H. Ishibuchi, H. Masuda, and Y. Nojima, "Sensitivity of performance evaluation results by inverted generational distance to reference points," *Proc. of 2016 IEEE Congress on Evolutionary Computation (Accepted)*.

Sensitivity of Performance Evaluation Results by Inverted Generational Distance to Reference Points

Hisao Ishibuchi, Hiroyuki Masuda, and Yusuke Nojima
Department of Computer Science and Intelligent Systems
Graduate School of Engineering, Osaka Prefecture University
Sakai, Osaka 599-8531, Japan
{hisaoi@, hiroyuki.masuda@ci., nojima@}cs.osakafu-u.ac.jp

***Abstract*—The inverted generational distance (IGD) indicator has been frequently used for performance evaluation of many-objective algorithms. In this paper, we discuss the sensitivity of performance evaluation results by the IGD to the specification of a reference point set. Through computational experiments, we demonstrate that misleading evaluation results can be obtained by the use of the IGD. The reason for the misleading evaluation results is that the IGD tends to favor a solution set with a similar distribution to the reference point set. We demonstrate that such an undesirable bias of the IGD can be decreased by increasing the size of the reference point set and the size of solution sets to be compared. It is also shown that the bias can be decreased by using a modified IGD indicator called the inverted generational distance plus (IGD$^+$). However, the bias becomes more severe by increasing the number of objectives. Our experimental results clearly demonstrate the necessity of very careful examination of performance comparison results by the IGD and IGD$^+$ indicators.**

***Keywords*—Inverted generational distance (IGD), inverted generational distance plus (IGD$^+$), evolutionary many-objective optimization, evolutionary multiobjective optimization.**

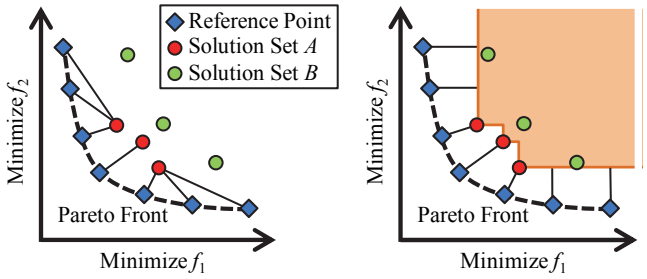
I. Introduction

In the EMO (Evolutionary Multiobjective Optimization) community, the hypervolume (HV) indicator [1] has been used for performance evaluation of EMO algorithms. The use of the HV for comparing solution sets has a strong theoretical support. That is, there is no other Pareto-compliant unary indicator [2]. When a solution set A is better than another solution set B in terms of Pareto dominance (i.e., by the relation "better" in [3]), A always has a larger HV value than B (i.e., $HV(A) > HV(B)$).

One difficulty in the use of the HV indicator is its large computation load especially for many-objective optimization. In the field of evolutionary many-objective optimization [4]-[6], the inverted generational distance (IGD) indicator has been frequently used. To the best of our knowledge, the name of "inverted generational distance" first appeared in 2004 [7], [8]. However, a similar indicator was used in the 1990s by Czyzak & Jaszkiewicz [9] where the distance between solutions was calculated by the weighted achievement scalarizing function. This indicator in [9] was referred to as $D1_R$ in [10] and I_D in [3].

The IGD with the Euclidean distance was also used in 2003 in [11], [12]. The Euclidean distance has been almost always employed in the IGD since 2004 [7], [8]. Currently, the IGD seems to be the most frequently used performance indicator in the field of evolutionary many-objective optimization (e.g., [13]-[16]). However, since the IGD is not Pareto-compliant, misleading or incorrect comparison results can be obtained. That is, a solution set A is not always evaluated as being better than another solution set B even when A is better than B in terms of Pareto dominance.

In a series of our former studies [17]-[19], we discussed the difficulty in the use of the IGD indicator. We also proposed its modified version called the "inverted generational distance plus" (IGD$^+$). More specifically, we demonstrated that different comparison results were obtained from the IGD when we used different reference point sets in [17]. In [18], we explained that the IGD$^+$ is weakly Pareto-compliant while the IGD is not weakly Pareto-compliant. In [19], we demonstrated that Pareto-incompliant comparison results were actually obtained by the IGD through computational experiments on a number of pairs of solution sets A and B where A was better than B in terms of Pareto dominance. Fig. 1 shows an example of such a pair of solution sets A and B for a two-objective minimization problem. In Fig. 1, the solution set A is better than B in terms of Pareto dominance (i.e., each solution in B is dominated by at least one solution in A). However, the better solution set A has a worse (i.e., larger) IGD value than B in Fig. 1 (a). That is, Fig. 1 (a) is an example of Pareto-incompliant comparison of two solution sets by the IGD. The basic idea in the IGD$^+$ is to calculate the distance from each reference point to the dominated region by a solution set as shown in Fig. 1 (b). This idea can be viewed as the incorporation of Pareto dominance between a reference point and a solution in their distance calculation.



(a) IGD calculation for solution set A. (b) IGD$^+$ calculation for solution set A.

Fig. 1. Example of a pair of solution sets A and B where A is better than B in terms of Pareto dominance.

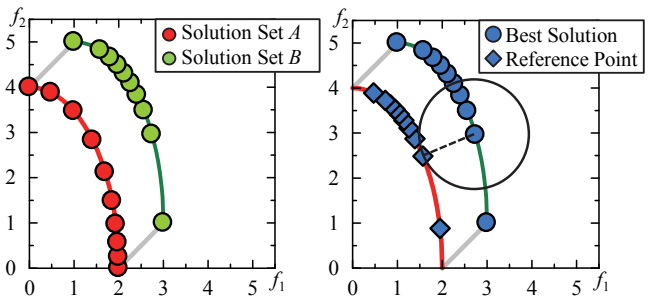
This work was supported by JSPS KAKENHI Grant Numbers 24300090 and 26540128.

The idea of the dominance-based distance calculation in the IGD$^+$ was further extended to the distance calculation between two solution sets by Li et al. [20]. Other ideas to calculate the distance between two solution sets have also been proposed. For example, Schütze et al. [21] combined the IGD with the generational distance (GD [22]) to define a new indicator Δ_p.

In this paper, we examine the sensitivity of performance evaluation results by the IGD and IGD$^+$ indicator to a reference point set. We generate two solution sets A and B such that A can be viewed as clearly being better than B. One solution set A is obtained by an EMO algorithm for minimizing all objectives in a test problem (e.g., WFG [23]) while the other solution set B is obtained by the same EMO algorithm for maximizing all objectives in the same test problem. When they are compared as solution sets of the original minimization problem, A should be evaluated as being better than B.

