R-Metric: Evaluating the Performance of Preference-Based Evolutionary Multi-Objective Optimization Using Reference Points*

Ke Li[1], Kalyanmoy Deb[2] and Xin Yao[3,4]

[1]Department of Computer Science, University of Exeter
[2]Department of Electrical and Computer Engineering, Michigan State University
[3]Department of Computer Science and Engineering, Southern University of Science and Technology
[4]CERCIA, School of Computer Science, University of Birmingham
*Email: k.li@exeter.ac.uk, kdeb@egr.msu.edu, xiny@sustc.edu.cn

Abstract: Measuring the performance of an algorithm for solving multi-objective optimization problem has always been challenging simply due to two conflicting goals, i.e., convergence and diversity of obtained trade-off solutions. There are a number of metrics for evaluating the performance of a multi-objective optimizer that approximates the whole Pareto-optimal front. However, for evaluating the quality of a preferred subset of the whole front, the existing metrics are inadequate. In this paper, we suggest a systematic way to adapt the existing metrics to quantitatively evaluate the performance of a preference-based evolutionary multi-objective optimization algorithm using reference points. The basic idea is to pre-process the preferred solution set according to a multi-criterion decision making approach before using a regular metric for performance assessment. Extensive experiments on several artificial scenarios and benchmark problems fully demonstrate its effectiveness in evaluating the quality of different preferred solution sets with regard to various reference points supplied by a decision maker.

Keywords: User-preference, performance assessment, reference point, multi-criterion decision making, evolutionary multi-objective optimization.

1 Introduction

Most real-world problem solving tasks usually involve multiple incommensurable and conflicting objectives which need to be considered simultaneously. Such problems are termed as multi-objective optimization problems (MOPs) that have earned considerable attention in engineering design, modeling, and operations research. Instead of a single solution that optimizes all objectives simultaneously, in multi-objective optimization, we often look for a set of Pareto-optimal solutions none of which can be considered better than another when all objectives are of importance.

Over the past two decades and beyond, evolutionary algorithms (EAs) have been widely accepted as a major approach for multi-objective optimization. Many efforts have been devoted to developing evolutionary multi-objective optimization (EMO) algorithms, such as elitist non-dominated sorting genetic algorithm (NSGA-II) [1–4], indicator-based EA (IBEA) [5–7] and multi-objective EA based on decomposition (MOEA/D) [8–11]. These algorithms, without any additional preference information (or intervention) from a decision maker (DM), are usually designed to obtain a set of solutions that approximate the whole Pareto-optimal set. However, the ultimate goal of multi-objective optimization is to help the DM find solutions that meet his/her own preference information. To facilitate the decision making process, it is desirable to integrate the DM's preference information into the search process of EMO for the following reasons:

1. Supplying a DM with a large amount of trade-off points not only increases his/her workload, but also provides many irrelevant or even noisy information to the decision making process. Rather than the whole Pareto-optimal front (PF), the DM usually interests in only a small set

of trade-off points most relevant to him/her. A biased search, according to the DM's preference information, is able to provide more acceptable solutions.

2. Due to the curse of dimensionality, the number of points used to accurately represent the whole PF increases exponentially with the number of objectives. This not only severely increases the computational burden of an EMO algorithm, but also causes extra difficulties for the DM to comprehend the obtained solutions and then to make decisions. Therefore, it is more practical to search for a fine-grained resolution of a preferred region of the PF by incorporating the DM's preference information.

3. In a high-dimensional space, the mushrooming of non-dominated solutions, even for a randomly generated population, renders the traditional Pareto dominance based selection useless [12]. However, by considering the DM's preference information, we can expect a necessary selection pressure additional to Pareto dominance [13].

In the past decade, there have been a number of studies on the preference-based EMO. Generally speaking, their ideas can be divided into four categories. The first one modifies the original Pareto dominance by classifying objectives into different levels and priorities (e.g., [14–16]) or expresses the DM's preference information by fuzzy linguistic terms according to different aspiration levels (e.g., [17–19]). The second sort modifies the diversity management module so that the density of Pareto-optimal solutions can be biased towards the region of interest (ROI) (e.g., [20–22]). The third approach combines the classical reference point based method [23] with EMO (e.g., [24–26]). The last category, as a recent trend, combines the DM's preference information with performance metrics (e.g., weight hypervolume [27], R2-indicator [28] and averaged Hausdorff distance [29]) in algorithm design. In this paper, our discussion focuses on the reference point based method, which has been recognized as one of most popular methods in this literature [30].

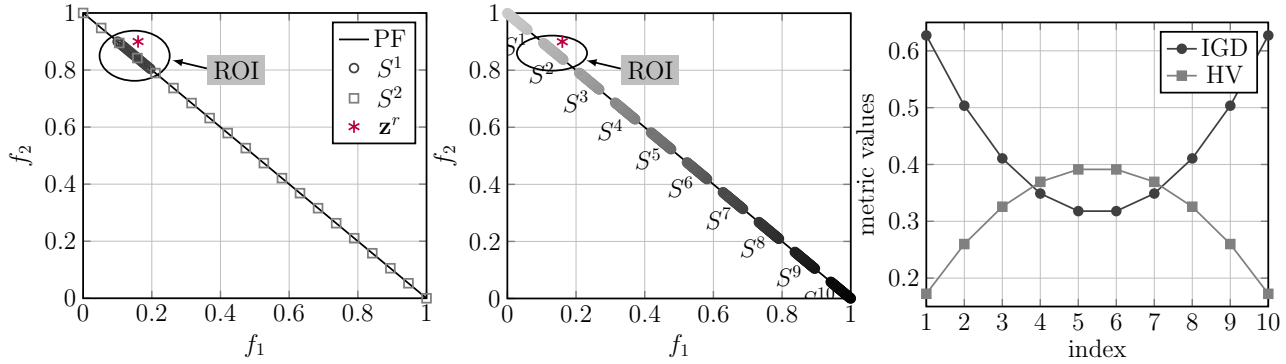
Despite the progress in algorithm design, few have been done on evaluating the quality of preferred solutions obtained by a preference-based EMO algorithm. Although a number of performance metrics have been suggested to evaluate the quality of solutions that approximate the whole PF, including metrics for evaluating convergence (e.g., [31–33]) and diversity (e.g., [34–36]) separately, and metrics that evaluate both aspects simultaneously (e.g., [37–39]), none of them can be directly applicable when only a partial PF is considered. Some attempts to adapt the regular metrics to serve the purpose of assessing the quality of a preferred solution set have been reported in [40] and [41]. However, they are ad-hoc and oversimplified which could make the assessments misleading. Possibly due to the lack of reliable metrics, many studies, if not all, on the preference-based EMO heavily rely on the visual plot in performance comparisons. These methods are rather subjective, and how to visualize data in a high-dimensional space is itself an open problem. This paper presents a systematic way, denoted as R-metric, to quantitatively evaluate the quality of preferred solutions obtained by a preference-based EMO algorithm using reference points. Our basic idea is to use a multi-criterion decision making (MCDM) approach to pre-process the obtained solutions, according to their satisfaction to the DM's preference information, before using a regular metric for performance assessment. The proposed method is so simple and general that any existing metric can be adapted with little modification.

The rest of this paper is organized as follows. Section 2 gives some preliminary concepts related to this paper. In Section 3, the motivations of this paper are delineated and discussed. Section 4 is devoted to the description of the proposed method. Section 5 and Section 6 present the empirical studies on several artificial scenarios and a series of benchmark problems respectively. Finally, Section 7 concludes this paper and provides some future directions.

2 Preliminary Concepts

This paper considers the following continuous MOP with box constraints:

$$\begin{array}{ll} \text{minimize} & \mathbf{F}(\mathbf{x}) = (f_1(\mathbf{x}), \cdots, f_m(\mathbf{x}))^T \\ \text{subject to} & \mathbf{x} \in \Omega \end{array} \tag{1}$$



(a) Partial region vs whole region. (b) Ten point sets along the whole EF. (c) IGD and HV values for point sets in Fig. 1(b).

Figure 1: Variations of IGD and HV values with respect to a DM specified reference point $\mathbf{z}^r = (0.16, 0.9)$.

where $\Omega = \prod_{i=1}^{n}[a_i, b_i] \subset \mathbb{R}^n$ is the *decision* (*variable*) *space*, $\mathbf{x} = (x_1, \ldots, x_n)^T \in \Omega$ is a candidate solution. $\mathbf{F} : \Omega \to \mathbb{R}^m_+$ constitutes of m real-valued objective functions, and $\mathbb{R}^m_+$ is called the *objective space*. The *attainable objective set* is defined as $\Theta = \{\mathbf{F}(\mathbf{x})|\mathbf{x} \in \Omega\}$.

Definition 1. *$\mathbf{x}^1$ is said to Pareto dominate $\mathbf{x}^2$, denoted as $\mathbf{x}^1 \preceq \mathbf{x}^2$, if and only if: $\forall i \in \{1, \cdots, m\}$, $f_i(\mathbf{x}^1) \leq f_i(\mathbf{x}^2)$; and $\exists j \in \{1, \cdots, m\}$, $f_j(\mathbf{x}^1) < f_j(\mathbf{x}^2)$.*

Definition 2. *$\mathbf{x}^* \in \Omega$ is said to be Pareto-optimal if there is no other $\mathbf{x} \in \Omega$ such that $\mathbf{x} \preceq \mathbf{x}^*$.*

Definition 3. *The set of all Pareto-optimal solutions is called the Pareto-optimal set (PS). The set of all Pareto-optimal objective vectors, $PF = \{\mathbf{F}(\mathbf{x})|\mathbf{x} \in PS\}$, is called the PF.*

For the ease of later discussion, we briefly introduce two widely used performance metrics in the EMO literature.

1. *Inverted Generational Distance* (IGD) metric [37]: Let P^* be a set of points uniformly sampled along the PF, and S be the set of solutions obtained by an EMO algorithm. The IGD value of S is calculated as:
$$\text{IGD}(S, P^*) = \frac{\sum_{\mathbf{x}^* \in P^*} dist(\mathbf{x}^*, S)}{|P^*|} \tag{2}$$
where $dist(\mathbf{x}^*, S)$ is the Euclidean distance between the point $\mathbf{x}^* \in P^*$ and its nearest neighbor of S in the objective space, and $|P^*|$ is the cardinality of P^*.

2. *Hypervolume* (HV) metric [33]: Let $\mathbf{z}^w = (z_1^w, \ldots, z_m^w)^T$ be a worst point in the objective space that is dominated by all Pareto-optimal objective vectors. HV metric measures the size of the objective space dominated by solutions in S and bounded by $\mathbf{z}^w$.
$$\text{HV}(S) = \text{VOL}(\bigcup_{\mathbf{x} \in S}[f_1(\mathbf{x}), z_1^w] \times \ldots [f_m(\mathbf{x}), z_m^w]) \tag{3}$$
where $\text{VOL}(\cdot)$ indicates the Lebesgue measure.

Both IGD and HV metrics are able to give a comprehensive information, including the convergence and diversity, of S simultaneously. The lower is the IGD value (or the larger is the HV value), the better is the quality of S for approximating the whole PF.

3 Motivation

This section first discusses the shortcomings of some existing metrics for evaluating the partial PF. Then, we develop the motivation for our proposed R-metric from the perspective of the achievement scalarization function (ASF), i.e., an MCDM-based scalarization approach.

