2013 IEEE Congress on Evolutionary Computation
June 20-23, Cancún, México

A New Performance Metric for User-preference Based Multi-objective Evolutionary Algorithms

Asad Mohammadi
School of Computer Science and IT
RMIT University
Melbourne, Australia
Email: asad.mohammadi@rmit.edu.au

Mohammad Nabi Omidvar
School of Computer Science and IT
RMIT University
Melbourne, Australia
Email: mohammad.omidvar@rmit.edu.au

Xiaodong Li
School of Computer Science and IT
RMIT University
Melbourne, Australia
Email: xiaodong.li@rmit.edu.au

***Abstract*—In this paper, we propose a metric for evaluating the performance of user-preference based evolutionary multi-objective algorithms by defining a preferred region based on the location of a user-supplied reference point. This metric uses a *composite front* which is a type of reference set and is used as a replacement for the Pareto-optimal front. This composite front is constructed by extracting the non-dominated solutions from the merged solution sets of all algorithms that are to be compared. A preferred region is then defined on the composite front based on the location of a reference point. Once the preferred region is defined, existing evolutionary multi-objective performance metrics can be applied with respect to the preferred region. In this paper the performance of a cardinality-based metric, a distance-based metric, and a volume-based metric are compared against a baseline which relies on knowledge of the Pareto-optimal front. The experimental results show that the distance-based and the volume-based metrics are consistent with the baseline, showing meaningful comparisons. However, the cardinality-based approach shows some inconsistencies and is not suitable for comparing the algorithms.**

I. Introduction

Several user-preference based evolutionary multi-objective algorithms (EMO) have been proposed in the past decade [1], [2], [3], [4], [5], [6], [7]. Incorporating the user-preference information into the optimization process is an appealing idea for several major reasons: 1) it allows a more directed search which results in faster convergence than in cases where the entire Pareto-optimal front needs to be approximated; 2) it allows tackling of problems with relatively higher numbers of objectives (better scalability); 3) it allows the user to search for better solutions in the vicinity of an existing solution.

Despite the growing interest in developing user-preference based algorithms, very few performance measures have been developed to facilitate a fair comparison of such algorithms. A metric which has been recently developed by Wickramasinghe *et al.* [8] is specifically designed for comparing user-preference based EMO algorithms. However, a major drawback of this metric is that its results can be misleading, depending on the choice of the reference point (cf. section III). An ideal metric for user-preference based algorithms should have the following properties:

1) Form a preferred region closest to the reference point provided by the user;
2) Measure both convergence and diversity of the solutions with respect to the preferred region;
3) Be independent of knowledge of Pareto-optimal front for its calculation;
4) Scale well as the number of objectives increases.

Many performance metrics have been proposed for comparing EMO algorithms. However, there has not been any suitable metric for evaluating user-preference based EMO algorithms that satisfy all four properties listed above. For example, *cardinality-based* metrics [9], [10] do not depend on knowledge of the Pareto-optimal front, but they cannot measure the diversity of obtained solutions. On the other hand, some *distance-based* metrics [9], [11], take both the convergence and the diversity of solutions into account, but they rely on knowledge of the Pareto-optimal front for their calculations. The techniques that rely on sampling of the Pareto-optimal front generally do not scale well as the number of objectives increases. This is because of exponential growth in the number of sample points required on the Pareto-optimal front.

In this paper, we propose a performance metric for comparing user-preference based EMO algorithms that borrows the idea of a *reference set* [12] from cardinality-based metrics to form a *composite front* that acts as a replacement for the Pareto-optimal front. This composite front is then used to define a preferred region based on the location of a user-supplied reference point. Once the preferred region is defined, existing EMO metrics can be applied to the preferred region.

The paper is organized in the following way. Section II briefly describes various existing performance metrics for the EMO algorithms. Section III gives a survey of the existing performance metrics for user-preference based EMO algorithms. Section IV describes the details of the proposed metric. Experimental results and their analysis are presented in Section V, and Section VI concludes the paper.

II. Background

This section gives an overview of some widely used metrics for evaluating multi-objective evolutionary algorithms.

The performance of EMO algorithms are typically measured on the following two aspects: 1) closeness of the solutions to the Pareto-optimal front (convergence); 2) the diversity and the spread of the solutions. A property which is often overlooked is that the metric should measure the performance of a set of algorithms without relying on knowledge of the Pareto-optimal front. This problem becomes more serious especially

when the Pareto-optimal front is difficult to compute, or when it is unknown, which is mostly the case in many real-world problems. In the remainder of this section, we review some existing performance metrics for EMO algorithms. The major classifications of metrics presented in this section are adopted from [13].

A. Cardinality-based Metrics

These metrics measure the performance of various algorithms by counting the total number of non-dominated solutions found by each algorithm [9], [10]. However, producing a large number of non-dominated solutions does not necessarily make an algorithm better than another. For example, an algorithm may have only one solution that dominates all the solutions of another algorithm. In order to alleviate this problem, many cardinality-based approaches rely on a *reference set* and measure the contribution from each algorithm with respect to this reference set [9], [14], [15], [16], [17].

There are many different ways of constructing a reference set. For example, a reference set may be formed by aggregating all known solutions to a problem by various means, or by merging the solution sets that are generated by a set of algorithms that are to be compared [12]. These reference sets may contain all possible solutions, or just the non-dominated solutions. It should be noted that the ranking of a set of algorithms may change depending on the choice of the reference set [12]. The reference set used in this paper is formed by taking the non-dominated solutions from the merged solution sets of several algorithms.

1) Set Convergence Metric: This metric is used to measure the relative convergence of two solution sets with respect to each other [18]. Let A, and B be the solution sets of two different algorithms. $\mathcal{C}(A, B)$ is calculated as follows:

$$\mathcal{C}(A, B) = \frac{|\{b \in B | \exists a \in A : a \preceq b\}|}{|B|}, \tag{1}$$

If $\mathcal{C}(A, B) = 1$ then A dominates all members of B, and if $\mathcal{C}(A, B) = 0$, none of the solutions from B are dominated by A. The result of $\mathcal{C}$ metric is not always reliable. For instance, there are cases where the surface covered by two fronts are equal, but one front is closer to the Pareto-optimal front than the other.