An example of two solution sets A and B is shown in Fig. 2 (a) where MOEA/D with the weighted Tchebycheff function [24] is applied to the two-objective WFG4 problem with the setting of the population size 10. In Fig. 2 (a), the red curve is the true Pareto front of the original WFG4 problem while the green curve is the Pareto front of its maximization problem variant. As shown in Fig. 2 (a), the solution set A is clearly better than B as a solution set of the minimization problem.

In our computational experiments, we generate reference points using the weighted Tchebycheff function in MOEA/D. First we analytically find the best solution for the weighted Tchebycheff function with each weight vector for the maximization problem. In Fig. 2 (b), ten blue circles on the green curve are obtained in this manner. The blue circles in Fig. 2 (b) look the same as the solution set B in Fig. 2 (a). This is because they are obtained for the same weighted Tchebycheff function with the same set of weight vectors by MOEA/D in (a) and by analytical optimization in (b). Then the nearest point on the red curve to each blue circle is found. In this manner, ten blue squares are obtained as reference points in Fig. 2 (b).



(a) Solution sets A and B. (b) Specification of reference points.

Fig. 2. Solution sets A and B for the two-objective WFG4 problem in (a), and the generated reference point set in (b).

As we can see from Fig. 2, the generated reference points have a similar distribution to the solution set B. However, since all reference points are on the red curve, they may be much closer to the solution set A than B. Intuitively, it is not likely that the solution set B (green circles) is evaluated as being better than A (red circles) as a solution set of the minimization problem. However, the solution set B is often evaluated as being better than A by the IGD as shown later in this paper.

This paper is organized as follows. In Section II, we briefly explain the IGD and IGD$^+$ indicators. In Section III, we explain how to generate two solution sets A and B where A can be viewed as clearly being better than B. In Section IV, we evaluate each solution set using the IGD and IGD$^+$ indicators for the WFG4-9 problems with 3-8 objectives. Whereas A is clearly better than B, B is often evaluated as being better than A by the IGD and IGD$^+$ indicators. Experimental results are discussed in Section V. In Section VI, we conclude this paper.

II. Inverted Generational Distance and Its Variant

A. Inverted Generational Distance: IGD

Let us assume that we have the following minimization problem with m objectives:

$$\text{Minimize } \boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}), ..., f_m(\boldsymbol{x})) \text{ subject to } \boldsymbol{x} \in \mathbf{X}, \quad (1)$$

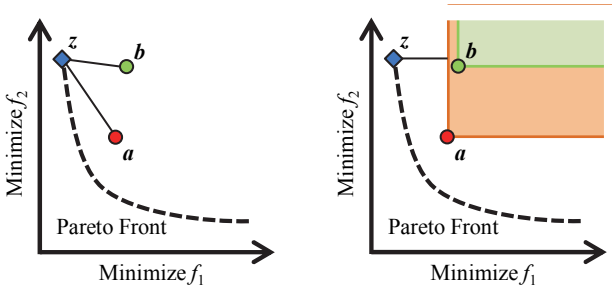
where $\boldsymbol{f}(\boldsymbol{x})$ is an m-dimensional objective vector, $f_i(\boldsymbol{x})$ is the ith objective to be minimized ($i = 1, 2, ..., m$), $\boldsymbol{x}$ is the decision vector, and $\mathbf{X}$ is the feasible region.

The IGD indicator measures the average distance from each reference point to its nearest solution in the objective space. Let us denote a reference point set as $Z = \{\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_{|Z|}\}$ where $\boldsymbol{z}_i$ is a point in the objective space. The reference point set can be viewed as an approximation of the Pareto front. When the Pareto front is known, each reference point is usually sampled from the Pareto front (i.e., $\boldsymbol{z}_i$ is on the Pareto front). Let us denote a solution set A as $A = \{\boldsymbol{a}_1, \boldsymbol{a}_2, ..., \boldsymbol{a}_{|A|}\}$ where $\boldsymbol{a}_j$ is a point in the objective space. The IGD indicator is defined as

$$IGD(A) = \frac{1}{|Z|}\sum_{i=1}^{|Z|}\min_{j=1}^{|A|} d(\boldsymbol{z}_i, \boldsymbol{a}_j), \quad (2)$$

where $d(\boldsymbol{z}_i, \boldsymbol{a}_j)$ is a distance between $\boldsymbol{z}_i$ and $\boldsymbol{a}_j$ in the objective space. In (2), $IGD(A)$ is calculated as the average value of the distance from each reference point $\boldsymbol{z}_i$ in Z to its nearest solution in A (see Fig. 1 (a)). As the distance $d(\boldsymbol{z}_i, \boldsymbol{a}_j)$, the Euclidean distance has been usually used in the field of evolutionary many-objective optimization (e.g., [13]-[16]).

As shown in Fig. 3 (a), even if a solution $\boldsymbol{b}$ is dominated by a solution $\boldsymbol{a}$, the distance from a reference point $\boldsymbol{z}$ to $\boldsymbol{b}$ can be smaller than the distance from $\boldsymbol{z}$ to $\boldsymbol{a}$. This explains why the IGD is not Pareto-compliant. A simple idea for avoiding such an inconsistent distance calculation result is to calculate the distance from the reference point $\boldsymbol{z}$ to the dominated region by $\boldsymbol{a}$ as shown in Fig. 3 (b). This is the basic idea of the IGD$^+$.



(a) Distance calculation in the IGD. (b) Distance calculation in the IGD$^+$.

Fig. 3. Distance calculation in the IGD and the IGD$^+$.

B. *Inverted Generational Distance Plus: IGD$^+$*

The distance calculation for the IGD$^+$ in Fig. 1 (b) and Fig. 3 (b) looks complicated. However, it is very simple even for many-objective problems. In the IGD$^+$, the Euclidean distance from the reference point $\boldsymbol{z} = (z_1, z_2, ..., z_m)$ to the dominated region by the solution $\boldsymbol{a} = (a_1, a_2, ..., a_m)$ is calculated for minimization problems as

$$d^+(\boldsymbol{z}, \boldsymbol{a}) = \sqrt{(\max\{a_1 - z_1, 0\})^2 + \cdots + (\max\{a_m - z_m, 0\})^2} \ . \quad (3)$$

In (3), the ith objective is used for the distance calculation only when its value a_i of the solution $\boldsymbol{a}$ is inferior to z_i of $\boldsymbol{z}$ (e.g., the first objective in Fig. 3 (b) for minimization problems). If its value a_i of the solution $\boldsymbol{a}$ is not inferior to z_i of $\boldsymbol{z}$ (e.g., the second objective in Fig. 3 (b)), the ith objective is not used for the distance calculation from $\boldsymbol{z}$ to $\boldsymbol{a}$ in the IGD$^+$.