3.1 Shortcomings of Regular Metrics

Let us use two toy examples to illustrate some shortcomings of IGD and HV metrics for assessing the partial PF. In particular, the example PF is a line (i.e., $f_2 = 1 - f_1$) having an intercept of one with each objective axis. The DM's preference information is specified as a reference point $\mathbf{z}^r = (0.16, 0.9)^T$ in the objective space. Points focusing on the region closest to $\mathbf{z}^r$ are most relevant to the DM's preference information. To calculate the IGD values, we sample 670 evenly distributed points along the PF; and we set the worst point as $\mathbf{z}^w = (1.1, 1.1)^T$ for calculating the HV value.

1. In Fig. 1(a), two sets of points S^1 and S^2 have the same cardinality ($|S^1| = |S^2| = 20$), but are with different spreads along the PF. S^1 crowds around $\mathbf{z}^r$, while S^2 evenly distributes along the whole PF. From the DM's perspective, S^1 is preferable than S^2. However, since S^2 has a wider spread over the PF, it obviously has better IGD and HV values than S^1. Specifically, IGD(S^1) = 3.476E-1 and IGD(S^2) = 4.610E-4; HV(S^1) = 0.2910 and HV(S^2) = 0.6837.

2. In Fig. 1(b), ten sets of points S^1 to S^{10} are created along the PF. Each set contains 40 evenly distributed points and has the same spread. Fig. 1(c) shows the IGD and HV values obtained by each point set. Since S^2 locates in the ROI, it was supposed to have the best metric values. However, as shown in Fig. 1(c), S^2 obtains the second worst metric values, whereas S^5 and S^6, far away from the ROI, obtain the best metric values.

In summary, neither IGD nor HV metric is reliable for evaluating the quality of a preferred solution set. A solution set with additional but unwanted points may obtain a better metric value, thereby making the IGD and HV metrics unsuitable for performance assessment in the toy example shown in Fig. 1(a). On the other hand, even for different point sets having the same spread along the PF, their IGD and HV values depend on their positions and the PF's geometric property. This makes the IGD and HV metrics unsuitable for performance assessment in the toy example shown in Fig. 1(b).

3.2 Shortcomings of Existing Preference-Based Metrics

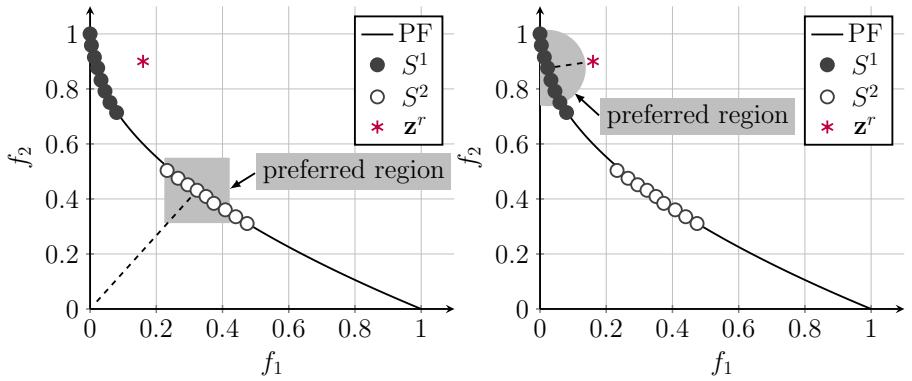
To the best of our knowledge, there are two previous attempts, i.e., [40] and [41], to adapt the regular HV metric for the preference-based EMO. Their basic ideas are similar. At first, they merge solutions obtained by all considered algorithms into a composite set. Then, they specify a preferred region within the composite set. Finally, only solutions falling within this preferred region are considered for performance assessment. The major difference between [40] and [41] is the setting of the preferred region. As shown in Fig. 2(a), [40] uses the closest point to the origin as the center of the preferred region. In contrast, as shown in Fig. 2(b), [41] uses the closest point to the DM supplied reference point as the center. Both these two metrics do not require any prior knowledge of the PF, and they work for some simple examples. However, they have some flaws that make them misleading:

1. It is obvious that [40] does not take the DM's preference information into consideration. For the example in Fig. 2(a), S^1 is obviously preferable than S^2 considering the given reference point $\mathbf{z}^r$. However, S^1 and S^2 are distant from each other, and the origin is closer to the points in S^2. Therefore, S^1 will be wrongly excluded from the preferred region for performance assessment.

2. On the other hand, although [41] considers the DM's preference information in computation, it treats points outside the preferred region equally redundant, e.g., in Fig. 2(b), no point in S^2 will be considered in performance assessment. Considering the example in Fig. 1(b), all ten point sets, except S^2, cannot get any meaningful metric value. This gives the DM a wrong information that S^1 to S^{10}, except S^2, are equally bad.

3.3 Intuitions of MCDM Approach

In the MCDM literature, there exists a number of methods for finding a preferred solution set according to the DM supplied reference point. In this subsection, we describe the basic idea of the ASF



(a) Mechanism of [40]. (b) Mechanism of [41].

Figure 2: Illustration of two preference-based metrics.

method [42], which is the foundation of our proposed R-metric, in detail. In particular, The ASF[1] considered in this paper is formulated as follows:

$$\begin{aligned} &\text{minimize} \quad \text{ASF}(\mathbf{x}|\mathbf{z}^r, \mathbf{w}) = \max_{1\leq i\leq m} \frac{f_i(\mathbf{x})-z_i^r}{w_i} \\ &\text{subject to} \quad \mathbf{x} \in \Omega. \end{aligned} \tag{4}$$

where $\mathbf{z}^r$ is the reference point that represents the DM's aspiration level for each objective, and $\mathbf{w}$ is the weight vector that implies the relative importance of objectives. Based on the ASF, each objective vector has a projection, called iso-ASF point, on the reference line originated from $\mathbf{z}^r$ and along $\mathbf{w}$, as shown in Fig. 3. Specifically, for a point $\mathbf{a} \in \Theta$, its corresponding iso-ASF point $\mathbf{a}^l$ is calculated as:

$$\mathbf{a}^l = \mathbf{z}^r + \delta\mathbf{w} \tag{5}$$

where $\delta = \max_{1\leq i\leq m} \frac{a_i-z_i^r}{w_i}$. This iso-ASF point gives us an information about the *closeness* of $\mathbf{a}$ to $\mathbf{z}^r$ along the preferred direction $\mathbf{w}$. Note that not all Pareto-optimal solutions are equally important when considering the DM's preference information. Based on the supplied reference point and a preferred direction, ASF is able to rank all Pareto-optimal solutions. As shown in Fig. 3, for $\mathbf{a}$, any point on its ASF contour line (e.g., point $\mathbf{b}$) has the same ASF value, i.e., they are equally good and have the identical rank. For another point $\mathbf{c}$, its iso-ASF point is $\mathbf{c}^l$. Comparing to $\mathbf{a}^l$, $\mathbf{c}^l$ is closer to $\mathbf{z}^r$ along the preferred direction. Thus, $\mathbf{c}$ should have a better rank than $\mathbf{a}$ and $\mathbf{b}$. Another nice property of this ASF-based ranking concept is promising scalability to many-objective problems.

4 R-metric Calculation Principle

The basic idea of our proposed method, denoted as R-metric, is to use an MCDM approach to pre-process the preferred solution set according to the DM supplied preference information. Thereafter, regular metrics, e.g., IGD and HV, can be applied for performance assessment. Note that the R-metric is specifically designed for evaluating the performance of a preference-based EMO algorithm using one or more reference points. In particular, we assume that the DM prefers the solutions lying toward the preferred direction, represented as a direct objective-wise weighting information or a worst point. In the R-metric calculation, the DM is required to provide three parameters relating to his/her preference information: *i*) a reference point $\mathbf{z}^r$ that represents his/her aspiration level or desired value for each objective, *ii*) a worst point $\mathbf{z}^w$ or a weight vector $\mathbf{w}$ that specifies the relative importance of each objective, and *iii*) a relative extent of the ROI, denoted as Δ ($0 < \Delta \leq 1$). Note that most of these parameters are used to elicit the DM's preference information and to help the preference-based optimization procedure find a set of trade-off solutions in the ROI. Our proposed R-metric calculation

[1]Here we use the classic weighted Chebyshev function for discussion. Without loss of generality, other ASF forms can also be adapted accordingly.

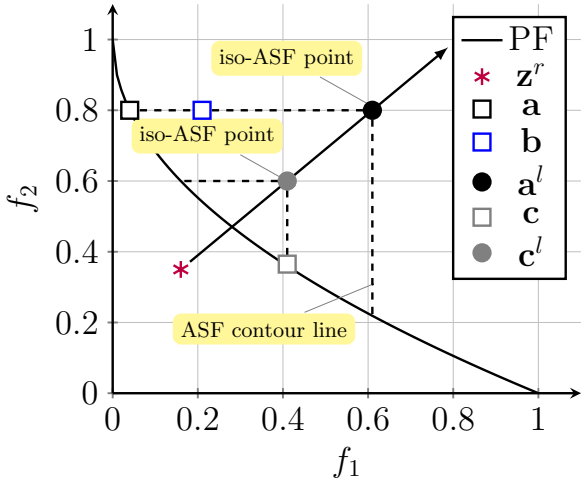


Figure 3: Illustration of the ASF ranking method. Considering the DM supplied reference point $\mathbf{z}^r$, $\mathbf{c}$ is preferable than $\mathbf{a}$ and $\mathbf{b}$; while $\mathbf{a}$ and $\mathbf{b}$ are equally important.

is simple in principle and its high level flowchart is given in Fig. 4. In the following paragraphs, we first describe each step in detail. Then, we provide some further comments followed by a time complexity analysis.

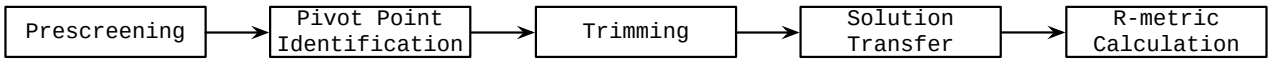


Figure 4: The flowchart of R-metric calculation.

4.1 Descriptions of Each Step

4.1.1 Prescreening Procedure

In multi-objective optimization, only the non-dominated solutions are of interest to the DMs and are meaningful for performance assessment. Assume that there are L ($L \geq 1$) preferred solution sets (denoted as $S^1, \cdots, S^L$), obtained by L different preference-based EMO algorithms, at hand. We at first merge these L preferred solution sets into a composite set S^c. For each S^i, $i \in \{1, \cdots, L\}$, only the non-dominated solutions, comparing to those in S^c, are retained for the R-metric calculation. The pseudo-code of this prescreening procedure is given in Algorithm 1.

Algorithm 1: Prescreening Procedure

Input: Preferred solution sets $S^1, \cdots, S^L$
Output: Processed $S^1, \cdots, S^L$

for $i \leftarrow 1$ **to** L **do**
  $S^c \leftarrow \bigcup_{l=1}^{L} S^l \backslash S^i$;
  for $j \leftarrow 1$ **to** $|S^i|$ **do**
    for $k \leftarrow 1$ **to** $|S^c|$ **do**
      if $S^c(k) \preceq S^i(j)$ **then**
        $S^i = S^i \backslash \{S^i(j)\}$;
        break;
return $S^1, \cdots, S^L$

4.1.2 Pivot Point Identification

As the name suggests, the pivot point (denoted as $\mathbf{z}^p$) of a given set of preferred solutions (denoted as S) is used as the representative that reflects the overall satisfaction of S with respect to the DM supplied preference information. In this paper, we use the best solution with respect to (4) to serve this purpose and thus $\mathbf{z}^p$ is:

$$\mathbf{z}^p = \underset{\mathbf{x} \in S}{\operatorname{argmin}} \operatorname{ASF}(\mathbf{x}|\mathbf{z}^r, \mathbf{w}) \tag{6}$$

4.1.3 Trimming Procedure

Instead of the whole PF, the ROI is a bounded region, i.e., a part of the PF, given the DM's preference information. Only solutions located in the ROI are of interest to the DM. In this paper, we define the ROI approximated by S as the cubic centered at the pivot point and is with a side length Δ. Only solutions located in this approximated ROI are valid for performance assessment. The pseudo-code of this trimming procedure is given in Algorithm 2.

