Proceedings of the 2025 SICE Festival with Annual Conference
September 9-12, 2025, Chiang Mai, Thailand

Comparison of Performance Indicators for Robust Multi-Objective Optimization: A Case Study Using Evolutionary Approaches

Takuro Tanaka[1†] and Masaya Nakata[1]

[1]Faculty of Engineering, Yokohama National University, Kanagawa, Japan
(Tel: +81-45-339-4139; E-mail: tanaka-takuro-cp@ynu.jp, nakata-masaya-tb@ynu.ac.jp)

Abstract: Many practical multi-objective optimization problems require obtaining robust Pareto solutions to perturbations in decision variables. While some efforts have been made to develop evolutionary algorithms for robust multi-objective optimization, little attention has been paid to designing sound experimental protocols, including appropriate performance indicators, for accurately assessing algorithm performance. To address this gap, this paper conducts a comparative study of performance indicators for robust multi-objective optimization. We compare two popular indicators based on Inverted Generational Distance to investigate which one offers a more reliable assessment of algorithms under the common goal of obtaining robust Pareto solutions. To this end, we employ multi-objective evolutionary algorithms as test algorithms, and evaluate the resulting Pareto solutions using the two performance indicators. Experimental results reveal that performance evaluations vary depending on the chosen indicator, and in some cases, yield results inconsistent with the goal of robust optimization. Based on these findings, we discuss the design of an appropriate experimental protocol, focusing particularly on the choice of performance indicators.

Keywords: Robust Optimization, Multi-objective Optimization, Evolutionary Algorithm

1. INTRODUCTION

Many real-world applications involve multi-objective optimization problems (MOPs), where multiple objectives must be optimized simultaneously [1]. The primary goal in MOPs is to obtain globally optimal Pareto solutions, which are not dominated by any other solutions. However, in practice, such optimal solutions may be sensitive to perturbations in decision variables, meaning that their solution quality can degrade depending on these perturbations. For example, in manufacturing, it is undesirable to obtain solutions whose quality is highly sensitive to production variations (correspond to variable perturbations). In such cases, robust multi-objective optimization [2]—that is, obtaining Pareto-optimal solutions that are of higher quality and less sensitive to variable perturbations—is often more desirable. A popular definition of robust multi-objective optimization is to optimize *mean effective objective functions* (MEOFs) [3], which return the mean value of objective values of all neighboring solutions of a given solution (see Section 2.2 for detailed formalizations).

Multi-objective evolutionary algorithms (MOEAs) [4] are a representative approach for solving MOPs. During the past two decades, numerous efforts have been devoted to developing efficient MOEA algorithms, applying them to industry, and establishing experimental protocols for fair comparison [5,6]. For example, popular algorithms include NSGA-II [7], MOEA/D [8], and IBEA [9], along with many of their variants. Regarding experimental protocols, a key focus has been the development of appropriate performance indicators (i.e., performance metrics) to properly assess the performance of algorithms based on the obtained Pareto solutions [10]. This pursuit stems from the fact that different performance indicators—such as Hypervolume (HV) [11] and Inverted Generational Distance (IGD) [12]—can lead to different rankings of algorithms [6, 13].

Despite numerous efforts in the field, few studies have addressed robust multi-objective optimization [2, 14]. In particular, limited attention has been given to assessing the reliability of performance indicators. Specifically, the sensitivity of Pareto solutions to variable perturbations is usually evaluated by defining the robustness of their solution quality. Accordingly, performance indicators adapted for robust multi-objective optimization are typically designed to incorporate this concept of robustness. Popular performance indicators in robust multi-objective optimization include the mean IGD [2] and the MEOF-based IGD [15], denoted as $\mathrm{IGD}_{\mathrm{mean}}$ and $\mathrm{IGD}_{\mathrm{MEOF}}$, respectively. These indicators are based on the IGD metric but differ in how they define and incorporate the robustness of solution quality. In a nutshell, $\mathrm{IGD}_{\mathrm{mean}}$ evaluates solutions based on how the original IGD values vary on average under variable perturbations. In contrast, $\mathrm{IGD}_{\mathrm{MEOF}}$ directly evaluates the robustness of Pareto solutions based on the optimal solutions of MEOFs. As aforementioned, different performance indicators may lead to different rankings of algorithms even in robust multi-objective optimization. However, to the best of our knowledge, this has not been thoroughly validated, and there is no in-depth analysis to investigate a proper performance indicator for robust multi-objective optimization.

