
$nGenUnCh$ ▷ Number of previous generations the p-values remain unchanged ($pVal_j > \alpha, j \in \{AHD, Div\}$)
$MaxGen$ ▷ Maximum number of allowed generations
$i = 0$
repeat
  $i \leftarrow i + 1$
  $AHD[i] \leftarrow$ AHD between Pareto front of i^{th} and $(i-1)^{th}$ generation
  $Div[i] \leftarrow$ Diversity of i^{th} generation
  if $i > nGenLT$ **then**
    $pAHD[i] \leftarrow pValue$ based on regression analysis of the previous $nGenLT$ values of AHD
    $pDiv[i] \leftarrow pValue$ based on regression analysis of the previous $nGenLT$ values of Div
  end if
until $i \geq MaxGen$ **or** $(\forall_j \in \{i, i-1, \ldots, i-nGenUnCh\} : pAHD[j] > \alpha \wedge pDiv[j] > \alpha\ )$

$(MaxGen - Gen_{AHD+Div}) * PS$, where $Gen_{AHD+Div}$ is the stopping generation number of the combined approach and PS is the population size.

Table 2: Standard maximum allowed number of generations for different problems [15]

Problem	ZDT family	DTLZ2	DTLZ5
Standard *MaxGen*	200	300	200

Table 3: Means and standard deviations of the evaluation metrics on the different problems for our proposed approach

Problem	Evaluation metrics	NSGA-II mean	NSGA-II SD	SPEA2 mean	SPEA2 SD
ZDT1	NoSFE	5527	2382	643	1153
	HV_d	-5.89E-03	3.54E-03	-6.24E-04	1.19E-03
	eps_d	-3.06E-03	3.33E-03	-5.40E-04	1.21E-03
	HV_{per}	0.979285	0.005466	0.986092	0.001775
ZDT2	NoSFE	2793	1887	0	0
	HV_d	-2.68E-03	2.04E-03	0.00E+00	0.00E+00
	eps_d	-1.63E-03	2.38E-03	0.00E+00	0.00E+00
	HV_{per}	0.966308	0.006051	0.957989	0.042851
ZDT3	NoSFE	5077	2427	1427	2509
	HV_d	-4.36E-03	2.88E-03	-3.46E-03	1.40E-02
	eps_d	-4.71E-03	4.35E-03	-5.37E-03	2.38E-02
	HV_{per}	0.984483	0.00566	0.983046	0.027052
ZDT4	NoSFE	777	1864	953	2582
	HV_d	-2.30E-02	6.62E-02	-4.03E-02	1.29E-01
	eps_d	-2.04E-02	6.17E-02	-7.41E-02	2.42E-01
	HV_{per}	0.925081	0.109336	0.874061	0.185557
DTLZ2	NoSFE	18323	2410	15173	6152
	HV_d	-3.13E-03	7.17E-03	-3.85E-03	4.81E-03
	eps_d	3.52E-03	3.21E-02	-2.45E-03	1.18E-02
	HV_{per}	0.790937	0.011261	0.857519	0.008931
DTLZ5	NoSFE	3113	3467	2767	4192
	HV_d	-8.90E-05	2.66E-04	-4.36E-04	7.50E-04
	eps_d	-2.66E-04	1.83E-03	-6.47E-04	1.58E-03
	HV_{per}	0.969181	0.003113	0.968569	0.00784

4.3 Discussion

Figure 3 and 4 present comparisons of the four different stopping approaches with respect to the stop generation (StopGen), HV_d and eps_d on the different test scenarios. Each row of the figures presents a problem and each column presents the performance with regard to a particular indicator. Horizontal lines in the plots of the first column indicate the standard $MaxGen$. Moreover, horizontal lines in the plots of the second and third columns are drawn to indicate the zero level for the corresponding indicators.

It is clear from the figures that for all the problems and algorithms, the combined approach runs longer than the individual approaches. At the same time, the combined approach has a better performance with regard to the Pareto front approximation. It is interesting to point out that there is no clear winner between AHD and Div. Therefore, relying on a particular space may detect converge prematurely. This clarifies the fact that the simultaneous monitoring of both spaces is more reliable to detect the convergence of an algorithm.

The figures also illustrate that the algorithms run much longer for OCD than for the combined approach on all ZDT problems. However, on DTLZ problems, the algorithms with OCD stop earlier. In contrast, using our method the algorithms stop more reliably with well-approximated Pareto fronts. Therefore, the simultaneous monitoring approach is more consistent, whereas OCD is not so stable when confronted with different problems.