Algorithm 2: Trimming Procedure

Input: Preferred solution set S, ROI's relative extent Δ
Output: Processed S

```
for i ← 1 to |S| do
    for j ← 1 to m do
        if |f_j(x^i) − z_j^p| > Δ/2 then
            S ← S\{x^i};
            break;
return S
```

4.1.4 Solution Transfer

This step is the main crux of our R-metric by which the trimmed points are transferred to a virtual position. Then, we can assess their *closeness* to the ROI along the preferred direction. To this end, we first compute the iso-ASF point of $\mathbf{z}^p$ (denoted as $\mathbf{z}^l$) on the reference line connecting $\mathbf{z}^r$ and $\mathbf{z}^w$. According to equation (5), this requires to identify the objective k that contributes to the ASF value:

$$k = \underset{1 \leq i \leq m}{\operatorname{argmax}} \frac{z_i^p - z_i^r}{z_i^w - z_i^r} \tag{7}$$

Then, we can compute $\mathbf{z}^l$ as:

$$z_i^l = z_i^r + \frac{z_k^p - z_k^r}{z_k^w - z_k^r}(z_i^w - z_i^r) \tag{8}$$

where $k \in \{1, \cdots, m\}$. Thereafter, all trimmed points are transferred, along the direction vector $\mathbf{z}^l - \mathbf{z}^p$ with the distance $\|\mathbf{z}^l - \mathbf{z}^p\|$, to a virtual position. The pseudo-code of this solution transfer procedure is given in Algorithm 3.

4.1.5 R-metric Calculation

In this paper, we choose the IGD and HV as the baseline metrics to evaluate the quality of a preferred solution set. The resulting R-metric is thus denoted as R-IGD or R-HV depending on the chosen baseline. For the R-HV, we simply compute the hypervolume of the transferred points with respect to the worst point $\mathbf{z}^w$. For the R-IGD, we need to pre-process P^* beforehand. More specifically, we first use the method developed in Section 4.1.2 to identify the pivot point of P^*. Then, we use the trimming procedure suggested in Section 4.1.3 to trim the points outside the ROI along the PF. In the end, the remaining points form the new P^* for the R-IGD calculation.

Algorithm 3: Solution Transfer

Input: Preferred solution set S

Output: Processed S

$k \leftarrow \underset{1 \le i \le m}{\text{argmax}}(\frac{z_i^p - z_i^r}{z_i^w - z_i^r})$;

for $i \leftarrow 1$ **to** m **do**

  $z_i^l \leftarrow z_i^r + \frac{z_k^p - z_k^r}{z_k^w - z_k^r}(z_i^w - z_i^r)$;// iso-ASF point

for $i \leftarrow 1$ **to** $|S|$ **do**

  Shift $S(i)$ along the direction vector $\mathbf{z}^l - \mathbf{z}^p$;

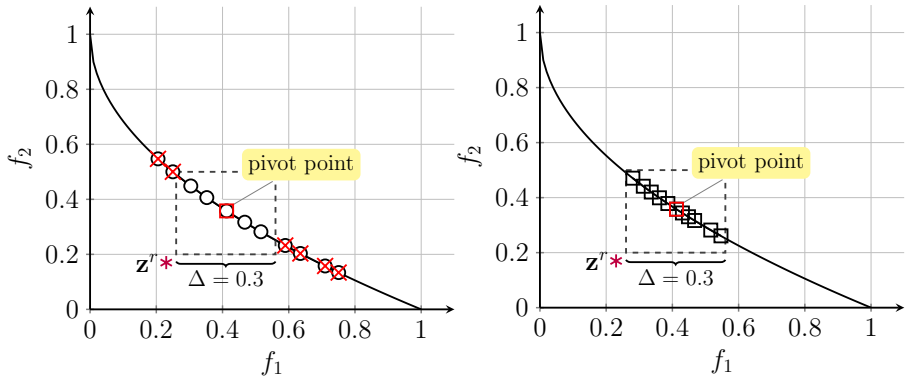
return S

4.2 Further Comments

1. In this paper, we set $\mathbf{z}^w = \mathbf{z}^r + 2.0 \times \mathbf{w}$ for proof of principle studies, where $\mathbf{w}$ is an unit vector. This setting implies that all objectives are equally important. However, in practice, different objectives might have various importance to the DM. For example, if we set $\mathbf{w} = (\frac{2}{\sqrt{5}}, \frac{1}{\sqrt{5}})^T$, the first objective is assumed to be twice less important than the second one. In particular, the importance is in the inverse order of the weights.

2. In practice, the DM has no idea about the range of the whole PF, not to mention the extent of ROI. Thus, Δ plays as an approximate expectation of the relative extent of ROI comparing to the whole PF. Note that the objective space is assumed to be normalized to $[0, 1]$.

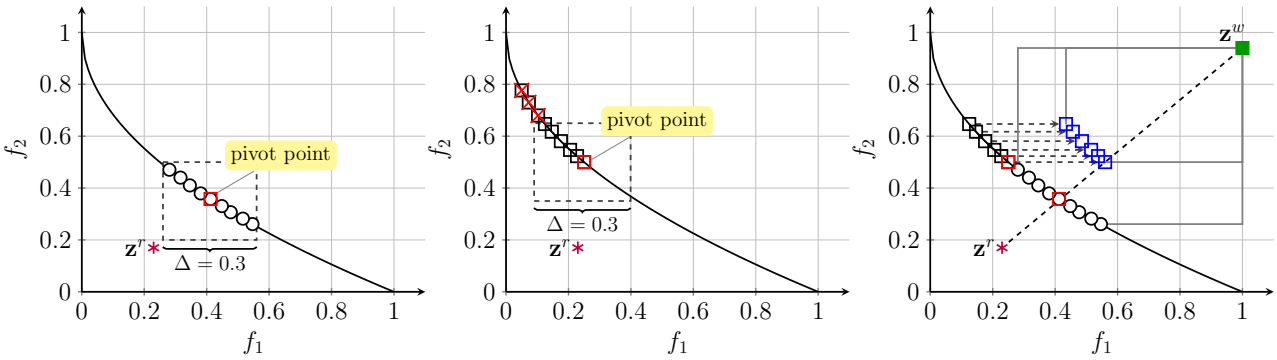
3. The trimming procedure penalizes the solution set, in term of the diversity, for having an excessive extent or deviating from the ROI. For example, in Fig. 5(a) and Fig. 5(b), both S^1 and S^2 have the same cardinality ($|S^1| = |S^2| = 11$), but S^1 has a wider spread. From the DM's perspective (given $\Delta = 0.3$), S^2 is preferable than S^1. After the trimming procedure, $|S^1|$ reduces to 5 while $|S^2|$ is still the same. Accordingly, S^1 will sacrifice its diversity when calculating the R-metric value. As for another example shown in Fig. 6(a) and Fig. 6(b), S^1 and S^2 again have the same cardinality. Since S^2 deviates from the ROI, its pivot point is identified as one of its extreme. After the trimming procedure, solutions far away from the ROI are excluded from further consideration.



(a) Trimming procedure for S^1. (b) Trimming procedure for S^2.

Figure 5: Illustration of the trimming procedure.

4. Given the DM's preference information, the convergence is not only the closeness to the PF, but also the *closeness* of the transferred points to the ROI along the preferred direction. This re-definition of convergence is fulfilled by transfering points to a virtual position along the iso-ASF line between $\mathbf{z}^p$ and its iso-ASF point $\mathbf{z}^l$. Let us consider the example shown in Fig. 6 again. The pivot point of S^1 is exactly its iso-ASF point, since this pivot point lies exactly



(a) Trimming procedure for S^1. (b) Trimming procedure for S^2. (c) Solution transfer.

Figure 6: Illustration of the R-metric calculation principle. Note that the solution transfer is according to iso-ASF lines for the ASF with $\mathbf{z}^r$ as the reference point and $\mathbf{z}^w - \mathbf{z}^r$ as the weight vector.

on the reference line connecting $\mathbf{z}^r$ and $\mathbf{z}^w$. In this case, solutions of S^1 stay in their original positions after the transfering procedure. In contrast, since the pivot point of S^2 deviates from $\mathbf{z}^r$, the remaining solutions of the processed S^2 after the trimming procedure are shifted, along the direction $\mathbf{z}^l - \mathbf{z}^p$, to a virtual position away from the PF, as shown in Fig. 6(c).

4.3 Time Complexity Analysis

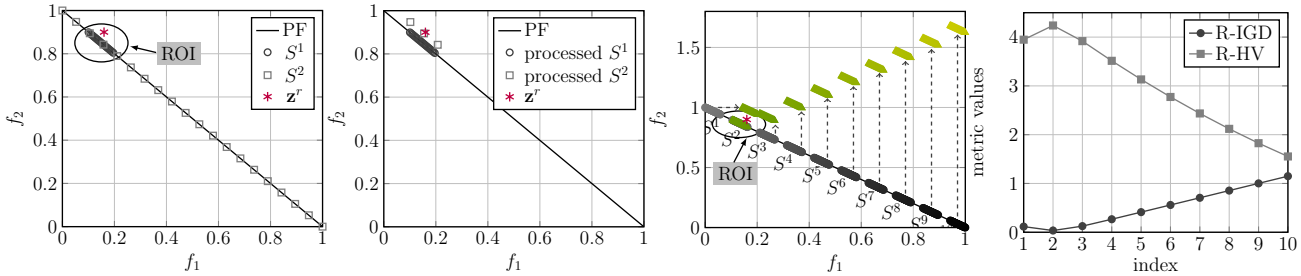
We now analyze the time complexity of R-metric calculation. The prescreening procedure requires at most $\mathcal{O}(L^2N^2)$ dominance comparisons. The identification of a pivot point $\mathbf{z}^p$ from N trade-off points requires $\mathcal{O}(N)$ computations. The trimming procedure can be achieved in $\mathcal{O}(N)$ computations. Transfering at most N points towards the reference line also requires $\mathcal{O}(N)$ computations. The complexity of R-IGD or R-HV computation in last step is the same as the regular IGD or HV. In summary, the time complexity of R-IGD is $\max\{\mathcal{O}(NM), \mathcal{O}(L^2N^2)\}$, where M is size of the newly formed P^* after the filtering procedure; the time complexity of R-HV is $\max\{\mathcal{O}(N^{m-1}), \mathcal{O}(L^2N^2)\}$ by using the method developed in [33].

5 Empirical Studies on Artificial Scenarios

In this section, we verify the proposed R-metrics on several artificial scenarios. First of all, the effectiveness of R-metrics is investigated on the two toy examples introduced in Section 3. Afterwards, four popular benchmark problems with different population distributions and different reference point settings are used to further examine R-metrics. Next, we investigate the applicability of R-metrics on a problem with disconnected PFs. Finally, R-metrics are used to evaluate the quality of a preferred solution set with respect to multiple reference points.