Accordingly, this paper conducts a comparative study of the performance indicators for robust multi-objective optimization. We compare the two indicators, $\mathrm{IGD}_{\mathrm{mean}}$ and $\mathrm{IGD}_{\mathrm{MEOF}}$, to investigate which indicator provides a more reasonable assessment of algorithms under the common goal of obtaining robust Pareto solutions. To this end, we employ NSGA-II and NSGA-II-DTI [3] as

† Takuro Tanaka is the presenter of this paper.

test algorithms. We run both algorithms on benchmark problems, and the resulting Pareto solutions are evaluated using the the performance indicators. The contribution of this paper is twofold:

- To the best of our knowledge, this paper provides the first in-depth analysis of the adequacy of performance indicators in robust multi-objective optimization, contributing to the improvement of experimental protocols for more accurate assessments of algorithms in the field.
- We further investigate the adequacy of a sampling-based variant of IGD_{MEOF}, where the integral calculation is replaced by a sampling-based approach, thereby expanding the applicability of IGD_{MEOF} to black-box problems.

The remainder of this paper is organized as follows. Section 2 introduces robust multi-objective optimization and explains the performance indicators: IGD_{mean} and IGD_{MEOF}. Section 3 presents the main experimental results of our comparative study. Section 4 provides an additional analysis for the sampling-based variant of IGD_{MEOF}. Finally, Section 5 provides our conclusion.

2. BACKGROUND

This section explains the definitions of MOPs and robust MOPs. Subsequently, it introduces the performance indicators adapted for robust MOPs.

2.1. MOPs

In this paper, we consider MOPs in the minimization form, given by

$$\begin{aligned} &\min_{\boldsymbol{x}} \boldsymbol{f}(\boldsymbol{x}) = (f_1(\boldsymbol{x}), \ldots, f_M(\boldsymbol{x})), \\ &\text{s.t. } \boldsymbol{x} \in \Omega = \prod_{i=1}^{D} [x_i^l, x_i^u], \end{aligned} \tag{1}$$

where $\Omega \in \mathbb{R}^D$ is a feasible region in the decision space; x_i^l and x_i^u are the lower and upper bounds of x_i; and M is the number of objectives.

Usually, because there is a trade-off relationship between objectives, the goal of Eq. (1) is to approximate the true Pareto front, which involves a set of non-dominated solutions (called Pareto optimal solutions). A non-dominated solution is defined as one that is not dominated[1] by any other solution, thereby representing a point on the true Pareto front.

2.2. Robust MOPs

Robust MOPs represent a category of MOPs that take into account perturbations in the decision variables, formalized as

$$\begin{aligned} &\min_{\boldsymbol{x}} \boldsymbol{f}(\boldsymbol{x}+\Delta\boldsymbol{x}) = (f_1(\boldsymbol{x}+\Delta\boldsymbol{x}), \ldots, f_M(\boldsymbol{x}+\Delta\boldsymbol{x})), \\ &\text{s.t. } \boldsymbol{x} + \Delta\boldsymbol{x} \in B(\boldsymbol{x}|\delta) = \prod_{i=1}^{D} [x_i - \delta_i, x_i + \delta_i], \end{aligned} \tag{2}$$

where $\Delta\boldsymbol{x}$ is the perturbation of the decision variables, $B(\boldsymbol{x}) \in \mathbb{R}^D$ is a neighborhood region of the solution $\boldsymbol{x}$,

[1]A dominance relationship is formalized as follows. A solution $\boldsymbol{u}$ dominates a solution $\boldsymbol{v}$, if $f_i(\boldsymbol{u}) \leq f_i(\boldsymbol{v})$, $\forall\, i \in \{1, \ldots, M\}$ and $f_j(\boldsymbol{u}) < f_j(\boldsymbol{v})$, $\exists\, j \in \{1, \ldots, M\}$.