2) Convergence Difference Of Two Sets: This metric, which is called $\mathcal{D}$ metric [15], is an improved version of the $\mathcal{C}$ metric. Let A and B be the solution sets of two different algorithms. Then $\mathcal{D}(A, B)$ is the size of the region which is *only* dominated by solutions in A, and $\mathcal{D}(B, A)$ is the size of the region which is *only* dominated by solutions in B. For a maximization problem if $\mathcal{D}(A, B) < \mathcal{D}(B, A)$, then it is concluded that B dominates A.

Both $\mathcal{C}$ and $\mathcal{D}$ metrics do not measure the diversity and spread of the solutions. Additionally, these two metrics are not efficient when comparing more than two algorithms. Another major drawback of these two metrics is that they become increasingly inaccurate as the number of objectives increases. The values for $\mathcal{C}(A, B)$ converge to $\mathcal{C}(\mathcal{B}, \mathcal{A})$ as the number of objectives increases. This is due to the fact that most of the solutions in many-objective problems are non-dominated to each other. Therefore, the areas covered by the two sets of solutions become equal.

B. Distance-based Metrics

Generational Distance (GD) [9] is a metric used widely to measure the convergence of EMO algorithms by calculating the average closest distances of obtained solutions to the Pareto-optimal front. More precisely, the GD value is calculated in the following way. Let Q be the obtained solution set and P^* a set of non-dominated solutions on the Pareto-optimal front. Then,

$$GD(Q, P^*) = \frac{\sum_{v \in Q} d(v, P^*)}{|Q|} \tag{2}$$

where $|Q|$ is the number of solutions in Q, and $d(v, P^*)$ is the closest Euclidean distance from point v to a point in P^*. Since GD is calculated based on the Pareto-optimal front, it gives an accurate measure for the convergence of an algorithm. However, GD does not measure the diversity and spread of the solutions on the Pareto-optimal front. Another disadvantage of GD is that it becomes difficult to calculate when dealing with many-objective problems. For a reasonable sampling of the Pareto-optimal front, a large number of points are required, which makes the calculation of GD computationally expensive. It should be noted that GD cannot be applied without the existence of a reference front such as the Pareto-optimal front.

Inverted Generational Distance (IGD) [11] is an improved version of GD that takes both diversity and convergence of solutions into account. However unlike GD, which calculates the average closest distance of solutions to the Pareto-optimal front, it calculates the average closest distance of sample points on the Pareto-optimal front to the obtained solutions. Therefore,

$$IGD(P^*, Q) = \frac{\sum_{v \in P^*} d(v, Q)}{|P^*|}. \tag{3}$$

A major advantage of IGD is that it can measure both convergence and diversity of the solutions simultaneously. Similar to GD, IGD becomes exponentially expensive as the number of objectives increases.

C. Volume-based Metrics

Hypervolume Metric (HV) [9], [19] is a metric which measures the volume between all solutions in an obtained non-dominated set and a nadir point. A nadir point is a vector of the worst objective function values obtained by the solution set. To calculate the HV value, a set of hypercubes (c_i) is constructed by taking a solution i and the nadir point as its diagonal corners. Finally, the HV value is the total volume of all hypercubes.

$$HV = \text{volume}\left(\bigcup_{i=1}^{|Q|} c_i\right), \tag{4}$$

where Q is the solution set. Higher HV values indicate a better convergence and diversity of solutions on the Pareto-optimal front. A major advantage of HV is that it does not depend on knowledge of the Pareto-optimal front.

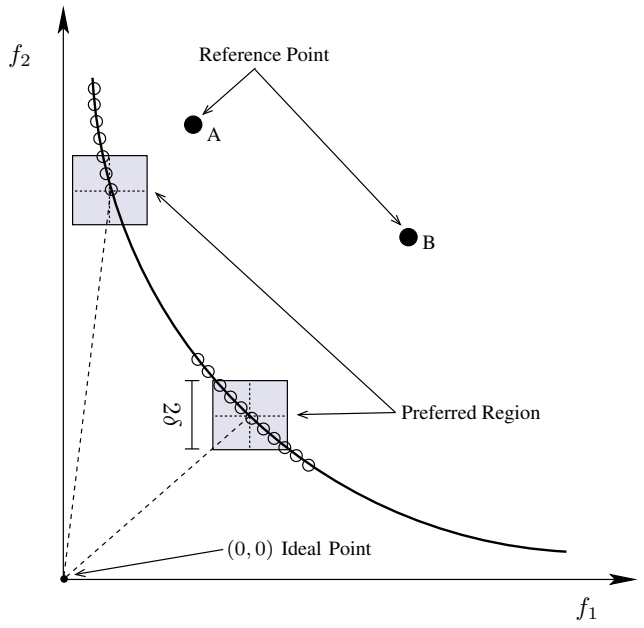


Fig. 1. An example to depict the deficiency of the metric proposed in [8].

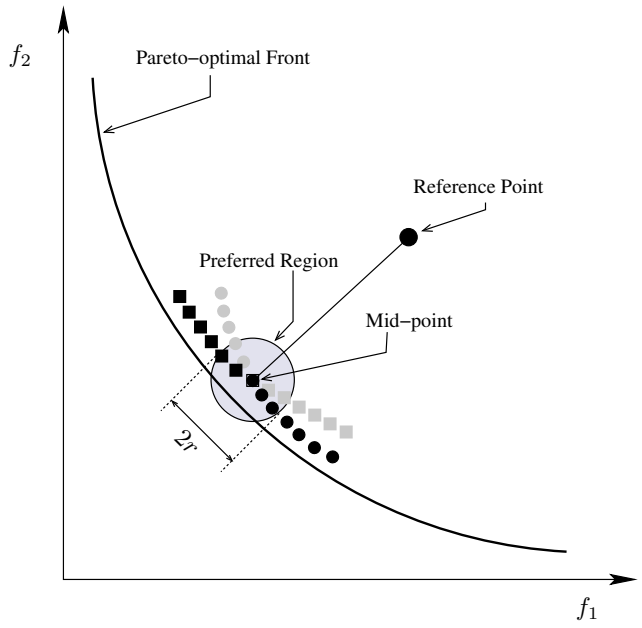


Fig. 2. An example of a composite front which is used to define a preferred region.