Since the IGD$^+$ is based on the dominance relation between the solution $\boldsymbol{a}$ and the reference point $\boldsymbol{z}$ for each objective, a different formulation is needed for maximization problems as

$$d^+(\boldsymbol{z}, \boldsymbol{a}) = \sqrt{(\max\{z_1 - a_1, 0\})^2 + \cdots + (\max\{z_m - a_m, 0\})^2} \ . \quad (4)$$

Except for the change from $(a_i - z_i)^2$ to $(\max\{a_i - z_i, 0\})^2$ in (3) or to $(\max\{z_i - a_i, 0\})^2$ in (4), there is no difference between the IGD and the IGD$^+$. The IGD$^+$ has the same advantage as the IGD: Simplicity of its computation. This is an advantage over the HV indicator. At the same time, the IGD$^+$ has an additional advantage: Weakly Pareto-compliant [18]. As shown in Fig. 3 (b), when the solution $\boldsymbol{b}$ is dominated by the solution $\boldsymbol{a}$, the distance from the reference point $\boldsymbol{z}$ to the dominated region by $\boldsymbol{b}$ is never smaller than that from $\boldsymbol{z}$ to the dominated region by $\boldsymbol{a}$. As a result, we can say that the IGD$^+$ is weakly Pareto-compliant [18]. However, the distance from $\boldsymbol{z}$ to the dominated region by $\boldsymbol{b}$ can be the same as that from $\boldsymbol{z}$ to the dominated region by $\boldsymbol{a}$ (e.g., this is the case in Fig. 3 (b) if $a_1 = b_1$). Thus we cannot say that the IGD$^+$ is Pareto-compliant. That is, even when a solution set A is better than another solution set B in terms of Pareto dominance, A and B can have the same IGD$^+$ value (whereas B cannot have a larger IGD$^+$ value than A). Recently, the IGD$^+$ has been added to the list of quality indicators in jMetal 5.0 [25].

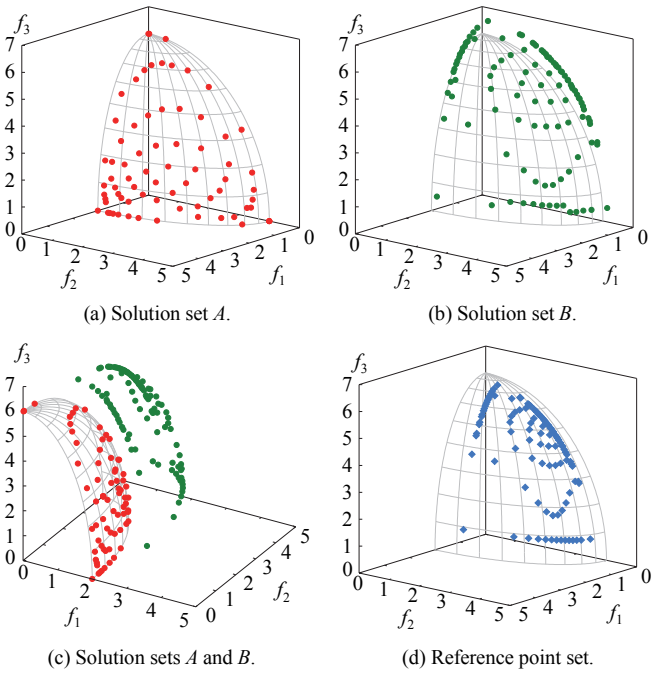
III. Specification of Solution Sets and Reference Points

In our computational experiments, we generate a pair of solution sets A and B in a manner that A can be always viewed as clearly being better than B. One solution set (called solution set A) is obtained by applying an EMO algorithm (called algorithm EMO-A) to a test problem. We use the MOEA/D with the weighted Tchebycheff function [24]. The solution set A is the set of all non-dominated solutions in the final population. As test problems, we use WFG4-9 [23]. It should be noted that all of them are minimization problems. We do not use any normalization of the objective space in this paper.

The other solution set (called solution set B) is obtained by applying the same EMO algorithm to the same test problem for maximizing all objectives. That is, the EMO algorithm is used for maximizing all objectives of the minimization problem. This algorithm is referred to as algorithm EMO-B. The two algorithms EMO-A and EMO-B are the same. However, the EMO-A is used to minimize the objectives of the minimization problem (i.e., its standard use) while the EMO-B is used to maximize the objectives of the same minimization problem (i.e., its meaningless use). The solution set B is the set of all non-dominated solutions with respect to the maximization of all objectives in the final population. Since all objectives are maximized by the EMO-B algorithm, the obtained solution set B can be viewed as a kind of the worst solution set of the original minimization problem (e.g., see Fig. 2 (a)).

As an example, we apply the MOEA/D with the weighted Tchebycheff function to the three-objective WFG4 problem in Fig. 4 where the population size is specified as 105 (i.e., the number of weight vectors is specified as 105). The obtained solution sets A and B are shown in Fig. 4 (a) and Fig. 4 (b), respectively. They are also shown from a different angle in Fig. 4 (c). It is clear from Fig. 4 (c) that the solution set A is better than B for the original minimization problem.

For generating reference points in Fig. 4 (d), we first find the best solution of the weighted Tchebycheff function for each of the 105 weight vectors for the maximization problem. The set of the 105 best solutions is similar to the solution set B in Fig. 4 (b). Then the nearest Pareto optimal solution on the Pareto front of the minimization problem to each of the 105 best solutions is sampled in the objective space as a reference point. Since all reference points in Fig. 4 (d) are on the Pareto front of the minimization problem, they should be close to the solution set A in Fig. 4 (a). However, Fig. 4 (d) has some similarity to the solution set B in Fig. 4 (b). In our computational experiments, we examine whether EMO-A is correctly evaluated as a better algorithm than EMO-B for the original minimization problem by measuring the quality of the solution sets A and B using the following four performance indicators: GD, IGD, IGD$^+$ and HV.



(a) Solution set A. (b) Solution set B.

(c) Solution sets A and B. (d) Reference point set.

Fig. 4. Solution sets A and B for the three-objective WFG4 problem and the generated reference point set (Population size: 105).

IV. Experimental Results

In the first computational experiment, we examine the effect of the population size on the performance evaluation results using the WFG4 problem with three, four and five objectives. Various values from 3 to 3876 are examined as the population size. The increase in the population size means the increase in the number of reference points as well as the increase in the size of solution sets to be compared. In the second computational experiment, we specify the population size as 105, 126 and 120 for the WFG4-9 problems with three, five and eight objectives, respectively. This is because such a small population has often been used in recent studies on evolutionary many-objective optimization. For example, in the NSGA-III paper [13], the population size was specified as 92, 212 and 156 for test problems with three, five and eight objectives, respectively. In the third computational experiment, each weight vector in the weighted Tchebycheff function is changed from $(w_1, w_2, ..., w_m)$ to $(1/w_1, 1/w_2, ..., 1/w_m)$.