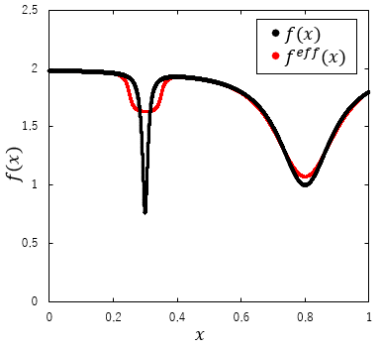

Fig. 1: The U-V valley landscape of $f(x)$.

and $\delta_i = \delta(x_i^u - x_i^l)$ is the maximum perturbation for the i-th decision variable x_i, determined by a perturbation ratio δ. Usually, a value of δ is given in advance. A main goal of robust MOPs is to obtain robust Pareto solutions that are of higher quality and less sensitive to variable perturbations. Specifically, robust MOPs of Eq. (2) can be re-formalized using mean effective objective functions (MEOFs) [3], as introduced in [15];

$$\min_{\boldsymbol{x}} \boldsymbol{f}^{\mathit{eff}}(\boldsymbol{x}) = (f_1^{\mathit{eff}}(\boldsymbol{x}), \ldots, f_M^{\mathit{eff}}(\boldsymbol{x})), \tag{3}$$

where $f_i^{\mathit{eff}}(\boldsymbol{x})$ is a MEOF for the i-th original objective function f_i, defined as

$$f_i^{\mathit{eff}}(\boldsymbol{x}) = \frac{1}{|B(\boldsymbol{x}|\delta)|} \int_{\boldsymbol{y} \in B(\boldsymbol{x}|\delta)} f_i(\boldsymbol{y}) d\boldsymbol{y}, \tag{4}$$

where $|B(\boldsymbol{x}|\delta)|$ is the volume of $B(\boldsymbol{x}|\delta)$. Using the MEOFs, both the quality and sensitivity of solutions to variable perturbations are quantified by integrating (i.e., averaging) f_i over the neighborhood region defined by $B(\boldsymbol{x}|\delta)$. Consequently, the robust MOPs formalized in Eq. (3) shift their goal to the minimization of the MEOFs of f_i, aiming to obtain high-quality, less sensitive Pareto solutions. Note that, due to this problem reformation, the robust MOPs of Eq. (3) yield a different true Pareto front from that of Eq. (1), referred to as the *robust Pareto front*, as it optimized with respect to $\boldsymbol{f}^{\mathit{eff}}(\boldsymbol{x})$.

For example, suppose the following function $f : \mathbb{R} \to \mathbb{R}$,

$$f(x) = 2 - \frac{1.2}{1 + \left(\frac{x-0.3}{0.01}\right)^2} - \frac{1}{1 + \left(\frac{x-0.8}{0.1}\right)^2}, \tag{5}$$

where its MEOF is obtained as;

$$\begin{aligned} f^{\mathit{eff}}(x) = 2 &- \frac{0.006}{\delta}\left(\arctan\left(\frac{x+\delta-0.3}{0.01}\right) - \arctan\left(\frac{x-\delta-0.3}{0.01}\right)\right) \\ &- \frac{0.05}{\delta}\left(\arctan\left(\frac{x+\delta-0.8}{0.1}\right) - \arctan\left(\frac{x-\delta-0.8}{0.1}\right)\right), \end{aligned} \tag{6}$$

The function f involves a U-V valley landscape, as shown by the black line in Fig. 1. The global optimal solution in the V-shaped valley, within the range $x \in [0.25, 0.35]$ is highly sensitive to variable perturbations when $\delta = 0.05$, which may be undesirable. In contrast, the local optimal solution in the U-shaped valley, within the range of $[0.6, 1.0]$ is more robust to perturbations. In this case, it becomes important to obtain the local optimum solution in the U-shaped valley, which is guided by optimizing f^{eff} (see the red line in Fig. 1).

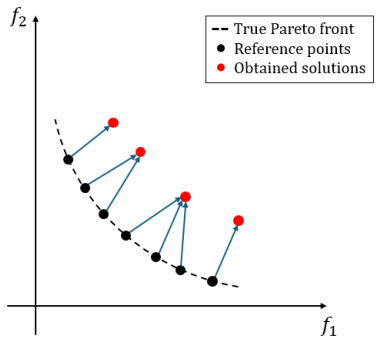

Fig. 2: IGD calculation.

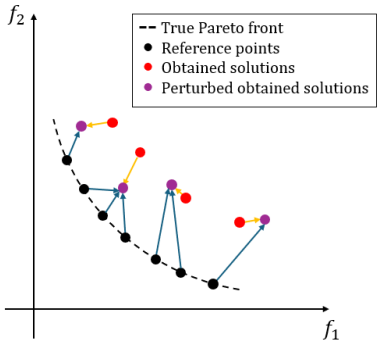

Fig. 3: IGD_{mean} calculation.