III. Related Work

All of the metrics discussed so far were designed to measure the performance of EMO algorithms that approximate the entire Pareto-optimal front. There has been limited work on developing metrics for comparing user-preference based EMO algorithms.

To the best of our knowledge, Wickramasinghe *et al.* [8] were the first to propose a metric for comparing user-preference based EMO algorithms. This metric works by combining the solution sets of all algorithms that need to be compared. Then, the closest solution to the ideal point is used as the center of a hypercube that defines a preferred region. Figure 1 shows how a preferred region is defined for two different reference points. The size of the preferred region is determined by a parameter, δ, which is half the edge length of the hypercube. Finally, for each of the algorithms, HV is calculated with respect to a nadir point for all the solutions that fall within the preferred region. To calculate the nadir point, this metric uses the solutions from all algorithms inside the preferred region. The choice of the ideal point is the origin of the coordinate system for minimization problems.

An advantage of this metric is that it does not require knowledge of the Pareto-optimal front. However, its major drawback is that it defines the preferred region based on the location of the ideal point. This causes misleading results when the reference point is biased towards one objective more than the other objectives. This effect is shown in Figure 1. It can be seen that the solutions for reference point A converged on the Pareto-optimal front with a minimum distance to the reference point. However, a bad choice of the ideal point causes many high quality solutions to fall outside the preferred region. This shows that the results of this metric can be misleading depending on the location of the reference point.

IV. Proposed Metric

In this section, we propose a metric to evaluate the performance of user-preference based multi-objective evolutionary algorithms.

In a nutshell, the proposed metric which hereafter is called user-preference metric based on a composite front (UPCF), merges the solution set of all algorithms and uses the non-dominated solutions of the merged solution sets as a replacement for the Pareto-optimal front. This so-called *composite front* is a type of reference set commonly used in several cardinality-based metrics. The composite front is then used to form a preferred region based on the position of a reference point provided by the decision maker. Finally, the performance of each algorithm is measured by calculating IGD or HV for solutions of each algorithm which are within the preferred region. UPCF can be coupled with either IGD or HV. In this paper both of these two popular techniques are used for the sake of comparison. Measuring both convergence and diversity of the solution set makes both IGD and HV desirable candidates for this new metric. The detailed procedure for applying UPCF is as follows:

Step 1 - Generating a Composite Front: The solution set of all the algorithms to be compared are merged, and all non-dominated solutions from this merged set are placed in another set called the *composite front*. In Figure 2, squares and circles show the solution sets for two different user-preference based algorithms. The solutions shown as black squares form the composite front, and the solutions shown as gray circles are those dominated by at least one solution in the composite front.

Step 2 - Generating a Preferred Region For Each Reference Point: To define the preferred region, the Euclidean distances between all the solutions in the composite front and a reference point is calculated. Then the solution with the

least distance to the reference point is identified. This point is called *mid-point* as shown in Figure 2. Finally, the solutions within r distance of the *mid-point* are considered to be in the preferred region. The parameter r is specified by the user, which determines the size of the preferred region. In the real-world applications where objectives do not have the same units, objectives should be normalized otherwise the parameter r will not be meaningful.

Step 3 - Calculating IGD and HV: IGD and HV are calculated based on the solutions inside the preferred region. To calculate the IGD values, instead of using sample points on the Pareto-optimal front, the solutions in the composite front are used. The IGD based on the composite front is abbreviated as IGD-CF.

A major advantage of UPCF is that it can be applied in situations where the Pareto-optimal front is unknown. This property has significant implications for the scalability and computational cost of the metric. For example, many distance-based approaches, such as GD and IGD, require a set of sample points on the Pareto-optimal front. For small problems with two or three objectives, it is easy to generate a set of points on the Pareto-optimal front. However, as the number of objectives increases, the cost of this process grows exponentially. In addition to the computational cost of sampling the Pareto-optimal front, for many real-world problems the Pareto-optimal front is either very difficult to generate or completely unknown.

V. Simulation Results

In order to evaluate the effectiveness of the proposed metric, it has been tested on three user-preference based algorithms namely R-NSGA-II [4], R-MEAD-Te [5], and R-MEAD-Ws [5]. R-NSGA-II is a modified version of the popular NSGA-II [20] algorithm that can handle multiple reference points. R-MEAD-Te and R-MEAD-Ws are two user-preference based algorithms which are based on the MOEA/D algorithm [21]. R-MEAD-Te and R-MEAD-Ws rely on the Tchebycheff [22] and Weighted-Sum [22] decomposition methods respectively, to convert a multi-objective optimization problem into a single-objective problem.

To understand whether UPCF coupled with IGD-CF and HV is an accurate metric to measure the performance of a user-preference based algorithm, their results have been compared with IGD based on the Pareto-optimal front (IGD-OF). In addition to IGD-CF and HV, the average number of solutions that each algorithm contributes to the composite front (NS-CF) is also reported for further analysis.

The benchmark problems that are used in this paper are two-objective ZDT1-ZDT4 and ZDT6 functions, and three-objective DTLZ1-DTLZ6 functions. We used $(0.7, 0.2)$ and $(0.2, 0.2, 0.6)$ as reference points for two-objective and three-objective test problems respectively. A different reference point $(0.2, 0.4, 0.9)$ is used for DTLZ1 since the point $(0.2, 0.2, 0.6)$ is located on its Pareto-optimal front. The population size for two-objective test problems has been set to 50. The number of iterations in each run is 150 for ZDT1 and ZDT2, 300 for ZDT3, and 500 for ZDT4 and ZDT6.

The population size is set to 200 for three-objective problems. The number of iterations in each run is 200 for DTLZ1, DTLZ2, DTLZ5 and DTLZ6, and 400 for DTLZ3 and DTLZ4.