5.1 Investigations on Toy Examples

For the first example discussed in Section 3, as shown in Fig. 7(b), S^1 and S^2 are rearranged according to our proposed R-metric calculation procedure. Since S^2 does not fully meet the DM's preference, points therein are penalized in terms of convergence and diversity. More specifically, after the filtering procedure, only three points have been left in the processed S^2. Then, the processed S^2 has been shifted to an area away from the PF in the solution transfer step. Thereafter, the R-metric values of S^1 should be better than that of S^2, i.e., R-IGD(S^1) = 0.0222 and R-IGD(S^2) = 0.0452; R-HV(S^1) = 1.1550 and R-HV(S^2) = 1.1108. As for the second example discussed in Section 3, now shown in Fig. 7(c), all ten point sets have been shifted to their corresponding virtual positions along the preferred direction (the dashed arrow denotes the transfer direction for each point set). Now R-IGD and R-HV successfully figure out the best point set S^2. This is because the processed S^2, denoted as P^2, is closest to the ROI. In addition, S^1 and S^3, which have a similar distance to the ROI



(a) Partial region vs whole region. (b) After data pre-processing. (c) Ten point sets along the PF. (d) R-IGD and R-HV values for each set.

Figure 7: Illustration of R-IGD and R-HV values with respect to a DM supplied reference point $\mathbf{z}^r = (0.16, 0.9)$.

after data pre-processing, achieve similar R-IGD and R-HV values. As for S^4 to S^{10}, their R-metric values become worse with their increasing distances to the ROI.

5.2 Investigations on Benchmark Problems

In this subsection, we investigate the effectiveness of R-metrics on four classic benchmark problems, including two-objective ZDT1 and ZDT2 [43], and three-objective DTLZ1 and DTLZ2 [44][2]. For ZDT1 and ZDT2, ten sets of points, S^1 to S^{10}, are sampled from different regions of their PFs. Each set contains 40 evenly distributed points. For DTLZ1 and DTLZ2, we sample 21 sets of points from different regions of their PFs, where each set has 25 evenly distributed points. For R-IGD calculation, we first sample 10,000 evenly distributed points from the corresponding PFs. Only those located in the ROI (Δ is set as 0.2) are chosen to form P^* in the R-IGD calculation.

5.2.1 Two-objective Cases

As shown in Fig. 8(a), we investigate three kinds of reference points for ZDT1.

- *Unattainable reference point* $\mathbf{z}^{r_1} = (0.2, 0.5)^T$: From the results shown in Fig. 8(b), we find that both R-IGD and R-HV are able to make a reasonable assessment on the quality of a point set with respect to the DM supplied preference information. For example, S^3 resides in the ROI with respect to $\mathbf{z}^{r_1}$. As shown in Fig. 8(b), the R-IGD and R-HV values obtained by S^3 are indeed the best. For the other point sets, the farther away from the ROI, the worse the R-metric values are.

- *Attainable reference point* $\mathbf{z}^{r_2} = (0.6, 0.3)^T$: Similar to the observations in the above scenario, the point set closest to the reference point, i.e., S^6, obtain the best R-IGD and R-HV values. And the R-metric values also depend on the distance to the ROI.

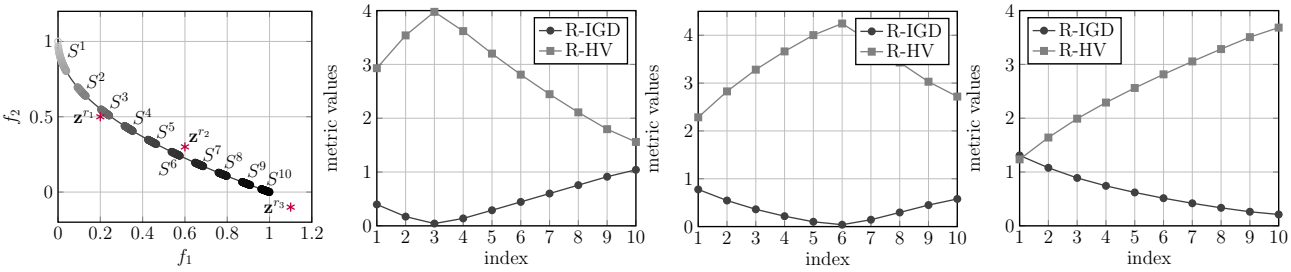
- *Extreme reference point* $\mathbf{z}^{r_3} = (1.1, -0.1)^T$: Since this reference point lies on one extreme, it is expected that the point set at the respective extreme boundary, i.e., S^{10}, is desirable. From the results shown in Fig. 8(d), we find that our proposed R-metrics are able to capture this fact and their trajectories show a monotone property.

5.2.2 Three-objective Cases

As shown in Fig. 9, we investigate two kinds of reference points for DTLZ1.

- *Unattainable reference point* $\mathbf{z}^{r_1} = (0.05, 0.05, 0.2)^T$: Different from the two-objective case, in a three-dimensional space, points are distributed in an ambient space where the neighboring

[2]Due to the page limit, the results on ZDT2 and DTLZ2, which are similar to the observations on ZDT1 and DTLZ1, are put in the supplementary file, which can be downloaded from http://coda-group.github.io/supp-rmetric.pdf.

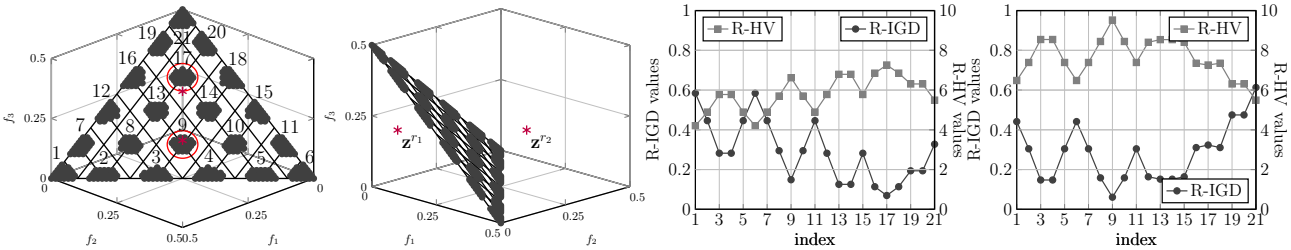


(a) Population distribution. (b) $\mathbf{z}^{r_1} = (0.2, 0.5)^T$. (c) $\mathbf{z}^{r_2} = (0.6, 0.3)^T$. (d) $\mathbf{z}^{r_3} = (1.1, -0.1)^T$.

Figure 8: Variations of R-IGD and R-HV with respect to an unattainable reference point $\mathbf{z}^{r_1} = (0.2, 0.5)^T$, an attainable reference point $\mathbf{z}^{r_2} = (0.6, 0.3)^T$ and a outside reference point $\mathbf{z}^{r_3} = (1.1, -0.1)^T$ on ZDT1 problem.

points can be in various directions. This explains the significant fluctuations of R-IGD and R-HV curves shown in Fig. 9(c). Nevertheless, the point set most relevant to the DM supplied preference information, i.e., S^{17}, obtains the best R-IGD and R-HV values. Furthermore, we also notice that the point sets close to the reference point obtain similar R-metric values, e.g., $S^9, S^{13}, S^{14}, S^{16}$ and S^{18} in DTLZ1 get similar R-IGD and R-HV values as shown in Fig. 9(d).

- *Attainable reference point* $\mathbf{z}^{r_2} = (0.3, 0.3, 0.2)^T$: From the results shown in Fig. 9(d), we find that R-metrics are able to provide a reliable assessment on different point sets. S^9, which is closest to this reference point, obtains the best R-IGD and R-HV values.

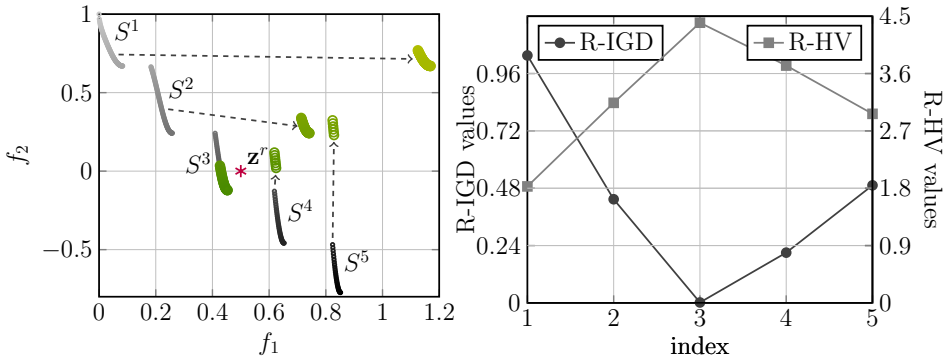


(a) Population distribution (front view). (b) Population distribution (side view). (c) $\mathbf{z}^r = (0.05, 0.05, 0.2)^T$. (d) $\mathbf{z}^r = (0.3, 0.3, 0.2)^T$.

Figure 9: Variations of R-IGD and R-HV with respect to an unattainable reference point $\mathbf{z}^r = (0.05, 0.05, 0.2)^T$, an attainable reference point $\mathbf{z}^r = (0.3, 0.3, 0.2)^T$ on DTLZ1 problem.

5.3 Investigations on Problems with Disconnected PFs

Although the investigations in Section 5.2.1 and Section 5.2.2 are based on problems with continuous PFs, our proposed R-metrics are also effective for problems with disconnected PFs. To validate this issue, this section chooses ZDT3, whose PF consists of five disconnected segments, for investigation. Five point sets, S^1 to S^5 as shown in Fig. 10(a), are respectively sampled from each segment and $\mathbf{z}^r = (0.5, 0.0)^T$. In addition, we also plot the transferred point sets, denoted as green circles, for illustration (the dashed arrow denotes the transfer direction for each point set). From the results shown in Fig. 10(b), we find that the R-metrics are able to provide a reasonable quality assessment for problems with disconnected PFs. In particular, S^3, which is closest to $\mathbf{z}^r$, obtains the best R-IGD and R-HV values. Note that the transferred points of S^3 still lie on S^3 and is indeed closest to $\mathbf{z}^r$. From Fig. 10(a), we find that S^4 is also very close to $\mathbf{z}^r$, the same for its transferred points. This explains that its R-metric values are very similar to that of S^3. On the other hand, S^1, which is farthest away from $\mathbf{z}^r$, obtains the worst R-metric values.



(a) Population distribution. (b) $\mathbf{z}^r = (0.5, 0.0)^T$.

Figure 10: Variations of R-IGD and R-HV with respect to a reference point $\mathbf{z}^r = (0.5, 0.0)^T$ on ZDT3 problem.