In this paragraph, we will compare our method ($AHD + Div$) with standard budget recommendations (table 2). Table 3 reports the means and standard deviations of the respective indicator values. We have found very well approximated Pareto fronts for all problems and both algorithms, except on *ZDT4*. The means and standard deviations of HV_d and eps_d are very low for all these problems.The good quality Pareto fronts are found while also obtaining reasonably good *NoSEF*s. The only exception is *ZDT2* for SPEA2, where the proposed method does not converge before $MaxGen$. Very good percentages of true HVs (HV_{per}) are obtained. HV_{per} has reached over 95% for all the problems, except *DTLZ2* and *ZDT4*. HV_{per}s are considerably low (0.7916) for NSGA-II and SPEA2 (0.8557) on the *DTLZ2* problem. However, please note that the average HV_{per}s are 0.7976 and 0.8657 for standard $MaxGen$ for NSGA-II and SPEA2, respectively. Therefore, there is almost no difference in terms of achieved HV_{per} when comparing our method with standard $MaxGen$. Nevertheless, sometimes the method detects convergence prematurely on the *ZDT4* problem. It should be noted that the *ZDT4* problem has 21^9 local Pareto fronts. Finally, in comparison to standard

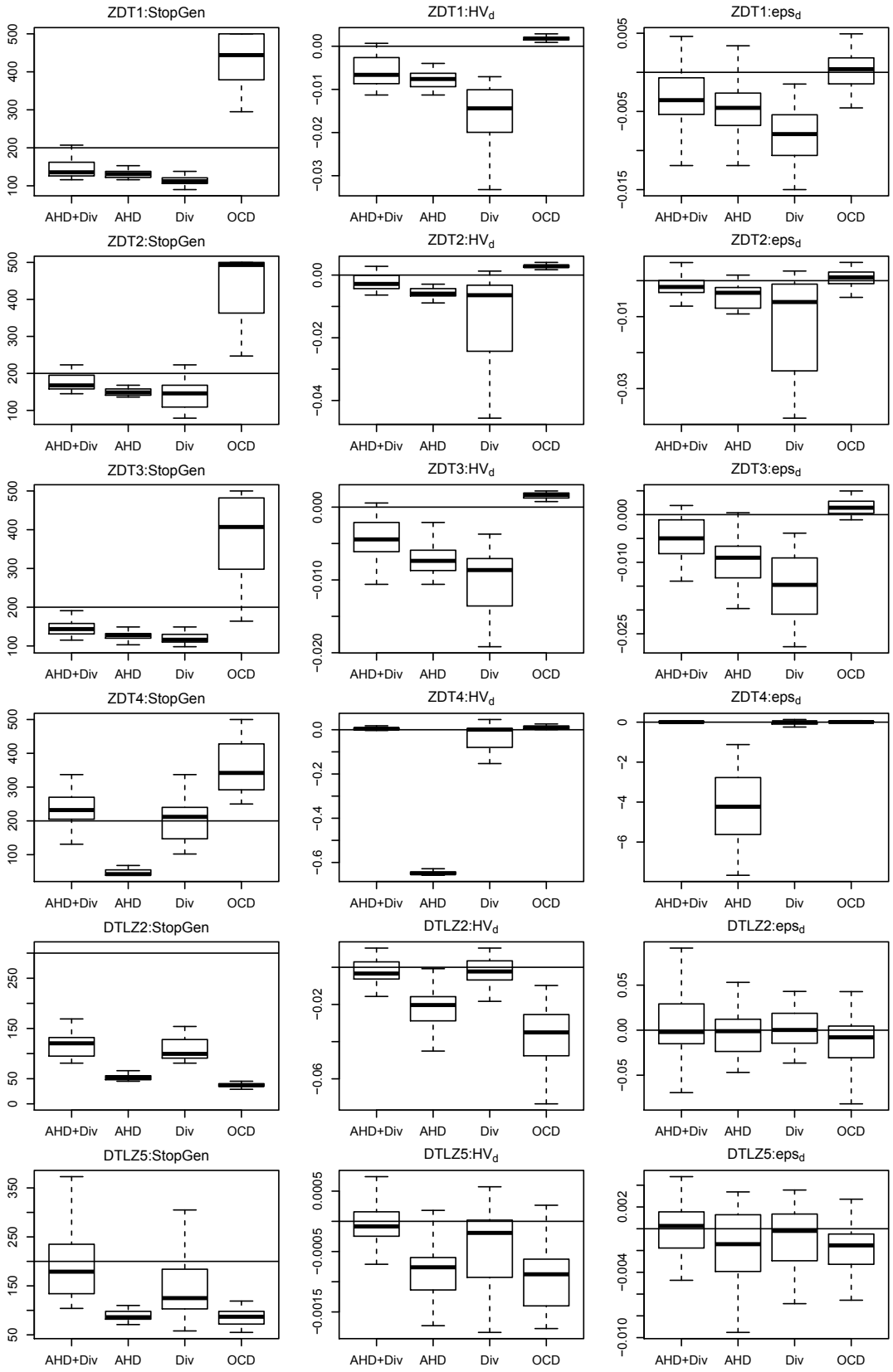

Figure 3: Boxplots for *stopping generation*, HV_d and eps_d on the different problems for NSGA-II