TABLE III. Nadir points for all test problems

Test problem	Nadir Point
ZDT1	(0.87, 0.30)
ZDT2	(1.00, 0.60)
ZDT3	(0.85, 1.00)
ZDT4	(1.00, 28.59)
ZDT6	(1.00, 2.77)
DTLZ1	(2.00, 1.03, 2.00)
DTLZ2	(0.37, 0.37, 1.00)
DTLZ3	(0.85, 1.00, 1.00)
DTLZ4	(1.00, 1.00, 0.96)
DTLZ5	(0.48, 0.48, 1.00)
DTLZ6	(0.43, 0.43, 1.00)

To specify the size of the preferred region on the composite front and the Pareto-optimal front, the parameter r (see Section IV) is set to 0.1 for all test problems. The ϵ parameter of R-NSGA-II is set to 0.001 and 0.002 for two- and three-objective problems respectively. The radius parameter of R-MEAD-Te and R-MEAD-Ws, which can be used to control the size of the preferred region, is set to 0.05 and 0.02 for Tchebycheff and Weighted-Sum approaches respectively. The initial population size is set to 100 and 250 for two- and three-objective problems respectively. The nadir point used by HV is calculated by taking the worst objective value for each of the objective functions from all solutions generated by all three algorithms in 25 independent runs. Table III shows the nadir points calculated for different problems.

Tables I and II show the mean and the standard deviation for 25 independent runs of R-NSGA-II, R-MEAD-Te and R-MEAD-Ws using four different performance measures. As mentioned previously, IGD-OF is not part of the proposed metric and is solely used as a baseline against which other algorithms are compared. The last three columns are the results of t-test (p-values) using a 95% confidence interval.

A. Two-Objective Test Problems

The ZDT1 test problem has a convex Pareto-optimal front. Table I shows the result of R-NSGA-II, R-MEAD-Te and R-MEAD-Ws on this test problem. The results of statistical test shows that R-NSGA-II and R-MEAD-Te are not significantly different using the HV, IGD-OF and IGD-CF measures. However, we can significantly distinguish the performance of R-MEAD-Ws from R-NSGA-II and R-MEAD-Te. Results of IGD-CF and HV are consistent with IGD-OF, which suggests that R-MEAD-Ws performs significantly better than the other two algorithms. However, the conclusion is different when NS-CF is used. Figures 3(a), 3(b), and 3(c) also show this behavior.

The next test problem is ZDT2, which has a non-convex Pareto-optimal front. According to p-values in Table I, all three methods are significantly different. The results of all three measures are consistent with IGD-OF, which suggests that R-MEAD-Te outperforms the other two algorithms. Figures 3(d), 3(f) and 3(e) visually confirm the results generated by the metrics. On ZDT3, the only measure which is consistent with IGD-OF is IGD-CF. The test results of IGD-CF and IGD-OF show that none of the algorithms performs significantly better than the others. However, the conclusion is different when HV and NS-CF measures are used. On the

TABLE I. RESULTS ON THE 2-OBJECTIVE TEST PROBLEMS. THE MEAN AND STANDARD DEVIATION OF 25 INDEPENDENT RUNS ARE REPORTED. THE STATISTICAL SIGNIFICANCE RESULTS ARE BASED ON THE T-TEST USING A 95% CONFIDENCE INTERVAL.

Func.	Metric	R-MEAD-Te	R-MEAD-Ws	R-NSGA-II	R-MEAD-Te vs R-NSGA-II	R-MEAD-Te vs R-MEAD-Ws	R-MEAD-Ws vs R-NSGA-II
ZDT1	HV	2.67e-02 (6.10e-03)	**3.66e-02 (8.56e-04)**	2.70e-02 (2.20e-03)	8.30e-01	4.79e-08	1.84e-16
	IGD-CF	5.70e-03 (1.43e-02)	**4.31e-04 (2.70e-04)**	2.50e-03 (1.00e-03)	2.74e-01	2.57e-09	1.24e-11
	NS-CF	3.96e+01 (1.21e+01)	3.68e+01 (3.60e+00)	**4.95e+01 (9.18e-01)**	4.44e-04	3.24e-01	2.32e-15
	IGD-OF	5.60e-03 (1.02e-02)	**7.44e-04 (3.28e-04)**	3.50e-03 (7.06e-04)	3.18e-01	2.81e-02	2.99e-14
ZDT2	HV	**5.81e-02 (1.20e-03)**	0.00e+00 (0.00e+00)	5.16e-02 (1.13e-02)	9.20e-03	1.88e-42	8.45e-18
	IGD-CF	**3.60e-03 (1.30e-03)**	1.04e-01 (1.23e-02)	1.05e-02 (2.42e-02)	6.60e-03	6.61e-23	4.42e-17
	NS-CF	**4.75e+01 (3.51e+00)**	0.00e+00 (0.00e+00)	2.77e+01 (1.61e+01)	3.93e-06	6.61e-29	8.90e-09
	IGD-OF	**1.00e-03 (1.01e-04)**	2.80e-02 (7.08e-18)	3.60e-03 (5.10e-03)	1.60e-09	5.40e-60	3.18e-18
ZDT3	HV	**1.12e-01 (1.33e-01)**	7.77e-02 (1.12e-01)	3.16e-02 (6.64e-02)	2.83e-02	1.39e-01	1.34e-01
	IGD-CF	3.12e-01 (7.46e-01)	3.12e-01 (7.47e-01)	3.11e-01 (7.29e-01)	8.49e-01	6.67e-01	9.13e-01
	NS-CF	**2.15e+01 (2.40e+01)**	**2.20e+01(2.53e+01)**	4.80e+00 (9.46e+00)	9.50e-03	9.40e-01	1.02e-02
	IGD-OF	1.74e-01 (4.84e-02)	1.80e-01 (4.42e-02)	1.78e-01 (6.04e-02)	8.65e-01	9.77e-01	8.80e-01
ZDT4	HV	**9.63e+00 (1.17e-01)**	0.00e+00 (0.00e+00)	**9.48e+00 (6.57e-01)**	3.40e-01	1.06e-47	1.43e-29
	IGD-CF	**7.20e-04 (6.49e-04)**	8.96e-02 (1.23e-02)	**1.00e-03 (7.42e-04)**	1.56e-01	1.38e-22	1.30e-22
	NS-CF	**3.32e+01 (1.34e+01)**	0.00e+00 (0.00e+00)	**3.26e+01 (1.96e+01)**	9.32e-01	6.94e-12	1.62e-08
	IGD-OF	**2.60e-03 (1.34e-04)**	5.42e-02 (2.83e-17)	**2.60e-03 (1.00e-03)**	8.67e-01	9.70e-64	8.66e-43
ZDT6	HV	**4.22e-01 (1.95e-02)**	0.00e+00 (0.00e+00)	2.94e-01 (2.32e-01)	1.01e-02	9.06e-34	1.54e-06
	IGD-CF	**8.50e-04 (1.10e-03)**	1.23e-01 (4.50e-03)	4.81e-02 (5.97e-02)	6.46e-04	4.51e-34	9.71e-07
	NS-CF	**5.00e+01 (0.00e+00)**	0.00e+00 (0.00e+00)	2.68e+00 (3.92e+00)	1.02e-27	0.00e+00	2.30e-03
	IGD-OF	**4.40e-03 (1.80e-03)**	9.86e-02 (5.67e-17)	3.94e-02 (4.54e-02)	7.86e-04	3.40e-43	9.78e-07