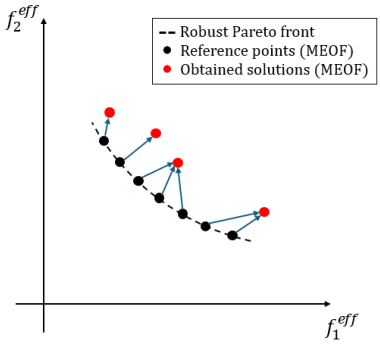

Fig. 4: IGD_{MEOF} calculation.

2.3. Performance indicators

In this section, we explain the the performance indicators adapted to robust multi-objective optimization, that is, the mean and MEOF-based IGDs.

To begin with, we firstly explain the IGD indicator as background information. IGD measures the average distance from *reference points* on the true Pareto front to their nearest solution among obtained Pareto solutions. Let $\mathcal{A}$ be a set of Pareto solutions obtained by an optimizer. An IGD score for $\mathcal{A}$ is then evaluated as

$$\text{IGD} = \frac{\sum_{i=1}^{|\mathcal{R}|} d_i}{|\mathcal{R}|}, \tag{7}$$

where $\mathcal{R} = \{r_i\}_{i=1}^{|\mathcal{R}|}$ is a set of $|\mathcal{R}|$ reference points; and d_i is the minimum Euclidean distance between r_i and its nearest solution in $\mathcal{A}$ (see Fig. 2). The reference points are sampled from the true Pareto front. Consequently, IGD evaluates how well the solutions in $\mathcal{A}$ approximate the true Pareto front; smaller IGD scores indicate higher approximation accuracy.

The remainder of this section describes the definition of the mean and MEOF-based IGDs.

2.3.1. Mean IGD

The mean IGD indicator, IGD_{mean}, evaluates the solutions in $\mathcal{A}$ based on how the original IGD values vary on average when variable perturbations are applied to those solutions. This indicator has been used in [2, 14, 16] to assess the performance of optimizers.

Algorithm 1 describes the pseudocode for IGD_{mean} calculation, where perturbation $\Delta\boldsymbol{x}$ is generated under the perturbation ratio δ, following the implementation provided in [2]. As show in this code, the perturbation is added to solutions in $\mathcal{A}$ for T times, resulting T IGD values calculated from Eq. (7). Then, the IGD_{mean} score is obtained as the mean of T IGD values.

Unlike the MEOF-based IGD, the mean IGD measures the robustness of the obtained Pareto solutions by evaluating the variance of the original IGD values. In other words, it assesses how the solutions move closer to or farther from the true Pareto front under perturbations, as shown in Fig. 3. Moreover, the mean IGD is broadly applicable to various problems, as it is a sampling-based indicator that does not require the objective functions to be integrable. A known drawback is the non-deterministic

Algorithm 1 IGD_{mean}

Input The set of reference points $\mathcal{R}$, The set of solutions obtained by the algorithm $\mathcal{A} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, The perturbation ratio δ, The number of test times under perturbations T
Ouput IGD_{mean}
for $t = 1$ **to** T **do**
  Generate a perturbation $\Delta\boldsymbol{x}^t$ under δ
  for $n = 1$ **to** N **do**
    Change the solution $\boldsymbol{x}_n$ to $\boldsymbol{x}_n^t$ by adding $\Delta\boldsymbol{x}^t$
  Select Pareto optimal solutions PS^t under the perturbations
  Calculate IGD^t from $\mathcal{R}$ and PS^t by Eq. (7)
$\text{IGD}_{\text{mean}} \leftarrow \text{mean}\ \{\text{IGD}^1, \ldots, \text{IGD}^T\}$
return IGD_{mean}

nature of IGD_{mean} calculation, as it depends on random sampling and the sample size.

2.3.2. MEOF-based IGD

The MEOF-based IGD indicator, IGD_{MEOF}, evaluates the robustness of Pareto solutions based on the robust Pareto front defined in MOPs of Eq. (3). This indicator has been used in [15].

The calculation of the MEOF-based IGD follows the same equation as that of IGD, i.e., Eq. (7), but reference points, $\mathcal{R} = \{r_i\}_{i=1}^{|\mathcal{R}|}$, are sampled from the robust Pareto front, as shown in Fig. 4. Unlike the mean IGD, the MEOF-based IGD can be calculated deterministically, as it does not rely on sampling.