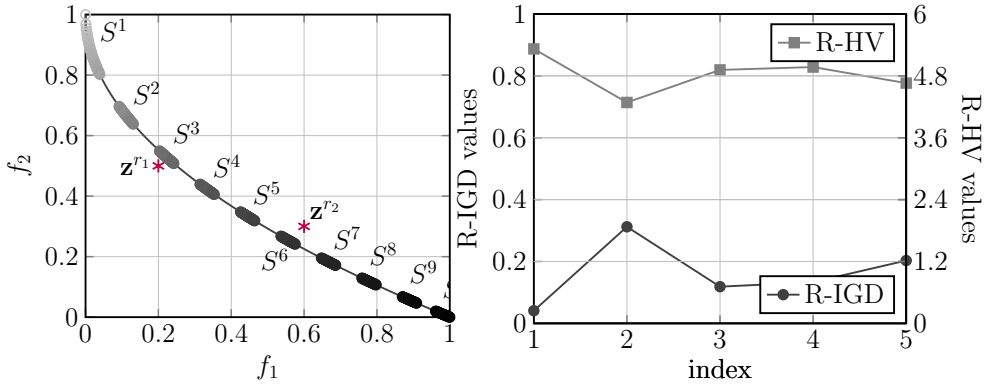
5.4 Investigations on Multiple Reference Points

In practice, the DM might not be sure about his/her exact preference beforehand. The DM would like to simultaneously explore several ROIs by supplying T ($1 < T \leq |S|$) reference points at the same time. Accordingly, a small modification is required to adapt the R-metrics to this circumstance. Generally speaking, when there are more than one reference point, the pre-processing procedure should take each reference point into consideration separately. At first, we use the prescreening procedure introduced in Section 4.1.1 to remove those dominated solutions from further R-metric calculation. Afterwards, we apply the k-means [45] algorithm to divide the remaining solutions into T clusters. Then, each cluster is associated with a reference point closest to its centroid. For each cluster, we use Step 2 to Step 4 introduced in Section 4.1 to pre-process the points and transfer them to a virtual position. Finally, we combine all pre-processed points together and evaluate their R-metric values as a whole. In particular, the worst point for R-HV calculation is chosen as the nadir point of the worst point for each reference point; for R-IGD calculation, P^* needs to be pre-processed for each reference point separately according to the method introduced in Section 4.1.

To validate the effectiveness of our strategy, we take the example in Fig. 11(a) for investigation. Here we set two reference points $\mathbf{z}^{r_1} = (0.2, 0.5)^T$ and $\mathbf{z}^{r_2} = (0.6, 0.3)^T$, simultaneously, in the objective space. Five point set combinations, i.e., $(S^3, S^6), (S^1, S^2), (S^4, S^5), (S^3, S^4), (S^6, S^7)$, are chosen as the candidates for performance assessment. From the results shown in Fig. 11(b), we find that (S^3, S^6) obtains the best R-IGD and R-HV values. From Fig. 11(a), we can see that S^3 and S^6 are in the corresponding ROI of $\mathbf{z}^{r_1}$ and $\mathbf{z}^{r_2}$, respectively. In contrast, (S^1, S^2) obtains the worst R-metric values. From Fig. 11(a), we find that both S^1 and S^2 are close to $\mathbf{z}^{r_1}$, but are far away from $\mathbf{z}^{r_2}$. Therefore, the R-metric values with respect to $\mathbf{z}^{r_1}$ can be acceptable, whereas the R-metric values with respect to $\mathbf{z}^{r_2}$ should be significantly bad. This makes its final R-metric values become the worst. Notice that sometimes DMs tend to use multiple reference points to discretely approximate the ROI. Therefore, the regions between these supplied reference points are also very important. From Fig. 11(b), we found that the R-IGD and R-HV values obtained by (S^4, S^5) and (S^3, S^4) are similar and are only inferior to (S^3, S^6). In contrast, although (S^6, S^7) has some part locating in the ROI of $\mathbf{z}^{r_2}$, S^7 is far away from $\mathbf{z}^{r_1}$. This makes its R-metric values not as good as (S^4, S^5) and (S^3, S^4). In summary, due to the existence of multiple reference points, a good point set should have a promising satisfaction for every reference point.

6 Empirical Studies on Preference-based EMO Algorithms

In this section, we apply our proposed R-metrics to evaluate the performance of the following four preference-based EMO algorithms. Notice that all multi-objective optimizers use reference points to articulate the DM's preference information, and all of them, except g-NSGA-II, are capable of handling more than one reference point. Here we choose the classic ZDT and DTLZ test suites as benchmark



(a) Population distribution. (b) Point sets combination.

Figure 11: Variations of R-IGD and R-HV with respect to two different reference point $\mathbf{z}^{r_1} = (0.2, 0.5)^T$ and $\mathbf{z}^{r_2} = (0.6, 0.3)^T$ simultaneously. Index 1 indicates (S^3, S^6), 2 indicates (S^1, S^2), 3 indicates (S^4, S^5), 4 indicates (S^3, S^4) and 5 indicates (S^6, S^7).

problems. For R-IGD computation, similar to Section 5, we at first sample 10,000 points from the corresponding PF. Then, points located in the ROI (Δ is set as 0.2) are used to form P^*.

1. *r-MOEA/D-STM* [46]: It is an extension of our recently proposed MOEA/D variant based on stable matching model [47]. Different from the original MOEA/D, where the selection of next parents is merely determined by the ASF value of a solution, MOEA/D-STM treats subproblems and solutions as two sets of agents and considers their mutual-preferences simultaneously. In particular, the preference of a subproblem over a solution measures the convergence issue, while the preference of a solution over a subproblem measures the diversity issue. Since the stable matching achieves an equilibrium between the mutual-preferences between subproblems and solutions, MOEA/D-STM strikes a balance between convergence and diversity of the search process. In order to incorporate the DM's preference information into the search process, we need to specify a population of weight vectors spread around reference points.

2. *R-NSGA-II* [24]: It is a variant of NSGA-II which modifies the crowding operator based on the idea of classic reference point based method. More specifically, solutions close to reference points have a larger chance to survive in the selection procedure. In addition, R-NSGA-II employs an ϵ-clearing idea to control the spread of the final obtained solutions in the ROI.

3. *g-NSGA-II* [48]: It modifies NSGA-II by replacing the Pareto dominance with a new dominance relation, called g-dominance. More specifically, g-dominance uses a reference vector to represent DM's desired value for each objective, i.e., aspiration levels. Solutions either satisfying all aspiration levels or fulfilling none of the aspiration levels are preferable than those merely satisfying some aspiration levels.

4. *r-NSGA-II* [25]: It uses a new dominance relation, called r-dominance, to replace the Pareto dominance in NSGA-II. When two solutions are non-dominated in terms of Pareto dominance, the one closer to the reference point is preferable. Moreover, its search behavior is adjusted by two parameters: one is the non-r-dominance threshold δ that controls the spread of the obtained solutions; the other is the ASF weight vector that controls the relative importance of different objectives.

All multi-objective optimizers use simulated binary crossover (SBX) [49] and polynomial mutation [50] as the reproduction operators. For proof of principle purpose, all four algorithms assume all objectives are equally important in our empirical studies. Note that although r-MOEA/D-STM, R-NSGA-II and r-NSGA-II are able to control the spread of obtained solutions, there is no specific guideline to set the corresponding parameter. Due to the page limit, the parameter settings and the specifications of reference points for different test instances are given in Section II of the supplementary file.

6.1 Empirical Studies on Benchmark Problems

Table 1: Comparison results of R-IGD and R-HV values on unattainable reference point.

R-metric	Test Instance	r-MOEA/D-STM	R-NSGA-II	g-NSGA-II	r-NSGA-II
R-IGD	ZDT1	**2.497E-2(3.51E-4)**	3.266E-2(2.01E-2)†	3.289E-2(1.27E-4)†	1.103E-1(2.72E-2)†
	ZDT2	3.164E-2(7.04E-6)†	4.065E-2(1.86E-2)†	**1.633E-3(4.83E-4)**	1.229E-1(1.37E-2)†
	ZDT3	2.231E-2(9.07E-4)†	4.908E-2(3.50E-2)†	**5.814E-3(5.19E-4)**	2.580E-1(1.98E-1)†
	ZDT4	**2.462E-2(5.85E-6)**	5.572E-2(3.50E-2)†	3.587E-2(4.07E-3)†	2.961E-1(1.01E-1)†
	ZDT6	3.701E-2(2.24E-2)	**1.583E-2(1.18E-2)**	1.879E-2(8.60E-3)†	1.084E-1(2.75E-2)†
	DTLZ1	**4.242E-2(2.00E-3)**	1.033E-1(1.65E-2)†	1.456E+2(1.59E+2)†	2.890E+1(1.31E+1)†
	DTLZ2	**3.517E-2(3.47E-4)**	8.163E-2(2.25E-3)†	3.959E-2(1.02E-2)	5.481E-2(1.03E-2)†
	DTLZ3	**3.724E-2(1.72E-3)**	8.465E-2(4.36E-3)†	2.047E+2(7.25E+1)†	1.755E+2(5.79E+1)†
	DTLZ4	5.285E-2(7.93E-2)	1.209E-1(1.00E-1)†	**3.583E-2(8.61E-3)**	1.189E+0(2.28E-16)†
	DTLZ5	1.712E-1(2.15E-5)†	1.863E-1(1.02E-2)†	**1.699E-1(1.38E-3)**	2.957E-1(5.76E-2)†
	DTLZ6	**2.669E-1(5.89E-3)**	4.395E-1(2.54E-2)†	2.819E+0(5.43E-1)†	6.450E+0(9.43E-1)†
	DTLZ7	1.915E+0(5.19E-4)†	1.894E+0(1.56E-1)†	**1.874E+0(6.52E-5)**	2.095E+0(1.24E-1)†
R-HV	ZDT1	4.0304(2.62E-3)	**4.0317(1.00E-1)**	3.9972(6.79E-4)	3.6857(1.02E-1)†
	ZDT2	3.7614(1.03E-4)†	3.6432(1.05E-1)†	**3.8954(3.38E-3)**	3.3403(6.85E-2)†
	ZDT3	3.7994(3.34E-3)†	3.7314(1.41E-1)†	**3.8945(4.68E-3)**	3.1460(5.22E-1)†
	ZDT4	**4.0329(4.10E-5)**	3.9258(1.64E-1)	3.9846(1.77E-2)†	3.1770(2.56E-1)†
	ZDT6	4.2267(1.03E-1)†	**4.3625(6.18E-2)**	4.3317(2.27E-2)†	4.0107(1.27E-1)†
	DTLZ1	**7.8684(2.18E-2)**	6.5847(6.00E-2)†	0†	0†
	DTLZ2	**7.4370(1.09E-2)**	6.6173(3.30E-2)†	7.4063(1.21E-1)	7.1551(1.68E-1)†
	DTLZ3	**7.4226(1.72E-2)**	6.5847(6.00E-2)†	0†	0†
	DTLZ4	7.3139(5.42E-1)	6.3934(6.23E-1)†	**7.4486(9.56E-2)**	1.7280(4.56E-16)†
	DTLZ5	5.3916(2.73E-4)†	5.1611(1.33E-1)†	**5.5771(2.12E-2)**	4.3066(3.43E-1)†
	DTLZ6	**4.5685(5.87E-2)**	3.6037(1.22E-1)†	0.0417(6.83E-2)†	0†
	DTLZ7	7.6506(4.36E-3)†	7.8907(1.04E+0)†	**7.9993(3.93E-4)**	6.3770(8.25E-1)†

† denotes the best mean metric value is significantly better than the others according to the Wilcoxon's rank sum test at a 0.05 significance level. – indicates all obtained solutions are dominated by the other counterparts, and thus no useful solution can be used for R-metric computation.

Each algorithm is performed 31 independent runs, and the R-metric values for two different reference point settings are respectively given in Table 1 and Table 2. In particular, the best mean metric values are highlighted in bold face with gray background, and the Wilcoxon's rank sum test at a 0.05 significance level is used to compare the statistical significance of the difference between the best mean metric value and the others. To have a visual comparison, we also plot the final solutions obtained by different algorithms having the best R-IGD value. Due to the page limit, they are presented in Section III of the supplementary file. In the following paragraphs, we will separately discuss the effectiveness of the R-metrics for evaluating the performance of different algorithms on problems with continuous and disconnected PFs.