TABLE II. RESULTS ON THE 3-OBJECTIVE TEST PROBLEMS. THE MEAN AND STANDARD DEVIATION OF 25 INDEPENDENT RUNS ARE REPORTED. THE STATISTICAL SIGNIFICANCE RESULTS ARE BASED ON THE T-TEST USING A 95% CONFIDENCE INTERVAL.

Func.	Metric	R-MEAD-Te	R-MEAD-Ws	R-NSGA-II	R-MEAD-Te vs R-NSGA-II	R-MEAD-Te vs R-MEAD-Ws	R-MEAD-Ws vs R-NSGA-II
DTLZ1	HV	7.33e-01 (9.98e-01)	0.00e+00 (0.00e+00)	**1.93e+00 (7.18e-02)**	4.00e-06	1.20e-03	5.12e-36
	IGD-CF	4.69e-02 (5.52e-02)	6.46e-02 (4.57e-02)	**1.20e-03 (2.70e-03)**	2.90e-04	2.00e-03	1.63e-07
	NS-CF	3.94e+01 (6.77e+01)	0.00e+00 (0.00e+00)	**1.70e+02 (5.98e+01)**	4.27e-07	7.70e-03	3.28e-13
	IGD-OF	4.12e-02 (2.73e-02)	6.12e-02 (1.42e-17)	**4.20e-03 (1.20e-03)**	6.54e-07	1.20e-03	9.25e-42
DTLZ2	HV	**5.06e-04 (1.01e-04)**	0.00e+00 (0.00e+00)	4.28e-04 (1.07e-05)	7.45e-04	1.06e-18	3.46e-04
	IGD-CF	**1.46e-04 (1.65e-05)**	5.12e-02 (3.60e-04)	8.83e-04 (3.73e-05)	7.97e-32	1.85e-53	2.36e-53
	NS-CF	1.82e+02 (5.34e+00)	0.00e+00 (0.00e+00)	**2.00e+02 (0.00e+00)**	1.04e-14	1.61e-38	0.00e+00
	IGD-OF	**2.10e-03 (4.42e-05)**	5.94e-02 (0.00e+00)	3.00e-03 (3.96e-05)	7.11e-03	1.95e-76	2.01e-77
DTLZ3	HV	2.11e-02 (1.94e-02)	0.00e+00 (0.00e+00)	**3.74e-02 (3.28e-04)**	3.08e-04	1.42e-05	4.44e-51
	IGD-CF	3.22e-02 (3.49e-02)	6.12e-02 (1.00e-02)	**7.64e-04 (8.05e-04)**	1.79e-04	1.25e-05	5.22e-02
	NS-CF	8.94e+01 (9.51e+01)	0.00e+00 (0.00e+00)	**2.00e+02 (0.00e+00)**	5.39e-06	8.82e-05	0.00e+00
	IGD-OF	2.85e-02 (2.80e-02)	5.94e-02 (0.00e+00)	**3.10e-03 (2.48e-04)**	1.46e-04	1.11e-05	2.70e-58
DTLZ4	HV	2.02e-02 (1.57e-02)	0.00e+00 (0.00e+00)	**2.97e-02 (1.80e-03)**	5.70e-03	1.10e-06	4.24e-31
	IGD-CF	2.69e-02 (3.36e-02)	6.09e-02 (9.00e-03)	**8.95e-04 (8.83e-04)**	8.82e-04	1.07e-06	4.32e-21
	NS-CF	8.74e+01 (8.45e+01)	0.00e+00 (0.00e+00)	**2.00e+02 (6.24e-01)**	7.24e-07	2.67e-05	7.43e-62
	IGD-OF	2.42e-02 (2.70e-02)	5.94e-02 (0.00e+00)	**3.00e-03 (1.16e-04)**	6.31e-04	2.90e-66	9.50e-07
DTLZ5	HV	2.90e-03 (4.86e-04)	0.00e+00 (0.00e+00)	**4.00e-03 (3.60e-05)**	4.00e-11	1.86e-02	8.82e-51
	IGD-CF	2.03e-04 (3.31e-05)	5.79e-02 (1.90e-03)	**4.92e-05 (5.45e-06)**	1.22e-18	2.32e-37	2.58e-37
	NS-CF	1.33e+02 (1.96e+01)	0.00e+00 (0.00e+00)	**1.66e+02 (2.33e+00)**	1.25e-08	8.61e-22	3.19e-46
	IGD-OF	2.16e-04 (3.70e-05)	2.76e-02 (7.08e-18)	**3.12e-05 (5.39e-06)**	9.09e-19	1.39e-07	9.75e-91
DTLZ6	HV	1.50e-03 (2.63e-05)	0.00e+00 (0.00e+00)	**2.00e-03 (8.13e-06)**	4.24e-32	1.13e-43	4.49e-59
	IGD-CF	1.77e-04 (2.86e-05)	5.02e-02 (1.35e-04)	**3.92e-05 (4.46e-06)**	8.28e-19	1.63e-63	1.99e-63
	NS-CF	**2.00e+02 (0.00e+00)**	0.00e+00 (0.00e+00)	1.98e+02 (2.14e+00)	9.06e-06	0.00e+00	6.75e-49
	IGD-OF	2.77e-04 (2.19e-05)	2.76e-02 (7.08e-18)	**3.42e-05 (9.85e-06)**	7.20e-26	5.06e-76	1.84e-84