Consequently, the MEOF-based IGD is designed to evaluate how well the obtained solutions in $\mathcal{A}$ approximate the robust Pareto front, rather than directly assessing the sensitivity of these solutions. This is because that the robust Pareto front itself involves optimal solutions that account for both quality and sensitivity, as the MEOFs are designed to quality sensitivity. However, a critical drawback is that the MEOF-based IGD in its basic form is applicable only to integrable objective functions, which are often less practical in real-world applications.

3. EXPERIMENT

This section compares the two performance indicators to investigate which indicator provides a more reasonable assessment of algorithms under the common goal of obtaining robust Pareto solutions. All the experiments conducted in this paper were derived using an evolutionary multi-objective optimization platform, PlatEMO [17].

3.1. Experiment settings

3.1.1. Benchmark problems

We used six benchmark problems developed for robust MOPs, namely TP8 [3] and RTP1–5 [15], all of which have a U-V valley landscape for the second decision variable x_2. The number of decision variables D was set to $D = \{5, 10, 20\}$. The ratio of perturbations δ was set to 0.05. All the problems are two-objective.

Note that RTP1–4 and RTP5 are modified versions of the well-known ZDT1–4 [18] and LZ1 [19] problems, respectively. For RTP1–4, the second objectives of these original problems were extended by adding $f(x_2)$ of Eq. (5), while for RTP5, both the first and second objectives were extended by this function. Accordingly, with $\delta = 0.05$, it is desirable to obtain robust Pareto solutions with x_2 values located in the U-shaped valley, as discussed in Section 2.2. Similarly, the second objective of TP8 was extended by adding another function $f'(x_2)$, and solutions with x_2 values located in the U-shaped valley (i.e., $x_2 = 0.35$) are the robust optimal solutions with $\delta = 0.05$.

3.1.2. Test algorithms

We used NSGA-II [7] and NSGA-II-DTI [3] as test algorithms, which are representative MOEAs designed for standard and robust MOPs, respectively. NSGA-II-DTI is an extension of NSAG-II, which evaluates solutions using sampling-based MEOFs instead of the original objective functions. In [3], NSGA-II-DTI has been demonstrated its effectiveness in producing robust Pareto solutions (i.e., solutions with x_2 values located in the U-shaped valley).

The following parameter settings were used; for NSGA-II, the population size $N = 100$, the distribution index of the simulated binary crossover $d_c = 20$, and that of the the polynomial mutation $\eta = 20$; for NSGA-II-DTI, $N = 100$, $d_c = 10$, $\eta = 50$, and the number of sample size for the sampling-based MEOFs $H = 30$. For both algorithms, the maximum number of generations was set to 300.

3.1.3. Evaluation criteria

To assess the adequacy of the mean and MEOF-based IGD indicators, we evaluated how closely the comparison results obtained using each indicator correspond to the *ground truth* comparison results. Specifically, we first ran both NSGA-II and NSGA-II-DTI on the selected benchmark problems for 21 trials to obtain the Pareto solutions derived by each algorithm. We calculated following comparison results.

- **The ground truth comparison results** were obtained based on the success counts, quantifying how many trials each algorithm produced solutions in the U-shaped valley. Technically, a success is defined as the average of x_2 values of solutions in the final population converging to 0.35 ± 0.01 for TP8 and to 0.8 ± 0.01 for RTP1–5, which corresponds to the optimum in the U-shaped valley (see Fig. 1).
- **IGD_{mean}-based comparison results** were obtained by calculating the IGD_{mean} scores for the obtained Pareto solutions by NSGA-II and NSGA-II-DTI. The number of sample size T was set to 1000, following existing works [2, 16]. As explained in Section 2.3.1, the IGD_{mean} score for a single trial is calculated by averaging 1000 original IGD values, each obtained by adding a different perturbation to the solutions. We then averaged the IGD_{mean} scores over 21 trials.
- **IGD_{MEOF}-based comparison results** were obtained by calculating the IGD_{MEOF} scores for the Pareto solutions by NSGA-II and NSGA-II-DTI. All the selected benchmark problems consist of integrable objective functions, making it possible to calculate IGD_{MEOF}. We then averaged the IGD_{MEOF} scores over 21 trials.

Furthermore, for the IGD_{mean}- and IGD_{MEOF}-based comparison results, we applied the Wilcoxon rank-sum test with a significant level of 5% to identify significant difference between the two algorithms.