Each WFG problem has k position-related variables and l distance-related variables [23]. For m-objective test problems, we specify k and l as $k = 2(m-1)$ and $l = 20$. The neighborhood size is specified using the population size N as $Int(N/10 + 0.9)$ where $Int(x)$ is the integer part of x. Roughly speaking, the neighborhood size is 10% of the population size. We use the SBX (distribution index: 15) with the crossover probability 0.8. We also use the polynomial mutation (distribution index: 20) with the mutation probability $1/(k+l)$ where $(k+l)$ is the number of decision variables (i.e., string length). In this section, we report average results over 20 runs.

A. Computational Experiment 1 on WFG4

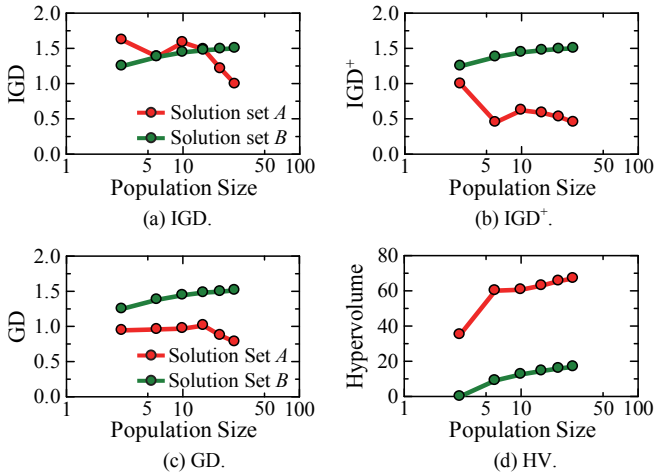
Our first computational experiment is performed on the WFG4 problems under the following settings:

Number of objectives (m): 3, 4, 5,
Termination condition: $m \times 20{,}000$ solution evaluations,
Population size:
Three-objective: 3, 6, 10, 15, 21, 28.
Four-objective: 4, 10, 20, 35, 56, 84, 120, 165, 220, 286.
Five-objective: 5, 15, 35, 70, 126, 210, 330, 495, 715, 1001, 1365, 1820, 2380, 3060, 3876.

For each test problem, we examine all possible values as the population size from its minimum value (i.e., 3 for $m = 3$, 4 for $m = 4$, and 5 for $m = 5$) until the correct comparison result is obtained by the IGD for at least two different specifications of the population size. For the three-objective WFG4 problem, the algorithm EMO-B is incorrectly evaluated as being better than EMO-A by the IGD when the population size is specified as 3, 6, 10 and 15. However, when the population size is specified as 21 and 28, EMO-A is correctly evaluated as being better than EMO-B. Thus we do not examine any larger value than 28 as the population size as shown in Fig. 5 (a). In the same manner, our first computational experiment is also performed for the WFG4 problems with four and five objectives.

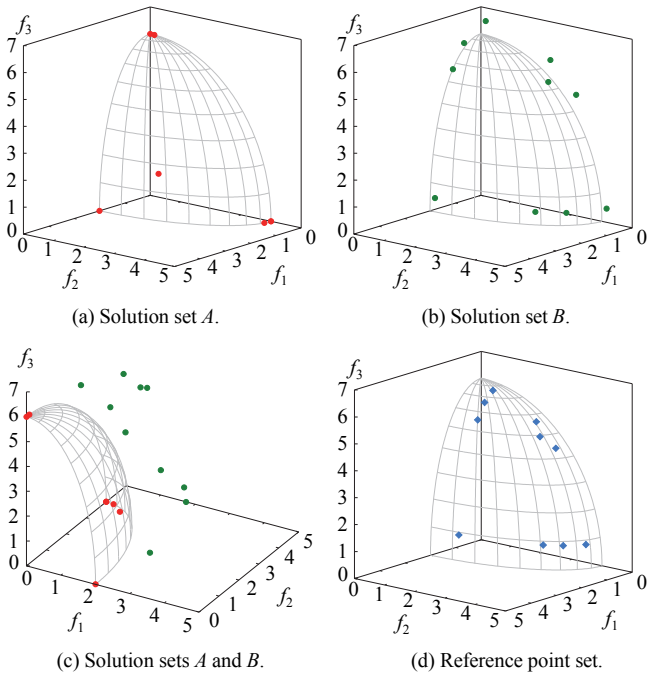
In Fig. 5, we show average results over 20 runs on the three-objective WFG4 problem. The same reference point set is used for the IGD, IGD$^+$ and GD calculations in all runs for each specification of the population size: 3, 6, 10, 15, 21 and 28. For the HV calculation, the reference point is specified as (3, 5, 7). In general, an m-objective problem, the reference point for the HV calculation is specified as $(3, 5, 7, ..., 2m+1)$. It should be noted that larger HV values mean better evaluation results in Fig. 5 whereas smaller IGD, IGD$^+$ and GD values mean better evaluation results.

When the population size is very small (e.g., 10) in Fig. 5 (a), the maximization algorithm EMO-B is evaluated as being better than the minimization algorithm EMO-A for the minimization problem (since the solution set B has the smaller IGD value than A). An example of two solution sets A and B and a reference point set is shown in Fig. 6 for the case of the population size 10. Whereas the solution set B is not close to the Pareto front as shown in Fig. 6 (c), B is evaluated as being better than the solution set A in Fig. 5 (a) by the IGD indicator.



(a) IGD. (b) IGD$^+$. (c) GD. (d) HV.

Fig. 5. Experimental results on the three-objective WFG4 problem.



(a) Solution set A. (b) Solution set B. (c) Solution sets A and B. (d) Reference point set.

Fig. 6. Solution sets A and B for the three-objective WFG4 problem and the generated reference point set from the solution set B (Population size: 10).

When the number of reference points and the number of solutions in each solution set are small, undesirable situations for the IGD indicator (e.g., Fig. 3 (a)) are more likely to happen. Thus the incorrect comparison result is obtained in Fig. 5 (a) when the population size is specified as 3, 6 and 10. However, those specifications are not realistic (i.e., unusually small).

By increasing the number of objectives, solutions and reference points become sparse in the objective space. As a result, undesirable situations in Fig. 3 (a) are more likely to happen in many-objective problems. In the same manner as Fig. 5, we show our experimental results for the WFG4 problems with four and five objectives in Fig. 7 and Fig. 8, respectively. In Fig. 7 (a), the maximization algorithm EMO-B is incorrectly evaluated as being better than EMO-A when the population size is about 100. Such an incorrect comparison result is obtained even when the population size is about 1000 in Fig. 8 (a). These observations show the difficulty of using the IGD indicator for many-objective optimization. When we used the other indicators in Fig. 5, Fig. 7 and Fig. 8, the minimization algorithm EMO-A is correctly evaluated as being better than EMO-B except for some cases where the population size is unusually small. However, unfortunately, this is not always the case as we show in the next two subsections.