ZDT4 function, the results of all three measures are consistent with IGD-OF. The t-test shows that R-MEAD-Te and R-NSGA-II exhibit similar performance, and both algorithms outperform R-MEAD-Ws. ZDT6 test problem has a concave Pareto-optimal front. Similar to ZDT4, all three measures are consistent with IGD-OF which suggests that R-MEAD-Te performs significantly better than the other algorithms. Figures 3(j), 3(k), and 3(l) also confirm the numerical results.

B. Three-Objective Test Problems

According to the t-test results shown in Table II, all algorithms are statistically distinguishable. Except for NS-CF on DTLZ2 and DTLZ6, all the other measures are consistent with IGD-OF on all functions. It can be seen from Table II that R-NSGA-II outperforms other algorithms on almost all of the functions. It is interesting to note that R-MEAD-Ws fails to converge on all functions. Consequently the HV and NS-CF for R-MEAD-Ws are consistently zero for all functions. Figure 4 shows the performance of all three algorithms on the two selected functions.

C. Result Analysis

According to Tables I and II we can see that the results of IGD-CF are consistent with IGD-OF on all benchmark functions. However, HV is inconsistent on 10 out of 11 functions and NS-CF on 7 out of 11 functions. These results are summarized in Table IV. This may suggest that IGD-CF is the best of the three performance measures reported in this paper. However, it should be noted that for almost all of the benchmarks used in this paper, the solutions which are obtained by the algorithms are fairly close to the Pareto-optimal front. In other words, the composite front is very similar to the Pareto-optimal front in the preferred region. Therefore, on more difficult problems with more complex Pareto-optimal fronts, or on problems with a large number of objectives, it is likely that the composite front is not very accurate, and this may affect the performance of IGD-CF.

It can be seen from Table IV that a cardinality-based approach is less reliable than HV and IGD-CF. The results tend to be worse when dealing with many-objective problems. As the number of objectives increases, a greater portion of the solutions becomes non-dominated, causing different algorithms to make a very close or even identical contribution to the composite front. This makes cardinality-based approaches less accurate as the number of objectives increases and may lead to an incorrect conclusion.

TABLE IV. THE TABLE SHOWS THE CONSISTENCY OF EACH MEASURE WITH IGD-OF (✓: CONSISTENT, ×: INCONSISTENT).

Function	IGD-CF	HV	NS-CF
ZDT1	✓	✓	×
ZDT2	✓	✓	✓
ZDT3	✓	×	×
ZDT4	✓	✓	✓
ZDT6	✓	✓	✓
DTLZ1	✓	✓	✓
DTLZ2	✓	✓	×
DTLZ3	✓	✓	✓
DTLZ4	✓	✓	✓
DTLZ5	✓	✓	✓
DTLZ6	✓	✓	×
Total	11	10	7

VI. CONCLUSION

This paper proposes a metric (UPCF) for measuring the performance of user-preference based EMO algorithms. UPCF works by combining the solution sets of the algorithms that are to be compared and extracting the non-dominated solutions into a *composite front* which is then used to define a preferred region based on the location of a user-supplied reference point. Once the preferred region is defined, existing EMO metrics such as IGD and HV can be used to measure the convergence and diversity of the solution set of each algorithm with respect to the preferred region. For distance-based metrics that require knowledge of the Pareto-optimal front, the *composite front* can be used as a replacement.

To ensure the effectiveness of the proposed metric, three metrics – namely IGD based on the composite front (IGD-CF), hypervolume (HV), and a cardinality-based approach (NS-CF) – were compared with a baseline metric that uses the Pareto-optimal front information (IGD-OF). It is assumed that an algorithm is accurate if its performance is consistent with IGD-OF. The experimental results suggest that both IGD-CF and HV are consistent with IGD-OF, but NS-CF shows some inconsistency. These inconsistencies tend to be magnified when dealing with problems with a higher number of objectives.

Although IGD-CF shows the highest level of consistency with IGD-OF, we speculate that this behavior is due to convergence of the algorithms on the Pareto-optimal front. In other words, the composite front happens to closely resemble a desired portion of the Pareto-optimal front. This might not be the case if the algorithms do not fully converge. This will be the subject of future investigations.

REFERENCES

[1] K. Deb and A. Kumar, "Interactive evolutionary multi-objective optimization and decision-making using reference direction method," in *Proceedings of the 9th annual conference on Genetic and evolutionary computation*, 2007, pp. 781–788.

[2] ——, "Light beam search based multi-objective optimization using evolutionary algorithms," in *In IEEE Congress on Evolutionary Computation (CEC)*, 2007, pp. 2125–2132.

[3] L. Thiele, K. Miettinen, P. J. Korhonen, and J. Molina, "A preference-based evolutionary algorithm for multi-objective optimization," *Evolutionary Computation*, vol. 17, no. 3, pp. 411–436, September 2009.

[4] K. Deb, J. Sundar, N. U. B. Rao, and S. Chaudhuri, "Reference Point Based Multi-Objective Optimization Using Evolutionary Algorithms," *International Journal of Computational Intelligence Research*, vol. 2, no. 3, pp. 273–286, 2006.