6.1.1 Problems with continuous PFs

ZDT1 and ZDT2 are two relatively simple test instances, where all four algorithms do not have too much difficulty in finding solutions around the DM supplied reference points. However, as shown in Fig. 3(d) and Fig. 4(d) of the supplementary file, the convergence of solutions found by r-NSGA-II is not satisfied enough. This makes most of its obtained solutions be trimmed during the prescreening step of the R-metric calculation. Accordingly, its R-IGD and R-HV values are the worst among all four algorithms. As for r-MOEA/D-STM and g-NSGA-II, their performance is visually similar on finding preferred solutions for ZDT1 and ZDT2. However, the R-IGD and R-HV values obtained by g-NSGA-II are better than r-MOEA/D-STM in 6 out of 8 comparisons. Let us look at Fig. 3(a) and Fig. 3(c) of the supplementary file, for the unattainable reference point, all solutions found by r-MOEA/D-STM well converge to the PF whereas some solutions found by g-NSGA-II are not fully converged. In this case, the R-metric values obtained by r-MOEA/D-STM are better than g-NSGA-II. For the other three cases (i.e., ZDT1 with an attainable reference point and ZDT2 with both unattainable and attainable reference points), although solutions found by r-MOEA/D-STM well converge to the PF, their overall distributions deviate from the ROI a bit. In contrast, solutions found by g-NSGA-II not only converge to the PF, but also have a well concentration on the ROI. Therefore, it should be preferable and our proposed R-metrics also make a reasonable assessment. ZDT4 has the same PF

Table 2: Comparison results of R-IGD values on attainable reference point.

R-metric	Test Instance	r-MOEA/D-STM	R-NSGA-II	g-NSGA-II	r-NSGA-II
R-IGD	ZDT1	1.930E-2(3.08E-4)†	3.340E-2(1.39E-2)†	**4.294E-3(1.81E-4)**	1.408E-1(3.22E-2)†
	ZDT2	1.839E-2(2.66E-3)†	3.389E-2(1.59E-2)†	**2.540E-3(2.18E-4)**	1.650E-1(3.41E-2)†
	ZDT3	3.062E-2(1.28E-4)†	5.188E-2(2.51E-2)†	**2.013E-2(1.25E-3)**	1.604E-1(3.15E-1)†
	ZDT4	**1.921E-2(1.55E-5)**	4.171E-2(1.96E-2)†	4.904E-2(3.34E-2)†	6.080E-1(1.64E-1)†
	ZDT6	4.181E-2(9.61E-3)†	**2.570E-2(1.17E-2)**	3.761E-2(2.56E-4)†	1.458E-1(9.44E-2)†
	DTLZ1	**2.650E-2(1.46E-3)**	7.121E-2(8.65E-3)†	–	2.430E+1(1.28E+1)†
	DTLZ2	**3.452E-2(2.69E-4)**	6.454E-2(2.74E-3)†	4.052E-2(1.23E-2)	1.127E-1(0.00E+0)
	DTLZ3	**3.528E-2(6.55E-4)**	9.186E-2(6.09E-3)†	–	1.711E+2(6.58E+1)†
	DTLZ4	**3.442E-2(2.96E-4)**	9.879E-2(9.02E-2)†	4.411E-2(1.14E-2)†	7.090E-1(0.00E+0)†
	DTLZ5	**9.413E-2(1.07E-5)**	1.030E-1(3.37E-3)†	–	2.356E-1(5.88E-3)†
	DTLZ6	**2.034E-1(1.22E-2)**	2.145E-1(9.18E-3)†	–	6.795E+0(1.61E+0)†
	DTLZ7	2.966E+0(3.02E-5)†	2.943E+0(5.65E-2)†	**2.889E+0(2.30E-2)**	–
R-HV	ZDT1	4.4556(1.45E-3)†	4.3811(8.22E-2)†	**4.5390(1.13E-3)**	3.9142(9.69E-2)†
	ZDT2	4.4803(1.30E-2)†	4.4341(1.03E-1)†	**4.6128(2.60E-3)**	3.9003(1.13E-1)†
	ZDT3	3.9489(9.07E-4)†	3.8572(1.10E-1)†	**3.9523(5.21E-3)**	3.6101(7.44E-1)†
	ZDT4	**4.4578(7.99E-5)**	4.3182(1.02E-1)†	4.3038(1.90E-1)†	2.7377(3.82E-1)†
	ZDT6	4.0521(2.73E-2)	**4.0679(7.44E-2)**	3.9975(1.23E-3)†	3.6650(3.00E-1)†
	DTLZ1	**10.0644(8.30E-3)**	9.3869(1.81E-1)†	–	0†
	DTLZ2	**10.3015(1.18E-1)**	9.4999(6.15E-2)†	10.2106(3.66E-3)†	8.8212(0.00E+0)†
	DTLZ3	**10.2039(7.15E-3)**	9.1023(8.38E-2)†	–	0†
	DTLZ4	**10.2343(1.33E-1)**	9.2082(6.68E-1)†	10.2119(3.49E-3)	4.9130(9.11E-16)†
	DTLZ5	**8.6172(2.11E-4)**	8.4080(5.82E-2)†	–	6.9070(4.12E-2)†
	DTLZ6	**7.9238(1.12E-1)**	7.1421(7.99E-2)†	–	0†
	DTLZ7	9.9382(3.03E-4)†	10.0296(4.71E-1)†	**10.6224(1.90E-1)**	–

† denotes the best mean metric value is significantly better than the others according to the Wilcoxon's rank sum test at a 0.05 significance level. – indicates all obtained solutions are dominated by the other counterparts, and thus no useful solution can be used for R-metric computation.

shape as ZDT1, but it is more difficult due to the presence of many local optima. All algorithms, except r-MOEA/D-STM, cannot find solution fully converge to the PF. Accordingly, r-MOEA/D-STM obtains the best R-metric values among all four algorithms. r-NSGA-II obtains the worst R-metric values since it only finds solutions close to the unattainable reference point. ZDT6 has a concave PF shape and a biased distribution in the search space. It is interesting to note that although the solutions found by r-MOEA/D-STM not only well converge to the PF but also have a uniform distribution, the R-metric values obtained by r-MOEA/D-STM are not as good as R-NSGA-II and g-NSGA-II. This might be explained as the representative point found by r-MOEA/D-STM is inferior to that found by R-NSGA-II and g-NGSA-II. In this case, the solution transfer step of the R-metric calculation can transfer the solutions found by r-MOEA/D-STM to a farther position. Due to the poor convergence property, the R-metric values obtained by r-NSGA-II is still the worst.

The PF of DTLZ1 is a simplex having an intercept of 0.5 at each coordinate. Due to the presence of $11^5 - 1$ local PFs, DTLZ1 causes difficulties for an EMO algorithm in reaching the global PF. From Fig. 7 of the supplementary file, only r-MOEA/D-STM well approximates the ROIs. Accordingly, it obtains the best R-metric values. It is interesting to note that solutions found by R-NSGA-II seem to have a nice concentration on the DM supplied reference points, but their spreads are too narrow. In contrast, neither g-NSGA-II nor r-NSGA-II finds any reasonable solution. DTLZ2 is a relatively simple test instance, where all four algorithms do not have much difficulty in finding solutions close to the ROIs. As shown in Fig. 8 of the supplementary file, it is clear that solutions found by r-MOEA/D-STM are preferable than the other candidates. Accordingly, its obtained R-metric values are also the best. Although solutions found by R-NSGA-II well converge to the PF, their spreads are too narrow. This explains its unsatisfied R-metric values. DTLZ3 has the same PF shape as DTLZ2, but is with $3^{10} - 1$ local PFs. From Fig. 9 of the supplementary file, it is obvious that g-NSGA-II and r-NSGA-II have some difficulties in converging to the PF. Thus their R-metric values are extremely poor. In contrast, solutions found by r-MOEA/D-STM and R-NSGA-II are similar to those in DTLZ2. DTLZ4 also has the same PF shape as DTLZ2, but it has a strong bias towards $f_3 - f_1$ plane. From Fig. 10 of the supplementary file, we can see that solutions found by r-MOEA/D-STM are significantly better than the other three algorithms. Accordingly, its R-IGD and R-HV values are the best. DTLZ5 and DTLZ6 are two degenerate problems, where the latter one has a strong bias away from the PF. From

Fig. 11 of the supplementary file, we find that solutions obtained by r-MOEA/D-STM are preferable since they converge well to the PF and have a good approximation to the ROIs. Accordingly, its obtained R-metric values are also the best among four algorithms. Although solutions found by R-NSGA-II converge well to the PF, their spreads are too narrow. For DTLZ6, all four algorithms have difficulties in converging to the PF. Solutions found by r-MOEA/D-STM seem to be closer to the PF and have a wide spread. Accordingly, it obtains the best R-metric values.

Table 3: Comparisons of R-IGD and R-HV values on DTLZ2 with 5 and 10 objectives.

R-metric	# of objectives	r-MOEA/D-STM	R-NSGA-II	g-NSGA-II	r-NSGA-II
R-IGD	$m = 5$	**2.155E-1(9.61E-3)**	3.861E-1(3.39E-3)†	4.012E+0(6.72E-1)†	2.995E+0(5.65E-1)†
	$m = 10$	5.924E-1(2.44E-2)†	**4.609E-1(4.44E-3)**	–	3.731E+0(8.41E-1)†
R-HV	$m = 5$	**27.4088(6.58E-1)**	15.1592(1.25E-1)†	0.0058(1.53E-2)†	0.1646(3.64E-1)†
	$m = 10$	**1054.1653(5.41E+1)**	1105.4065(4.61E+1)†	–	3.9650(1.03E+1)†

† denotes the best mean metric value is significantly better than the others according to the Wilcoxon's rank sum test at a 0.05 significance level.

6.1.2 Problems with disconnected PFs

After the empirical studies on problems with continuous PFs, this subsection investigates the effectiveness of our proposed R-metrics on two problems with disconnected PFs. For ZDT3, as shown in Fig. 13 of the supplementary file, all four algorithms are able to find solutions close to the reference points, but those found by g-NSGA-II are visually better where most solutions converge to the PF and the spread is satisfactory. Accordingly, the R-metric values obtained by g-NSGA-II are better than the other three algorithms. For DTLZ7, as shown in Fig. 14 of the supplementary file, although solutions found by r-NSGA-II have a well focus on the ROIs, they are away from the PF. This explains its poorest R-metric values. All the other three algorithms find some solutions on the PF segments outside the ROIs. In particular, solutions found by r-MOEA/D-STM spread over all four PF segments, while those found by g-NSGA-II are the visual best as shown in Fig. 14 of the supplementary file. Accordingly, g-NSGA-II obtains the best R-IGD and R-HV values on both unattainable and attainable reference points.

6.2 Empirical Studies on Many-objective Problems

Recently, problems with more than three objectives have become one of the hottest topics in EMO. Due to the expansion of the objective space in size, many-objective problems cause several challenges to the traditional EMO algorithm design. For example, the mushrooming of non-dominated solutions in a population significantly weaken the selection pressure of Pareto-based EMO methods, and the sparse distribution of a limited number of solutions in a high-dimensional space makes the density estimation and diversity management become even more difficult than the two- and three-objective cases. As discussed in [51], instead of searching for the whole PF, finding a preferred subregion satisfying the DM's preference information is more practical in many-objective optimization.