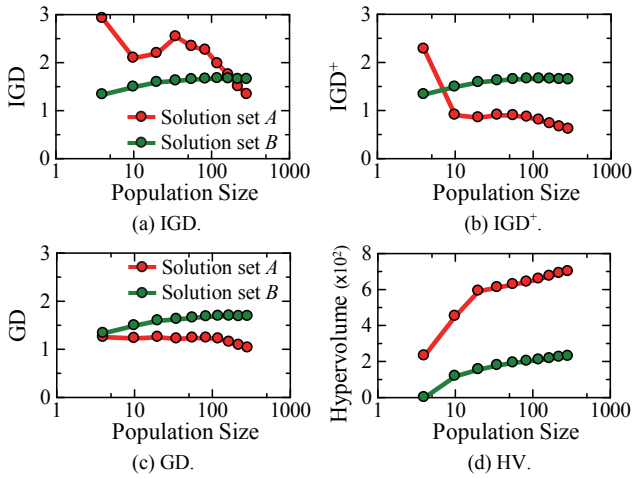


Fig. 7. Experimental results on the four-objective WFG4 problem.

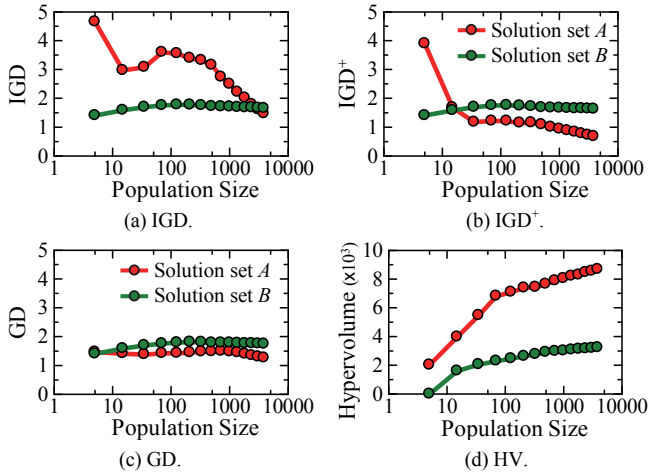


Fig. 8. Experimental results on the five-objective WFG4 problem.

One may wonder why the inconsistency comparison result is obtained by the weakly Pareto-compliant IGD^+ indicator in Fig. 7 (b) and Fig. 8 (b). This is because the Pareto dominance relation does not always hold between the solution sets A and B (whereas all solutions in A are close to the Pareto front and all solutions in B are far from the Pareto front).

B. Computational Experiment 2 on WFG4-9

In our second computational experiment, we examine whether the incorrect comparison result (i.e., the maximization algorithm EMO-B is evaluated as being better than the minimization algorithm EMO-A for minimization problems) is obtained for the WFG4-9 problems with three, five and eight objectives. The following settings are used:

The number of objectives (m): 3, 5, 8,
Termination condition: $m \times 200$ generations,
Population size: 105 ($m = 3$), 126 ($m = 5$), 120 ($m = 8$).

All the other settings such as the crossover and mutation operators are the same as in the first computational experiment. Experimental results are summarized in Tables I-VI. Each entry in each table shows the average indicator value over 20 runs for each test problem. For example, 0.83E7 in the right-bottom corner cell of Table I means that the average HV value of the solution set B for the eight-objective WFG4 problem over 20 runs is 0.83×10^7. The incorrect comparison result is highlighted by the bold red font in each table (larger HV values mean better evaluation whereas smaller IGD, IGD^+ and GD values mean better evaluation).

In Tables I-VI, EMO-B is incorrectly evaluated as being better than EMO-A by the IGD for all test problems with five and eight objectives. When the IGD^+ is used, the incorrect comparison result is not obtained for the five-objective WFG4-9 problems. However, EMO-B is incorrectly evaluated as being better than EMO-A by the IGD^+ indicator for the eight-objective WFG4-9 problems. When GD is used, the incorrect comparison result is obtained for the eight-objective WFG4, WFG7 and WFG8 problems. Except for the eight-objective WFG9, the HV indicator works well. The GD indicator looks better than the IGD^+ in Tables I-VI. However, this totally depends on the situation as shown in the next subsection.

TABLE I. EXPERIMENTAL RESULTS ON THE WFG4 PROBLEM.

Indicator	Three-objective		Five-objective		Eight-objective	
	Set A	Set B	Set A	Set B	Set A	Set B
IGD	0.33	1.52	**3.53**	**1.78**	**6.58**	**1.96**
IGD^+	0.20	1.52	1.16	1.78	**4.01**	**1.96**
GD	0.57	1.54	1.43	1.79	**1.98**	**1.97**
HV	73.2	19.5	7.19E3	2.44E3	1.01E7	0.83E7

TABLE II. EXPERIMENTAL RESULTS ON THE WFG5 PROBLEM.

Indicator	Three-objective		Five-objective		Eight-objective	
	Set A	Set B	Set A	Set B	Set A	Set B
IGD	0.34	1.54	**3.40**	**1.79**	**5.74**	**1.98**
IGD^+	0.24	1.54	1.19	1.79	**3.54**	**1.98**
GD	0.55	1.55	1.28	1.80	1.57	1.99
HV	69.3	19.2	6.82E3	2.40E3	1.11E7	0.79E7

TABLE III. EXPERIMENTAL RESULTS ON THE WFG6 PROBLEM.

Indicator	Three-objective		Five-objective		Eight-objective	
	Set *A*	Set *B*	Set *A*	Set *B*	Set *A*	Set *B*
IGD	0.34	1.52	**3.77**	**1.78**	**6.33**	**1.94**
IGD^+	0.23	1.52	1.35	1.77	**3.68**	**1.94**
GD	0.53	1.54	1.45	1.79	1.84	1.94
HV	70.5	19.5	6.42E3	2.44E3	8.95E6	8.20E6

TABLE IV. EXPERIMENTAL RESULTS ON THE WFG7 PROBLEM.

Indicator	Three-objective		Five-objective		Eight-objective	
	Set *A*	Set *B*	Set *A*	Set *B*	Set *A*	Set *B*
IGD	0.33	1.51	**3.39**	**1.76**	**5.97**	**1.94**
IGD^+	0.19	1.51	1.17	1.76	**2.62**	**1.94**
GD	0.54	1.53	1.49	1.77	**2.12**	**1.95**
HV	73.3	19.7	7.20E3	2.48E3	1.34E7	0.84E7

TABLE V. EXPERIMENTAL RESULTS ON THE WFG8 PROBLEM.

Indicator	Three-objective		Five-objective		Eight-objective	
	Set *A*	Set *B*	Set *A*	Set *B*	Set *A*	Set *B*
IGD	0.36	1.52	**3.80**	**1.77**	**6.52**	**1.92**
IGD^+	0.26	1.52	1.52	1.76	**3.90**	**1.92**
GD	0.59	1.54	1.24	1.78	**1.93**	**1.92**
HV	67.9	19.6	5.19E3	2.46E3	9.10E6	8.26E6

TABLE VI. EXPERIMENTAL RESULTS ON THE WFG9 PROBLEM.