[5] A. Mohammadi, M. N. Omidvar, and X. Li, "Reference Point Based Multi-objective Optimization Through Decomposition," in *in Proceedings of Congress of Evolutionary Computation (CEC 2012)*, 2012, pp. 1150–1157.

[6] U. K. Wickramasinghe and X. Li, "Integrating user preferences with particle swarms for multi-objective optimization," in *Proceedings of the 10th annual conference on Genetic and evolutionary computation*. New York, NY, USA: ACM, 2008, pp. 745–752.

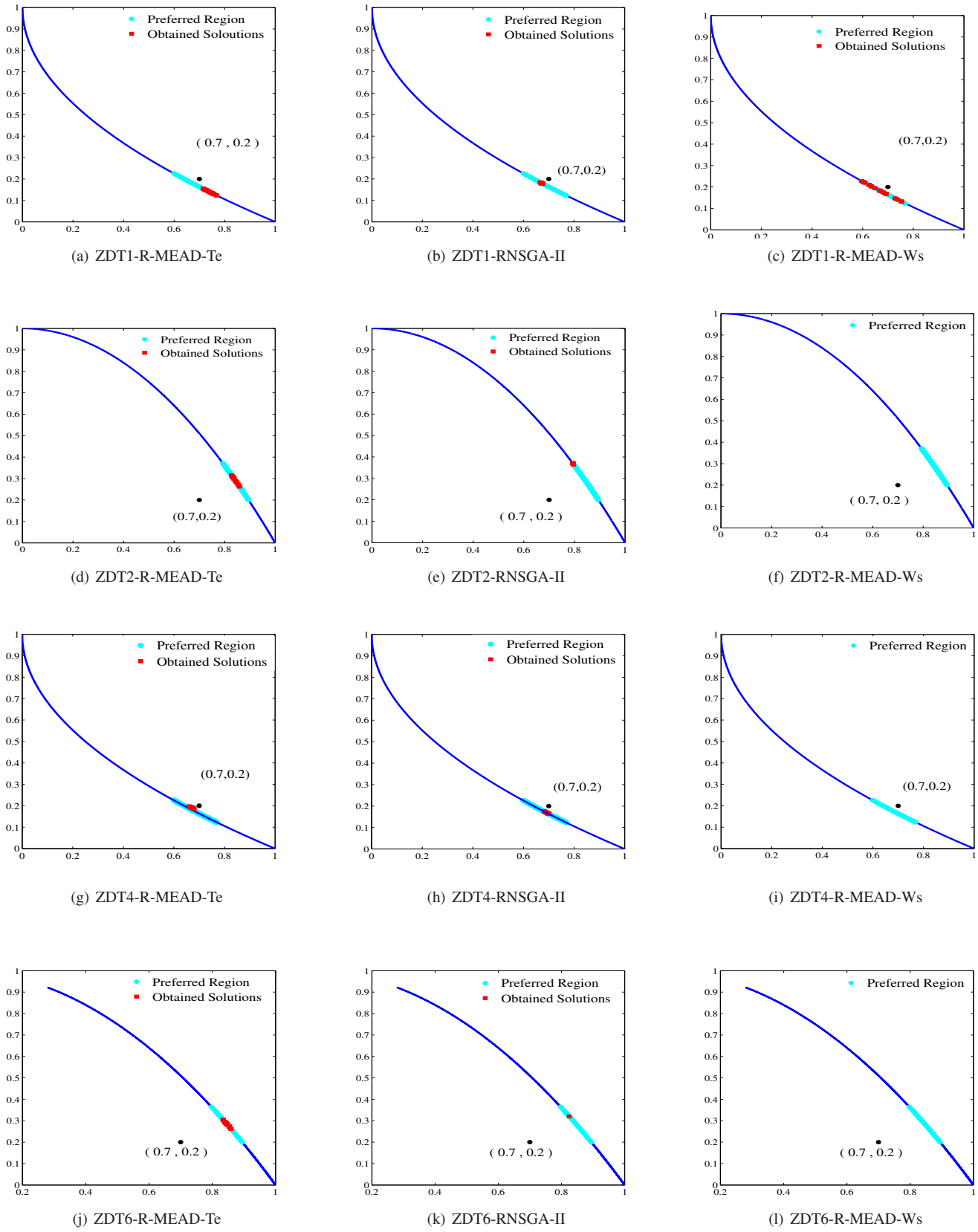
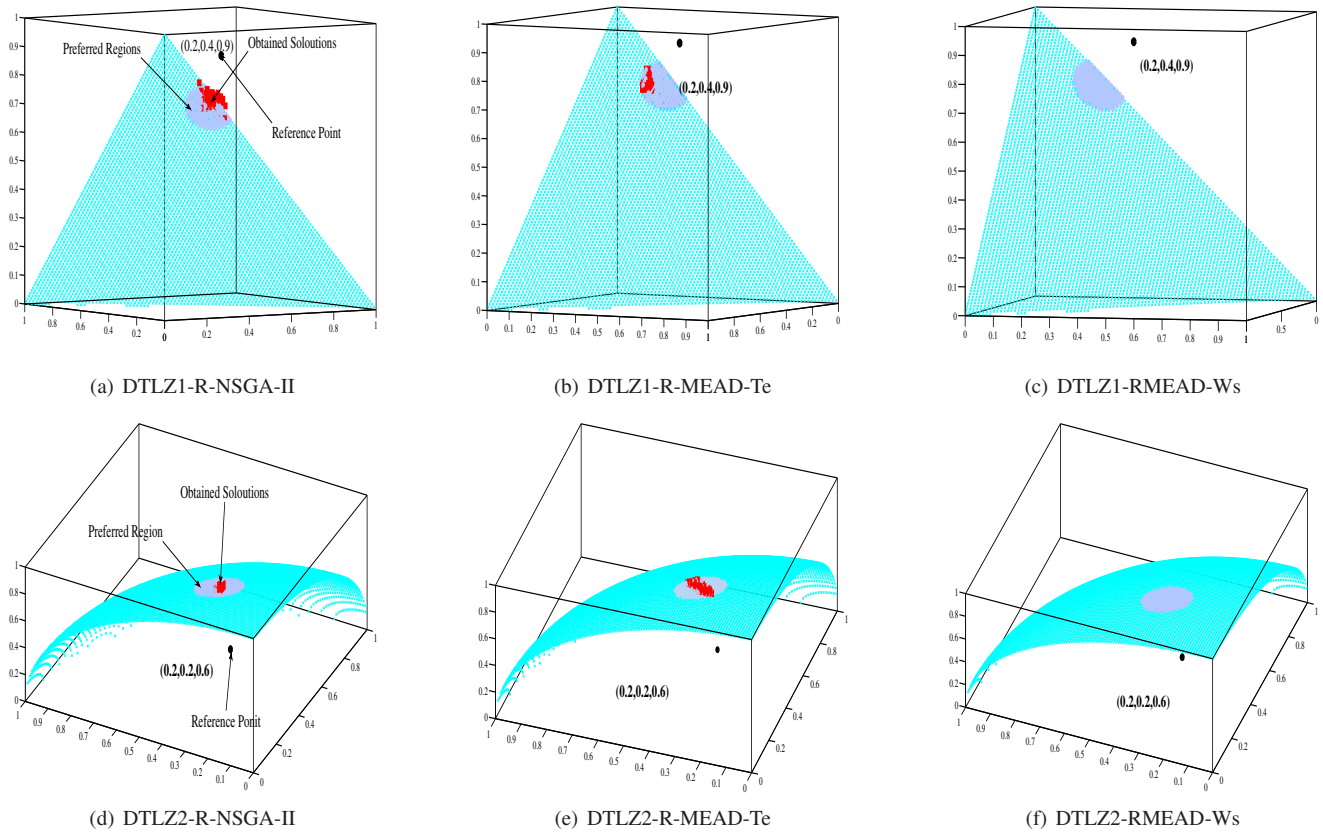