In this section, we investigate the scalability of R-metrics for quantitatively assessing the quality of a preferred solution set in a high-dimensional space. DTLZ2 with 5 and 10 objectives are chosen as the test instances. The reference point is set as $\mathbf{z}^r = (0.1, 0.3, 0.2, 0.4, 0.2)^T$ for the 5-objective test instance, and $\mathbf{z}^r = (0.3, 0.3, 0.3, 0.1, 0.3, 0.55, 0.35, 0.35, 0.25, 0.45)^T$ for the 10-objective case. For the R-IGD calculation, we employ the method suggested in [51] to sample 101,270 points from DTLZ2's PF in the 5-objective case, and 3,124,550 points in the 10-objective case. Moreover, Δ is increased to 0.5 in the R-metric computation due to the sparse distribution of solutions in a high-dimensional space. For preference-based EMO algorithms, all parameters are kept the same as Section 6, except the number of function evaluations. Specifically, it is set as 80,000 and 150,000 for the 5- and 10-objective case respectively. Table 3 shows the comparisons of R-metric values and the parallel coordinates of the populations with the medium R-IGD value are plotted in Fig. 15 and Fig. 16 of the supplementary file. In particular, the red dotted line represents the reference point. From the results shown in these two figures, it is clear that g-NSGA-II and r-NSGA-II are the worst optimizers. This observation is

also confirmed by their worst R-IGD and R-HV values. Both r-MOEA/D-STM and R-NSGA-II are able to find solutions around the ROIs. However, solutions found by R-NSGA-II almost concentrate on the reference point, while those found by r-MOEA/D-STM have a wider spread. This explains the better R-HV values obtained by r-MOEA/D-STM in these two test instances.

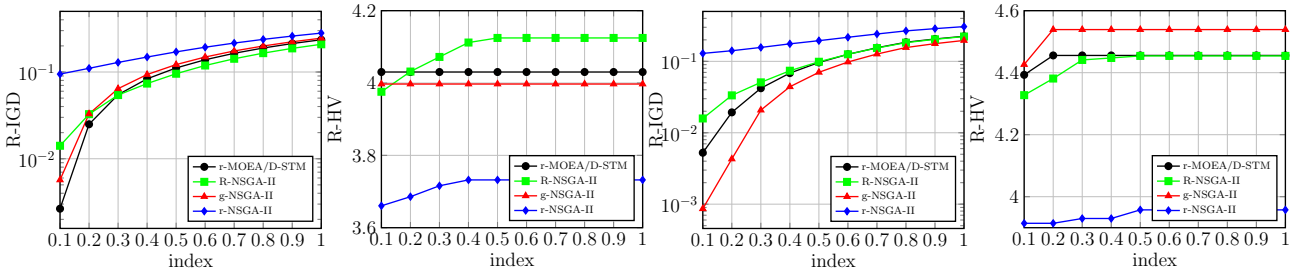
6.3 Further Investigations

In this section, we further investigate some other interesting properties of R-metrics. ZDT1 and DTLZ2 are chosen as the test instances, since all four EMO algorithms have no difficulty on solving them. For each test instance, we keep a record of R-IGD and R-HV values of an intermediate population every 10 consecutive generations. Fig. 17 of the supplementary file plots the variations of R-IGD and R-HV values versus the number of generations on ZDT1 with $\mathbf{z}^r = (0.3, 0.4)^T$. From this figure, we find that all algorithms, except r-NSGA-II, converge to their optimal R-IGD and R-HV values within a few generations. In contrast, the R-metric trajectories of r-NSGA-II grow slowly with the number of generations. Fig. 18 of the supplementary file plots some intermediate populations for different algorithms. From these four subfigures, we find that r-NSGA-II approximates the preferred region in a layer-wise manner while the other three algorithms converge to the preferred region rapidly (with around 100 generations). These observations are in accord with the corresponding R-metric trajectories. For DTLZ2, we have a similar observation. As shown in Fig. 19 of the supplementary file, the R-metric trajectories of r-MOEA/D-STM and R-NSGA-II converge to a stable value within a few generations. These observations are also validated by the plots of intermediate populations in Fig. 20(a) and Fig. 20(b) of the supplementary file. As for g-NSGA-II, we also notice some fluctuations in its R-metric trajectories. This observation is also in line with the fluctuations of the evolutionary population as shown in Fig. 20(c) of the supplementary file. The R-metric trajectories of r-NSGA-II are rather rugged. From Fig. 20(d) of the supplementary file, we find that the intermediate populations of r-NSGA-II vibrate significantly during the search process.

From the above experiments, we have another interesting observation that some algorithms do not need the predefined number of generations to find preferred solutions. For example, as shown in Fig. 18(a) and Fig. 20(a) of the supplementary file, r-MOEA/D-STM only uses around 100 generations to converge to the preferred region for ZDT1 and around 80 generations for DTLZ2. Furthermore, an algorithm almost converges to the preferred region when the R-metric trajectories become stable. Based on these observations, we keep a record of the standard deviations of the R-IGD and R-HV values obtained by different algorithms for every 10 and 25 consecutive generations. In addition, we set two thresholds $\tau = 0.1$ and $\tau = 0.01$ and to see how many generations an algorithm needs to have a R-metric's standard deviation less than τ. From the empirical results shown in Table III of the supplementary file, we observe that the standard deviation of R-IGD can be reduced to the given thresholds when the time window is set to 10 generations. More interestingly, the number of generations that makes the standard deviation of R-IGD reduce to 0.01 is similar to the required budgets of the corresponding algorithm converges to the preferred region. Moreover, we also notice that the standard deviation of R-HV cannot always be reduced to the given thresholds on DTLZ2. However, if we extend the time window to 25 generations, even the standard deviation of R-IGD cannot drop down to the expected thresholds in many cases. From this experiment, we find that our proposed R-metrics are not only reliable metrics to evaluate the performance of a preference-based EMO algorithm, more interestingly, the variation of a certain R-metric (e.g., R-IGD with a time window of 10 generations and standard deviation's threshold $\tau = 0.1$) can be used as a stopping criterion in searching for a preferred solution set.

6.4 Influence of Δ

As described in Section 4, Δ represents the DM's expectation of the ROI's relative extent comparing to the whole PF. It is also used to trim the irrelevant solutions for the R-metric calculation. A large Δ means that the DM prefers solutions having a wide spread, while a small Δ indicates that the the DM prefers solutions having a good concentration on the ROI. This section takes ZDT1 as an example to investigate the influence of Δ on R-metric values, where Δ varies from 0.1 to 1.0 with an



(a) $\mathbf{z}^r = (0.3, 0.4)^T$ (b) $\mathbf{z}^r = (0.3, 0.4)^T$ (c) $\mathbf{z}^r = (0.65, 0.3)^T$ (d) $\mathbf{z}^r = (0.65, 0.3)^T$

Figure 12: Variations of R-metric values for different Δ.

increment of 0.1. Fig. 12 shows the variations of R-metric values obtained by four preference-based EMO algorithms for different Δ.

Let us start from $\mathbf{z}^r = (0.3, 0.4)^T$. As shown in Fig. 12(a), the R-IGD value obtained by R-NSGA-II is worse than r-MOEA/D-STM and g-NSGA-II when Δ is small. But it becomes the best in case Δ is larger than 0.2. Moreover, the R-IGD value obtained by r-NSGA-II is the worst when Δ is small. However, when Δ is larger than 0.4, the R-IGD values obtained by r-MOEA/D-STM, g-NSGA-II and r-NSGA-II are almost the same. Let us refer to Fig. 3 of the supplementary file, as for $\mathbf{z}^r = (0.3, 0.4)^T$, solutions found by R-NSGA-II and r-NSGA-II have a wider spread than those of r-MOEA/D-STM and g-NSGA-II. If the DM expects the ROI to be concentrated on his/her provided reference point, i.e., Δ is set to be small, solutions found by r-MOEA/D-STM and g-NSGA-II are preferable. Accordingly, the R-metric values obtained by the previous two algorithms should be better. On the flip side, if the DM expects the ROI to be widely spread, i.e., Δ is set to be large, solutions found by r-MOEA/D-STM and g-NSGA-II are not satisfactory any longer. Even though the solutions found by R-NSGA-II and r-NSGA-II are not well converged, their wide spread meet the DM's expectation and provide him/her more choices. This explains their better R-metric values when Δ becomes large. As for $\mathbf{z}^r = (0.65, 0.3)$, since solutions found by g-NSGA-II not only well converge to the PF, but also have a wide spread around the DM supplied reference point, its R-metric values are constantly better than the other competitors. In contrast, although solutions obtained by R-NSGA-II still have a wide spread along the PF, their convergence is poor. Therefore, the R-metric values of R-NSGA-II are worse than g-NSGA-II.

It is worth noting that although some preference-based EMO algorithms (e.g., [24], [25] and [46]) claim to be able to control the ROI's extent by setting an appropriate parameter, to the best of our knowledge, there is no rule-of-thumb to set the corresponding parameter. We believe that Δ, used in the R-metric calculation, is able to provide a general guideline to tune the corresponding parameters in a posterior manner.

7 Conclusions and Future Works

Given the DM's preference information, approximating a partial and preferred PF, rather than the whole PF, has been one of the most important topics in the modern EMO research. Besides developing effective algorithms that drive solutions towards the ROI, how to evaluate the quality of a set of preferred trade-off solutions is of the same importance but has rarely been studied in this literature. In this paper, we presented a systematic way to evaluate the quality of a set of preferred solutions obtained by a preference-based EMO using reference points. More specifically, we pre-process the preferred solutions, according to a MCDM approach, before using a regular metrics for performance assessment. In particular, according to the DM's expectation of the ROI's extent, our proposed R-metric has a trimming procedure that penalizes the population diversity of a preferred solution set having an excessive extent. Furthermore, inspired by the ASF-based ranking from the MCDM literature, our proposed R-metric has a transferring procedure that transfers the preferred trade-off solutions to a virtual position according to their satisfaction degree to the DM supplied preference information. Extensive experiments on several artificial scenarios and benchmark problems fully demonstrate the

efficacy of our proposed R-metrics for evaluating the quality of a preferred solution set according to the DM supplied preference information.

This work is a very first attempt to systematically and quantitatively evaluating the quality of a preferred solution set. Much more attention and effort should be required on this topic. In future, we want to explore the following issues:

1. Note that there is no single way of expressing the DM's preference information, so we may not be able to expect a universal way for performance assessment. This paper assumes that the DM's preference information is expressed in terms of a reference point. However, there exist other types of preferences to which our proposed R-metric may not be directly useful. To solve this drawback, one may consider the method presented in [52] to adapt the R-metric to other types of preferences.

2. As described in Section 4.2, the setting of Δ assumes the objective space is normalized to $[0, 1]$. However, in practice, this assumption might not always hold. It is interesting to investigate other method to specify the DM's expectation of the relative extent of ROI with respect to objectives in different scales.

3. To avoid wasting computational resources and to examine the formal convergence and optimality achieved by an EMO algorithm, the research on the online stopping criteria (OSC) has obtained increasing popularity. Empirical studies in Section 6.3 shows a simple application of our proposed R-metrics for designing OSC. Nevertheless, more sophisticated and advanced techniques [53] are worth being studied in future.

4. In addition to the empirical studies, it is of importance to have a rigorous analysis of the *optimal archive* with respect to the R-metric. This is not only useful for better understanding the behavior of R-metric itself, but also for providing foundations in the case of using the R-metric to design an indicator-based algorithm in future.

Acknowledgment

This work was supported by EPSRC (grant no. EP/K001523/1) and NSFC (grant no. 61329302). Xin Yao was also supported by a Royal Society Wolfson Research Merit Award.

References

[1] 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, Apr. 2002.

[2] K. Li, S. Kwong, R. Wang, K.-S. Tang, and K.-F. Man, "Learning paradigm based on jumping genes: A general framework for enhancing exploration in evolutionary multiobjective optimization," *Information Sciences*, vol. 226, pp. 1–22, 2013.