Indicator	Three-objective		Five-objective		Eight-objective	
	Set *A*	Set *B*	Set *A*	Set *B*	Set *A*	Set *B*
IGD	0.35	1.51	**3.47**	**1.81**	**6.09**	**2.22**
IGD^+	0.26	1.51	1.37	1.81	**3.76**	**2.22**
GD	0.55	1.53	1.00	1.85	1.41	2.28
HV	66.8	19.6	5.67E3	2.39E3	**7.30E6**	**7.75E6**

C. *Computational Experiment 3 on WFG4-9*

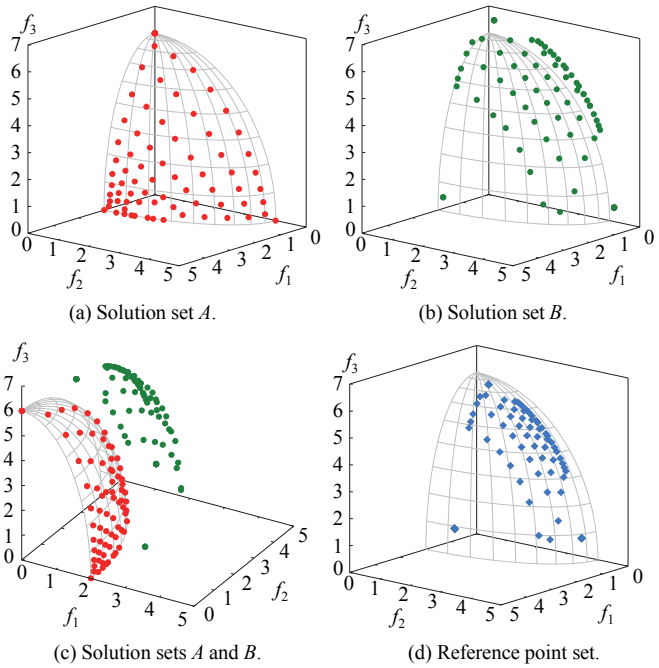
Our experimental results in the previous subsection showed that the maximization algorithm EMO-B was often incorrectly evaluated as being better than EMO-A for the minimization problems by the IGD indicator. One may think that such an incorrect comparison result was obtained since the solution set *A* included less solutions than *B* (e.g., see Fig. 4 and Fig. 6). Actually, the number of obtained non-dominated solutions is often smaller than the population size when the MOEA/D with the weighted Tchebycheff function is used. For example, it was reported in [26] that the number of different non-dominated solutions was about 50-60% of the population size for the many-objective DTLZ2 [27] problems with 4-6 objectives.

One may also think that the incorrect comparison result was obtained by the IGD indicator since the solutions in *A* were not uniformly distributed (see Fig. 2, Fig. 4 and Fig. 6). It has been known that uniformly distributed solutions cannot be obtained by the weighted Tchebycheff function with uniformly distributed weight vectors (e.g., see [28], [29]). The following modification has often been used in the weighted Tchebycheff function for the weight vector $\boldsymbol{w} = (w_1, w_2, ..., w_m)$:

$$\lambda_i = \frac{1/w_i}{\sum_{j=1}^{m} 1/w_j}, \quad i = 1, 2, ..., m. \tag{5}$$

For w_i with $w_i = 0$ in (5), a very small positive real number is used instead of 0. In this paper, we use 10^{-6} as w_i if $w_i = 0$.

Using this modification, we apply the MOEA/D with the weighted Tchebycheff function to the three-objective WFG4 problem. Fig. 9 (a) shows the obtained solution set *A*. Using the same EMO algorithm for maximizing all objectives of the three-objective WFG4 problem, we obtain the solution set *B* in Fig. 9 (b). These two solution sets are shown in Fig. 9 (c) from a different angle. We can see from Fig. 9 (c) that the solution set *A* is clearly better than *B* as a solution set of the original minimization problem. Fig. 9 (d) shows the reference point set obtained from the weighted Tchebycheff function with the modified weight vectors for the maximization variant of WFG4. In Fig. 9 (a), many solutions in the solution set *A* are obtained around the bottom-left corner of the Pareto front. However, there are many reference points in the top-right region of the Pareto front in Fig. 9 (b). This may imply that the IGD, IGD^+ and GD indicators have large values for the solution set *A*. The solutions in *B* in Fig. 9 (b) have some similarity to the reference point set in Fig. 9 (d). However, all solutions in *B* are far from the Pareto front as shown in Fig. 9 (c). So, it is likely that the solution set *A* will be correctly evaluated as being better than *B* by most performance evaluation indicators.



(a) Solution set *A*. (b) Solution set *B*.

(c) Solution sets *A* and *B*. (d) Reference point set.

Fig. 9. Solution sets *A* and *B* for the three-objective WFG4 problem and the generated reference point set from the solution set *B* (Population size: 105).

In the same manner as in the previous subsection, we evaluate the two solution sets *A* and *B* using the four indicators (i.e., IGD, IGD^+, GD and HV) for the WFG4-9 problems with three, five and eight objectives. Average results over 20 runs for each test problem are shown in Tables VII-XII.

TABLE VII. RESULTS WITH EQ. (5) ON THE WFG4 PROBLEM.

Indicator	Three-objective		Five-objective		Eight-objective	
	Set *A*	Set *B*	Set *A*	Set *B*	Set *A*	Set *B*
IGD	0.27	1.45	**1.93**	**1.64**	**6.54**	**2.52**
IGD^+	0.15	1.45	0.68	1.64	1.84	2.47
GD	0.69	1.50	**1.74**	**1.68**	**6.40**	**2.51**
HV	73.1	17.8	7.88E3	1.73E3	2.64E7	0.52E7

TABLE VIII. RESULTS WITH EQ. (5) ON THE WFG5 PROBLEM.

Indicator	Three-objective		Five-objective		Eight-objective	
	Set *A*	Set *B*	Set *A*	Set *B*	Set *A*	Set *B*
IGD	0.26	1.47	**1.85**	**1.68**	**6.97**	**2.52**
IGD^+	0.19	1.47	0.69	1.68	1.87	2.48
GD	0.26	1.47	**1.85**	**1.68**	**6.97**	**2.52**
HV	69.4	17.5	7.68E3	1.69E3	2.47E7	0.50E7

TABLE IX. RESULTS WITH EQ. (5) ON THE WFG6 PROBLEM.

Indicator	Three-objective		Five-objective		Eight-objective	
	Set *A*	Set *B*	Set *A*	Set *B*	Set *A*	Set *B*
IGD	0.26	1.45	**1.73**	**1.64**	**4.62**	**2.50**
IGD^+	0.19	1.45	0.68	1.64	1.81	2.47
GD	0.65	1.52	**1.82**	**1.68**	**6.16**	**2.49**
HV	70.0	17.8	7.86E3	1.74E3	2.81E7	0.50E7

TABLE X. RESULTS WITH EQ. (5) ON THE WFG7 PROBLEM.