Fig. 3. Results on ZDT1, ZDT2, ZDT4 and ZDT6 functions using R-NSGA-II, R-MEAD-Te and R-MEAD-Ws. Preferred region on Pareto-optimal front is shown in light blue color and solutions found by each algorithm in the region are shown in red.



(a) DTLZ1-R-NSGA-II (b) DTLZ1-R-MEAD-Te (c) DTLZ1-RMEAD-Ws

(d) DTLZ2-R-NSGA-II (e) DTLZ2-R-MEAD-Te (f) DTLZ2-RMEAD-Ws

Fig. 4. Results on DTLZ1 and DTLZ2 functions using R-MEAD-Te, R-MEAD-Ws and R-NSGA-II.

[7] J. Mehnen, H. Trautmann, and A. Tiwari, "Introducing user preference using desirability functions in multi-objective evolutionary optimisation of noisy processes," in *IEEE Congress on Evolutionary Computation, 2007*, 2007, pp. 2687–2694.

[8] U. K. Wickramasinghe, R. Carrese, and X. Li, "Designing Airfoils using a Reference Point based Evolutionary Many-objective Particle Swarm Optimization Algorithm," in *in Proceedings of Congress of Evolutionary Computation (CEC 2010)*. IEEE, 2010, pp. 1857–1864.

[9] D. A. V. Veldhuizen, "Multiobjective evolutionary algorithms: Classifications, analyses, and new innovations," Ph.D. dissertation, School of Engineering of the Air Force Institute of Technology, 1999.

[10] D. A. V. Veldhuizen and G. Lamont, "On measuring multiobjective evolutionary algorithm performance," in *In 2000 Congress on Evolutionary Computation*. IEEE press, 2000, pp. 204–211.

[11] E. Zitzler, L. Thiele, M. Laumanns, C. M. Fonseca, and V. G. da Fonseca, "Performance Assessment of Multiobjective Optimizers: An Analysis and Review," *Evolutionary Computation, IEEE Transactions on*, vol. 7, no. 2, pp. 117–132, 2003.

[12] C. K. Mohan and K. G. Mehrotra, "Reference set metrics for multi-objective algorithms," in *Proceedings of the Second international conference on Swarm, Evolutionary, and Memetic Computing - Volume Part I*, 2011, pp. 723–730.

[13] T. Okabe, Y. Jin, and B. Sendhoff, "A critical survey of performance indices for multi-objective optimisation," in *Evolutionary Computation, 2003. CEC '03. The 2003 Congress on*, vol. 2, December 2003, pp. 878–885.

[14] D. A. van Veldhuizen and G. B. Lamont, "Multiobjective evolutionary algorithm test suites," in *Proceedings of the 1999 ACM symposium on Applied computing*. ACM, 1999, pp. 351–357.

[15] E. Zitzler, "Evolutionary algorithms for multiobjective optimization: Methods and applications," Ph.D.Thesis, Swiss Federal Institute of Technology Zurich, 1999.

[16] P. Czyzak and A. Jaszkiewicz, "Pareto simulated annealing - a metaheuristic technique for multiple-objective combinatorial optimization," *Journal of Multi-Criteria Decision Analysis*, vol. 7, no. 1, pp. 34–47, 1998.

[17] M. P. Hansen and A. Jaszkiewicz, "Evaluating the quality of approximations to the non-dominated set," Institute of Mathematical Modeling, Technical University of Denmark, Tech. Rep. IMM-REP-1998-7, 1998.

[18] E. Zitzler and L. Thiele, "Multiobjective evolutionary algorithms: A comparative case study and the strength pareto approach," *Evolutionary Computation, IEEE Transactions on*, vol. 3, no. 4, pp. 257–271, 1999.

[19] ——, "Multiobjective optimization using evolutionary algorithms – A comparative case study," in *In Parallel Problem Solving from Nature*. PPSN-V, 1998, pp. 292–301.

[20] K. Deb, a. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, 2002.

[21] Q. Zhang and H. Li, "MOEA/D: A multiobjective evolutionary algorithm based on decomposition," *Evolutionary Computation, IEEE Transactions*, pp. 712–731, 2007.

[22] K. Miettinen, *Nonlinear Multiobjective Optimization*. Norwell, MA: Kluwer, 1999.