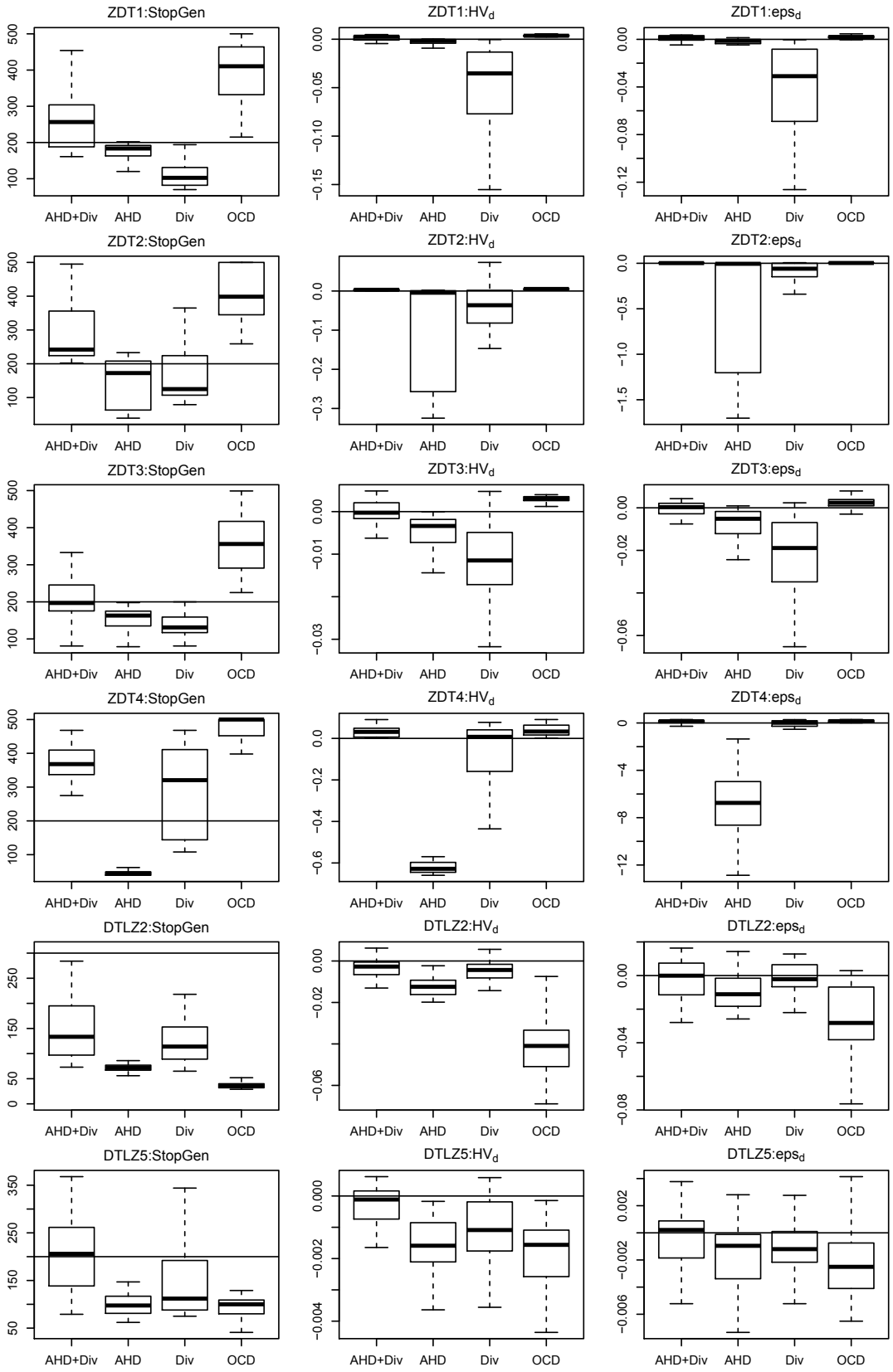

Figure 4: Boxplots for *stopping generation*, HV_d and eps_d on the different problems for SPEA2

budget recommendation, on average our proposed method saves 29% and 17% of function evaluations for NSGA-II and SPEA2, respectively.

5. CONCLUSION AND FUTURE WORK

In this paper, we have shown that the simultaneous monitoring of the objective and decision space of a problem, can improve the robustness of stopping MOEA. Two metrics have been proposed for the different spaces. The stagnation has been detected by a two-sided t-test on the regression coefficients estimated from the matrices in the two spaces.