Indicator	Three-objective		Five-objective		Eight-objective	
	Set *A*	Set *B*	Set *A*	Set *B*	Set *A*	Set *B*
IGD	0.26	1.44	**1.73**	**1.64**	**4.17**	**2.49**
IGD^+	0.14	1.44	0.61	1.64	1.75	2.44
GD	0.67	1.49	**1.80**	**1.69**	**6.04**	**2.46**
HV	73.4	18.0	8.28E3	1.74E3	3.06E7	0.52E7

TABLE XI. RESULTS WITH EQ. (5) ON THE WFG8 PROBLEM.

Indicator	Three-objective		Five-objective		Eight-objective	
	Set *A*	Set *B*	Set *A*	Set *B*	Set *A*	Set *B*
IGD	0.31	1.45	**2.78**	**1.65**	**8.59**	**2.48**
IGD^+	0.22	1.45	1.31	1.65	**4.04**	**2.43**
GD	0.66	1.48	1.51	1.66	**6.71**	**2.47**
HV	67.6	17.8	4.60E3	1.74E3	1.37E7	0.52E7

TABLE XII. RESULTS WITH EQ. (5) ON THE WFG9 PROBLEM.

Indicator	Three-objective		Five-objective		Eight-objective	
	Set *A*	Set *B*	Set *A*	Set *B*	Set *A*	Set *B*
IGD	0.29	1.45	**2.23**	**1.68**	**8.67**	**2.66**
IGD^+	0.23	1.45	1.09	1.68	2.25	2.65
GD	0.61	1.52	1.04	1.73	**6.22**	**2.65**
HV	65.8	18.2	5.34E3	1.71E3	1.68E7	0.42E7

The incorrect comparison result is highlighted by red bold font in Tables VII-XII. The incorrect comparison result is obtained by the IGD for all WFG4-9 problems with five and eight objectives and by the GD for most of those test problems. The IGD^+ looks a better performance indicator than the IGD in Tables VII-XII (the same observation was also obtained in Tables I-VI). The IGD^+ also looks better than the GD in Tables VII-XII. However, we cannot say that the IGD^+ is always a better performance indicator than the GD (see Tables I-VI in the previous subsection). The comparison between these two measures (i.e., GD and IGD^+) totally depends on the situation. With respect to the comparison between the IGD and the IGD^+, the IGD^+ looks a better indicator than the IGD in all tables. This is due to the incorporation of the dominance relation between a solution and a reference point for each objective into the distance calculation in IGD^+ (i.e., due to the weakly Pareto-compliant property of the IGD^+). In Tables I-XII, the correct comparison result was almost always obtained by the Pareto-compliant performance indicator HV.

V. DISCUSSIONS ON EXPERIMENTAL RESULTS

One may think that reference points should be uniformly sampled from the entire Pareto front. As shown in Fig. 4 (d) and Fig. 9 (d), the reference points in our computational experiments are biased. This may be the main reason for the incorrect comparison result by the distance-based performance indicators (i.e., IGD, IGD^+ and GD) in this paper. However, even when we uniformly sample reference points from the entire Pareto front, the incorrect comparison results can be obtained by these distance-based indicators as shown in Fig. 10 (a similar example was also shown in [30]). In Fig. 10, nine reference points are uniformly sampled as (2, 18), (4, 16), ..., (18, 2). Eight solutions in the solution set *A* are located at the midpoint between each pair of the adjacent reference points as (3, 17), (5. 15), ..., (17, 3). Seven solutions in the solution set *B* are located at (4.5, 16.5), (6.5, 14.5), ..., (16.5, 4.5). From Fig. 10, we can say that *A* is a better solution set than *B* in terms of both the convergence of solutions to the Pareto front and the spread of solutions. However, the solution set *B* is incorrectly evaluated as being better than *A* by the GD and the IGD (while *A* is evaluated as being better than *B* by IGD^+):

$$GD(A) = \sqrt{2} > GD(B) = \sqrt{2}/2,$$
$$IGD(A) = \sqrt{2} > IGD(B) = (7\sqrt{2}/2 + 2\sqrt{8.5})/9 \cong 1.198,$$
$$IGD^+(A) = 1 < IGD^+(B) = (7\sqrt{2}/2 + 2\times 2.5)/9 \cong 1.106.$$

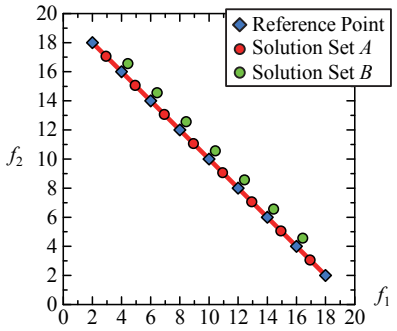


Fig. 10. Two solution sets *A* and *B* and a reference point set *Z* where *A* = {(3, 17), (5, 15), ..., (17, 3)}, *B* = {(4.5, 16.5), (6.5, 14.5), ..., (16.5, 4.5)} and *Z* = {(2, 18), (4, 16), ..., (18, 2)}.

VI. Concluding Remarks

Through computational experiments, we demonstrated that the incorrect comparison result was often obtained by the IGD indicator for many-objective WFG4-9 problems. The incorrect comparison result was more likely to be obtained by increasing the number of objectives and decreasing the size of solution sets. More reliable comparison results were obtained from IGD^+ than IGD in our computational experiments. However, for some test problems, the incorrect comparison result was obtained by all the GD, IGD and IGD^+ indicators. That is, misleading results can be obtained even when we use all of these three indicators. From our experimental results, we can see that performance comparison results by these distance-based indicators should be examined very carefully especially for many-objective problems. The use of a large number of uniformly distributed reference points increases the reliability of the distance-based indicators. However, as shown in Fig. 10, incorrect comparison results can be obtained even when sets of uniformly distributed solutions are compared using uniformly distributed reference points. It may be advisable to use the hypervolume together with the distance-based indicators.

References

[1] E. Zitzler and L. Thiele, "Multiobjective optimization using evolutionary algorithms – A comparative case study," *Lecture Notes in Computer Science*, vol. 1498, pp. 292-301, Springer, Berlin, September 1998.

[2] E. Zitzler, D. Brockhoff, and L. Thiele, "The hypervolume indicator revisited: On the design of Pareto-compliant indicators via weighted integration," *Proc. of 4th International Conference on Evolutionary Multi-Criterion Optimization (Lecture Notes in Computer Science 4403: EMO 2007)*, pp. 862-876, Matsushima, Japan, March 5-8, 2007.

[3] E. Zitzler, L. Thiele, M. Laumanns, C. M. Fonseca, and V. G. da Fonseca, "Performance assessment of multiobjective optimizers: An analysis and review," *IEEE Trans. on Evolutionary Computation*, vol. 7, no. 2, pp. 117-132, April 2003.