[3] K. Li, K. Deb, Q. Zhang, and Q. Zhang, "Efficient non-domination level update method for steady-state evolutionary multiobjective optimization," *IEEE Transactions on Cybernetics*, 2016, accepted for publication.

[4] K. Li, S. Kwong, R. Wang, J. Cao, and I. J. Rudas, "Multi-objective differential evolution with self-navigation," in *SMC'12: Proc. of the 2012 IEEE International Conference on Systems, Man, and Cybernetics*. Seoul, Korea(South): IEEE, Oct. 2012, pp. 508–513.

[5] E. Zitzler and S. Künzli, "Indicator-based selection in multiobjective search," in *PPSN'04: Proc. of the 8th International Conference on Parallel Problem Solving from Nature*, 2004, pp. 832–842.

[6] K. Li, S. Kwong, J. Cao, M. Li, J. Zheng, and R. Shen, "Achieving balance between proximity and diversity in multi-objective evolutionary algorithm," *Information Sciences*, vol. 182, no. 1, pp. 220–242, 2012.

[7] N. Beume, B. Naujoks, and M. Emmerich, "SMS-EMOA: Multiobjective selection based on dominated hypervolume," *European Journal of Operational Research*, vol. 181, no. 3, pp. 1653–1669, 2007.

[8] Q. Zhang and H. Li, "MOEA/D: A multiobjective evolutionary algorithm based on decomposition," *IEEE Transactions on Evolutionary Computation*, vol. 11, pp. 712–731, Dec. 2007.

[9] K. Li, Á. Fialho, S. Kwong, and Q. Zhang, "Adaptive operator selection with bandits for a multiobjective evolutionary algorithm based on decomposition," *IEEE Transactions on Evolutionary Computation*, vol. 18, no. 1, pp. 114–130, 2014.

[10] K. Li, S. Kwong, and K. Deb, "A dual population paradigm for evolutionary multiobjective optimization," *Information Sciences*, vol. 309, pp. 50–72, 2015.

[11] K. Li, S. Kwong, Q. Zhang, and K. Deb, "Interrelationship-based selection for decomposition multiobjective optimization," *IEEE Transactions on Cybernetics*, vol. 45, no. 10, pp. 2076–2088, 2015.

[12] H. Ishibuchi, N. Tsukamoto, and Y. Nojima, "Evolutionary many-objective optimization: A short review," in *CEC'08: Proc. of the 2008 IEEE Congress on Evolutionary Computation*, 2008, pp. 2419–2426.

[13] K. Deb and H. Jain, "An evolutionary many-objective optimization algorithm using reference-point-based nondominated sorting approach, part I: solving problems with box constraints," *IEEE Transactions on Evolutionary Computation*, vol. 18, no. 4, pp. 577–601, 2014.

[14] C. M. Fonseca and P. J. Fleming, "Multiobjective optimization and multiple constraint handling with evolutionary algorithms—Part I: A unified formulation," *IEEE Transactions on Systems, Man, and Cybernetics, Part A: Systems and Humans*, vol. 28, pp. 26–37, 1998.

[15] J. Branke, T. Kaussler, and H. Schmeck, "Guidance in evolutionary multi-objective optimization," *Advances in Engineering Software*, vol. 32, no. 6, pp. 499–507, 2001.

[16] K. C. Tan, E. F. Khor, T. H. Lee, and R. Sathikannan, "An evolutionary algorithm with advanced goal and priority specification for multi-objective optimization," *Journal of Artificial Intelligent Research*, vol. 18, pp. 183–215, 2003.

[17] Y. Jin, T. Okabe, and B. Sendhoff, "Adapting weighted aggregation for multiobjective evolution strategies," in *EMO'01: Proc. of the first International Conference on Evolutionary Multi-Criterion Optimization*, 2001, pp. 96–110.

[18] Y. Jin and B. Sendhoff, "Incorporation of fuzzy preferences into evolutionary multiobjective optimization," in *GECCO'02: Proc. of the 2002 Genetic and Evolutionary Computation Conference*, 2002, p. 683.

[19] I. C. Parmee and D. Cvetkovic, "Preferences and their application in evolutionary multiobjective optimization," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 1, pp. 42–57, 2002.

[20] K. Deb, "Multi-objective evolutionary algorithms: Introducing bias among pareto-optimal solutions," in *Advances in Evolutionary Computing*, A. Ghosh and S. Tsutsui, Eds. Springer Berlin Heidelberg, 2003, pp. 263–292.

[21] J. Branke and K. Deb, "Integrating user preferences into evolutionary multi-objective optimization," in *Knowledge Incorporation in Evolutionary Computation*, Y. Jin, Ed. Springer Berlin Heidelberg, 2005, vol. 167, pp. 461–477.

[22] K. Deb and A. Kumar, "Interactive evolutionary multi-objective optimization and decision-making using reference direction method," in *GECCO'07: Proc. of the 9th Genetic and Evolutionary Computation Conference*, 2007, pp. 781–788.

[23] A. Wierzbicki, "The use of reference objectives in multiobjective optimisation," in *MCDM theory and Application, Proceedings*, ser. Lecture notes in economics and mathematical systems, F. G. and G. T., Eds., no. 177. Hagen: Springer Verlag, 1980, pp. 468–486.

[24] K. Deb, J. Sundar, U. Bhaskara, and S. Chaudhuri, "Reference point based multiobjective optimization using evolutionary algorithms," *International Journal of Computational Intelligence Research*, vol. 2, no. 3, pp. 273–286, 2006.

[25] L. B. Said, S. Bechikh, and K. Ghédira, "The r-dominance: A new dominance relation for interactive evolutionary multicriteria decision making," *IEEE Transactions on Evolutionary Computation*, vol. 14, no. 5, pp. 801–818, 2010.

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

[27] D. Brockhoff, J. Bader, L. Thiele, and E. Zitzler, "Directed Multiobjective Optimization Based on the Hypervolume Indicator," *Journal of Multi-Criteria Decision Analysis*, vol. 20, pp. 291–317, 2013.

[28] T. Wagner, H. Trautmann, and D. Brockhoff, "Preference articulation by means of the R2 indicator," in *EMO'13: Proc. of the 7th International Conference on Evolutionary Multi-Criterion Optimization*. Springer, 2013, pp. 81–95.

[29] G. Rudolph, O. Schütze, C. Grimme, and H. Trautmann, "A multiobjective evolutionary algorithm guided by averaged hausdorff distance to aspiration sets," in *EVOLVE - A Bridge between Probability, Set Oriented Numerics, and Evolutionary Computation V*. Springer International Publishing, 2014, vol. 288, pp. 261–273.

[30] S. Bechikh, M. Kessentini, L. B. Said, and K. Ghédira, "Chapter four - preference incorporation in evolutionary multiobjective optimization: A survey of the state-of-the-art," *Advances in Computers*, vol. 98, pp. 141–207, 2015.

[31] D. V. Veldhuizen and G. B. Lamont, "Evolutionary computation and convergence to a pareto front," in *Late Breaking Papers at the Genetic Programming Conference*, 1998, pp. 221–228.

[32] J. Knowles, L. Thiele, and E. Zitzler, "A Tutorial on the Performance Assessment of Stochastic Multiobjective Optimizers," Computer Engineering and Networks Laboratory (TIK), ETH Zurich, Switzerland, Tech. Rep., 2006.

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

[34] J. R. Schott, "Fault tolerant design using single and multicriteria genetic algorithm optimization," Master's thesis, Massachusetts Institute of Technology, May 1995.

[35] A. Farhang-Mehr and S. Azarm, "An information-theoretic entropy metric for assessing multi-objective optimization solution set quality," *Journal of Mechanical Design*, vol. 125, no. 4, pp. 655–663, 2004.

[36] K. C. Tan, T. H. Lee, and E. F. Khor, "Evolutionary algorithms for multi-objective optimization: Performance assessments and comparisons," *Artificial Intelligence Review*, vol. 17, no. 4, pp. 251–290, 2002.

[37] P. Bosman and D. Thierens, "The balance between proximity and diversity in multiobjective evolutionary algorithms," *IEEE Transactions on Evolutionary Computation*, vol. 7, no. 2, pp. 174–188, Apr. 2003.

[38] R. L. While, L. Bradstreet, and L. Barone, "A fast way of calculating exact hypervolumes," *IEEE Transactions on Evolutionary Computation*, vol. 16, no. 1, pp. 86–95, 2012.

[39] O. Schütze, X. Esquivel, A. Lara, and C. A. Coello Coello, "Using the averaged hausdorff distance as a performance measure in evolutionary multi-objective optimization," *IEEE Transactions on Evolutionary Computation*, vol. 16, no. 4, pp. 504–522, Aug. 2012.

[40] U. K. Wickramasinghe, R. Carrese, and X. Li, "Designing airfoils using a reference point based evolutionary many-objective particle swarm optimization algorithm," in *CEC'10: Proc. of the 2010 IEEE Congress on Evolutionary Computation*, 2010, pp. 1–8.

[41] A. Mohammadi, M. N. Omidvar, and X. Li, "A new performance metric for user-preference based multi-objective evolutionary algorithms," in *CEC'13: Proc. of the 2013 IEEE Congress on Evolutionary Computation*, 2013, pp. 2825–2832.

[42] A. P. Wierzbicki, "The use of reference objectives in multiobjective optimization," in *Multiple Criteria Decision Making Theory and Applications*, G. Fandel and T. Gal, Eds. Berlin: Springer-Verlag, 1980, pp. 468–486.

[43] E. Zitzler, K. Deb, and L. Thiele, "Comparison of multiobjective evolutionary algorithms: Empirical results," *Evolutionary Computation*, vol. 8, no. 2, pp. 173–195, 2000.

[44] K. Deb, L. Thiele, M. Laumanns, and E. Zitzler, "Scalable test problems for evolutionary multiobjective optimization," in *Evolutionary Multiobjective Optimization*, ser. Advanced Information and Knowledge Processing, A. Abraham, L. Jain, and R. Goldberg, Eds. Springer London, 2005, pp. 105–145.

[45] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," *ACM Computing Survey*, vol. 31, no. 3, pp. 264–323, 1999.

[46] K. Li, Q. Zhang, S. Kwong, M. Li, and R. Wang, "Stable matching based selection in evolutionary multiobjective optimization," *IEEE Transactions on Evolutionary Computation*, vol. 18, no. 6, pp. 909–923, 2014.

[47] D. Gale and L. S. Shapley, "College admissions and the stability of marriage," *American Mathematical Monthly*, vol. 69, pp. 9–15, 1962.

[48] J. Molina, L. V. Santana, A. G. Hernández-Díaz, C. A. C. Coello, and R. Caballero, "g-dominance: Reference point based dominance for multiobjective metaheuristics," *European Journal of Operational Research*, vol. 197, pp. 685–692, 2009.

[49] K. Deb and R. B. Agrawal, "Simulated binary crossover for continuous search space," *Complex Systems*, vol. 9, pp. 1–34, 1994.

[50] K. Deb and M. Goyal, "A combined genetic adaptive search (GeneAS) for engineering design," *Computer Science and Informatics*, vol. 26, pp. 30–45, 1996.

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

[52] M. Luque, R. Caballero, J. M. Luque, and F. Ruiz, "Equivalent information for multiobjective interactive procedures," *Management Science*, vol. 53, no. 1, pp. 125–134, 2007.

[53] T. Wagner, H. Trautmann, and L. Martí, "A taxonomy of online stopping criteria for multiobjective evolutionary algorithms," in *EMO'11: Proc. of the 6th International Conference on Evolutionary Multi-Criterion Optimization*, 2011, pp. 16–30.