We have validated our method by investigating the performance on six different benchmark problems for two well-established MOEAs. Our proposed method performs consistently well for all problems, also compared to state-of-the-art stopping criteria. Additionally, on average a decent number of function evaluations has been saved compared to fixed budget recommendations without losing significant approximation accuracy.

In future work, we will investigate the combination of monitoring the decision space diversity with other objective space metrics, such as the hypervolume.

6. REFERENCES

[1] L. T. Bui, S. Wesolkowski, A. Bender, H. A. Abbass, and M. Barlow. A dominance-based stability measure for multi-objective evolutionary algorithms. In *IEEE Congress on Evolutionary Computation*, pages 749–756. IEEE, 2009.

[2] K. Deb. *Multi-objective optimization using evolutionary algorithms*, volume 16. John Wiley & Sons, 2001.

[3] K. Deb and S. Jain. Running performance metrics for evolutionary multi-objective optimizations. In *Proceedings of the Fourth Asia-Pacific Conference on Simulated Evolution and Learning (SEAL'02)*, pages 13–20, 2002.

[4] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, 2002.

[5] K. Deb, L. Thiele, M. Laumanns, and E. Zitzler. Scalable test problems for evolutionary multiobjective optimization. In *Evolutionary Multiobjective Optimization*, Advanced Information and Knowledge Processing, pages 105–145. Springer London, 2005.

[6] J. J. Durillo and A. J. Nebro. jMetal: A Java framework for multi-objective optimization. *Advances in Engineering Software*, 42:760–771, 2011.

[7] T. Goel and N. Stander. A non-dominance-based online stopping criterion for multi-objective evolutionary algorithms. *International Journal for Numerical Methods in Engineering*, 84(6):661–684, 2010.

[8] L. Martí, J. García, A. Berlanga, and J. M. Molina. An approach to stopping criteria for multi-objective optimization evolutionary algorithms: the MGBM criterion. In *IEEE Congress on Evolutionary Computation*, pages 1263–1270. IEEE, 2009.

[9] A. J. Nebro, J. J. Durillo, C. A. C. Coello, F. Luna, and E. Alba. A study of convergence speed in multi-objective metaheuristics. In *Parallel Problem Solving from Nature–PPSN X*, pages 763–772. Springer, 2008.

[10] D. Roche, D. Gil, and J. Giraldo. Detecting loss of diversity for an efficient termination of EAs. In *15th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, pages 561–566, 2013.

[11] O. Rudenko and M. Schoenauer. A steady performance stopping criterion for Pareto-based evolutionary algorithms. In *6th International Multi-Objective Prpgramming and Goal Programming Conference*, 2004.

[12] O. Schutze, X. Esquivel, A. Lara, and C. A. Coello Coello. Using the averaged Hausdorff distance as a performance measure in evolutionary multiobjective optimization. *IEEE Transactions on Evolutionary Computation*, 16(4):504–522, 2012.

[13] H. Trautmann, T. Wagner, B. Naujoks, M. Preuss, and J. Mehnen. Statistical methods for convergence detection of multi-objective evolutionary algorithms. *Evolutionary computation*, 17(4):493–509, 2009.

[14] T. Wagner, H. Trautmann, and L. Martí. A taxonomy of online stopping criteria for multi-objective evolutionary algorithms. In *Proceedings of the 6th International Conference on Evolutionary Multi-criterion Optimization*, EMO'11, pages 16–30. Springer, 2011.

[15] T. Wagner, H. Trautmann, and B. Naujoks. OCD: Online convergence detection for evolutionary multi-objective algorithms based on statistical testing. In *Evolutionary Multi-Criterion Optimization*, volume 5467 of *Lecture Notes in Computer Science*, pages 198–215. Springer, 2009.

[16] M. Wineberg and F. Oppacher. The underlying similarity of diversity measures used in evolutionary computation. In *Genetic and Evolutionary Computation-GECCO 2003*, pages 1493–1504. Springer, 2003.

[17] E. Zitzler, K. Deb, and L. Thiele. Comparison of multiobjective evolutionary algorithms: Empirical results. *Evolutionary Computation*, 8(2):173–195, 2000.

[18] E. Zitzler, M. Laumanns, and L. Thiele. SPEA2: Improving the strength pareto evolutionary algorithm. Technical Report TIK-Report No. 103, Computer Engineering and Networks Laboratory (TIK), Swiss Federal Institute of Technology (ETH), Zurich, 2001.

[19] E. Zitzler and L. Thiele. Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach. *IEEE Transactions on Evolutionary Computation*, 3(4):257–271, 1999.

[20] E. Zitzler, L. Thiele, M. Laumanns, C. M. Fonseca, and V. G. Da Fonseca. Performance assessment of multiobjective optimizers: An analysis and review. *IEEE Transactions on Evolutionary Computation*, 7(2):117–132, 2003.