[4] H. Ishibuchi, N. Tsukamoto, and Y. Nojima, "Evolutionary many-objective optimization: A short review," *Proc. of 2008 IEEE Congress on Evolutionary Computation*, pp. 2419-2426, Hong Kong, China, June 1-6, 2008.

[5] C. von Lücken, B. Barán, and C. Brizuela, "A survey on multi-objective evolutionary algorithms for many-objective problems," *Computational Optimization and Applications*, vol. 58, no. 3, pp. 707-756, July 2014.

[6] B. Li, J. Li, K. Tang, and X. Yao, "Many-objective evolutionary algorithms: A survey," *ACM Computing Surveys*, vol. 48, no. 1, Article 13, pp. 1-35, September 2015.

[7] C. A. C. Coello and M. R. Sierra, "A study of the parallelization of a coevolutionary multi-objective evolutionary algorithm," *Lecture Notes in Computer Science 2972: MICAI 2004*, pp. 688-697, April 2004.

[8] M. R. Sierra and C. A. C. Coello, "A new multi-objective particle swarm optimizer with improved selection and diversity mechanisms," *Technical Report of CINVESTAV-IPN*, November 23, 2004. (available from http://delta.cs.cinvestav.mx/~ccoello/techreports/evocinv-05-2004.pdf).

[9] 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, January 1998.

[10] J. D. Knowles and D. W. Corne, "On metrics for comparing non-dominated sets," *Proc. of 2002 Congress on Evolutionary Computation*, pp. 711-716, Honolulu, May 12-17, 2002.

[11] P. A. N. Bosman and D. Thierens, "The balance between proximity and diversity in multiobjective evolutionary algorithms," *IEEE Trans. on Evolutionary Computation*, vol. 7, no. 2, pp. 174-188, April 2003.

[12] H. Ishibuchi, T. Yoshida, and T. Murata, "Balance between genetic search and local search in memetic algorithms for multiobjective permutation flowshop scheduling," *IEEE Trans. on Evolutionary Computation*, vol. 7, no. 2, pp. 204-223, April 2003.

[13] K. Deb and H. Jain, "An evolutionary many-objective optimization algorithm using reference-point-based non-dominated sorting approach, Part I: Solving problems with box constraints," *IEEE Trans. on Evolutionary Computation*, vol. 18, no. 4, pp. 577-601, August 2014.

[14] M. Asafuddoula, T. Ray, and R. Sarker, "A decomposition-based evolutionary algorithm for many objective optimization," *IEEE Trans. on Evolutionary Computation*, vol. 19, no. 3, pp. 445-460, June 2015.

[15] K. Li, K. Deb, Q. Zhang, and S. Kwong, "An evolutionary many-objective optimization algorithm based on dominance and decomposition," *IEEE Trans. on Evolutionary Computation*, vol. 19, no. 5, pp. 694-716, October 2015.

[16] Y. Yuan, H. Xu, B. Wang, and X. Yao, "A new dominance relation based evolutionary algorithm for many-objective optimization," *IEEE Trans. on Evolutionary Computation*, vol. 20, no. 1, pp. 16-37, February 2016.

[17] H. Ishibuchi, H. Masuda, Y. Tanigaki, and Y. Nojima, "Difficulties in specifying reference points to calculate the inverted generational distance for many-objective optimization problems," *Proc. of 2014 IEEE Symposium on Computational Intelligence in Multi-Criteria Decision-Making*, pp. 170-177, Orlando, USA, December 9-12, 2014.

[18] H. Ishibuchi, H. Masuda, Y. Tanigaki, and Y. Nojima, "Modified distance calculation in generational distance and inverted generational distance," *Proc. of 8th International Conference on Evolutionary Multi-Criterion Optimization (Lecture Notes in Computer Science 9018: EMO 2015)*, pp. 110-125, Guimarães, Portugal, March 29-April 1, 2015.

[19] H. Ishibuchi, H. Masuda, and Y. Nojima, "A study on performance evaluation ability of a modified inverted generational distance indicator," *Proc. of 2015 Genetic and Evolutionary Computation Conference*, pp. 695-702, Madrid, Spain, July 11-15, 2015.

[20] M. Li, S. Yang, and X. Liu, "A performance comparison indicator for Pareto front approximations in many-objective optimization," *Proc. of 2015 Genetic and Evolutionary Computation Conference*, pp. 703-710, Madrid, Spain, July 11-15, 2015.

[21] O. Schütze, X. Esquivel, A. Lara, and C. A. C. Coello, "Using the averaged Hausdorff distance as a performance measure in evolutionary multiobjective optimization," *IEEE Trans. on Evolutionary Computation*, vol. 16, no. 4, pp. 504-522, August 2012.

[22] D. A. Van Veldhuizen, "Multiobjective evolutionary algorithms: Classifications, analyses, and new innovations," *Ph. D dissertation*, Air Force Institute of Technology, Wright-Patterson AFB, Ohio, May 1999.

[23] S. Huband, P. Hingston, L. Barone, and L. While, "A review of multiobjective test problems and a scalable test problem toolkit," *IEEE Trans. on Evolutionary Computation*, vol. 10, no. 5, pp. 477-506, October 2006.

[24] Q. Zhang and H. Li, "MOEA/D: A multiobjective evolutionary algorithm based on decomposition," *IEEE Trans. on Evolutionary Computation*, vol. 11, no. 6, pp. 712-731, December 2007.

[25] jMetal 5 Web Site http://jmetal.github.io/jMetal/

[26] H. Ishibuchi, K. Doi, H. Masuda, and Y. Nojima, "Relation between weight vectors and solutions in MOEA/D," *Proc. of 2015 IEEE Symposium on Computational Intelligence in Multi-Criteria Decision-Making*, pp. 861-868, Cape Town, December 8-10, 2015.

[27] K. Deb, L. Thiele, M. Laumanns, and E. Zitzler, "Scalable multi-objective optimization test problems," *Proc. of 2002 IEEE Congress on Evolutionary Computation*, pp. 825-830, May 12-17, 2002.

[28] X. Ma, Y. Qi, L. Li, F. Liu, L. Jiao, and J. Wu, "MOEA/D with uniform decomposition measurement for many-objective problems," *Soft Computing*, vol. 18, no. 12, pp. 2541-2564, December 2014.

[29] Y. Qi, X. Ma, F. Liu, L. Jiao, J. Sun, and J. Wu, "MOEA/D with adaptive weight adjustment," *Evolutionary Computation*, vol. 22, no. 2, pp. 231-264, Summer 2014.

[30] H. Ishibuchi, N. Akedo, and Y. Nojima, "Behavior of multi-objective evolutionary algorithms on many-objective knapsack problems," *IEEE Trans. on Evolutionary Computation*, vol. 19, no. 2, pp. 264-283, April 2015.