Table 1: The success counts for NSGA-II and NSGA-II-DTI.

Problem	D	NSGA-II	NSGA-II-DTI
TP8	5	10	21
	10	11	21
	20	8	21
RTP1	5	5	21
	10	7	21
	20	0	21
RTP2	5	9	21
	10	5	21
	20	1	21
RTP3	5	5	21
	10	7	21
	20	2	21
RTP4	5	13	21
	10	10	21
	20	2	21
RTP5	5	0	21
	10	3	21
	20	0	21
Percentage		25.93%	100.0%

3.2. Experiment results

3.2.1. Ground truth comparison results

Table 1 shows the success counts for NSGA-II and NSGA-II-DTI. Fig. 5 shows the distributions of x_2 values in the final solutions obtained by NSGA-II and NSGA-II-DTI on RTP1 with $D = 10$ for a certain trial. As shown in Table 1, NSGA-II-DTI succeeded in producing robust solutions with x_2 values located in the U-shaped valley in all 21 trials across all problems, resulting in a 100% success rate (see the bottom of the table). In contrast, NSGA-II frequently failed to produce robust solutions, with an average success rate of only 25.93%. This indicates that NSGA-II often produced sensitive solutions with x_2 values located in the V-shaped valley, as shown Fig. 5.

The overall results demonstrate the strong effectiveness of NSGA-II-DTI compared to NSGA-II on robust MOPs.

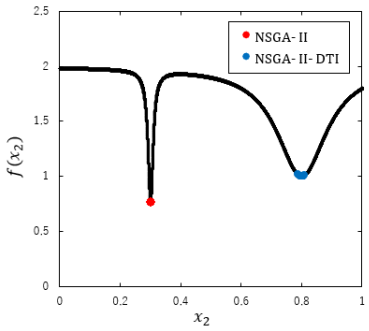

Fig. 5: Distributions of x_2 values in the final solutions obtained by NSGA-II and NSGA-II-DTI on RTP1 with $D = 10$.

3.2.2. Indicator-based comparison results

Table 2 shows the IGD_{mean}- and IGD_{MEOF}-based comparison results, where the best value for each problem instance is highlighted. In each table, the results of the Wilcoxon rank sum test were summarized as follows; the three symbols, "$+$", "$-$", and "$\approx$" indicate that NSGA-II is statistically better than, worse than or competitive with NSGA-II-DTI, respectively.

From Table 2-a, the IGD_{mean}-based comparison results indicate that IGD_{mean} values of NSGA-II were statistically *better* (i.e., smaller) than those of NSGA-II-DTI in 14 out of the 18 problem instances. The average rank of NSGA-II was also better than that of NSGA-II-DTI. These comparison results were clearly different from the grand truth results, suggesting that the IGD_{mean} indicator tends to underestimate the effectiveness of NSGA-II-DTI and misrepresents the actual algorithmic performances. In contrast, the IGD_{MEOF}-based comparison results are consistent with the grand truth results, showing that NSGA-II-DTI outperformed NSGA-II in 11 problem instances and was never outperformed by it (see Table 2-b). Moreover, for TP8 with $D = 5, 10, 20$, RTP1 with $D = 10$, RTP2 with $D = 5$, and RTP4 with $D = 5, 10$, NSGA-II frequently succeeded in producing robust solutions more than 6 of 21 trials (c.f. Table 1), which reasonably explains why no significant difference was detected for these problem instances.

The inaccurate performance assessment using IGD_{mean} can be attributed to the nature of the original IGD calculation. As explained in Section 2.3, IGD averages the minimum Euclidean distance between a reference point and its nearest solution among the obtained solutions. However, owing to this definition, a solution perturbed far from the true Pareto front does not significantly worsen the IGD value, as illustrated in Fig. 6; only perturbed solutions that are close to the true Pareto front (i.e., the reference points) are emphasized in the IGD score calculation. In the case illustrated in Fig. 6, the IGD value is not bad because some perturbed solutions become close to the true Pareto front, thereby improving the IGD scores. Although the goal in robust MOPs is to obtain high-quality, less sensitive Pareto solutions, IGD_{mean} cannot

Table 2: Average IGD_{mean} and IGD_{MEOF} values on TP and RTP problems.

		(a) IGD_{mean}		(b) IGD_{MEOF}	
Problem	D	NSGA-II	NSGA-II-DTI	NSGA-II	NSGA-II-DTI
TP8	5	1.053e−01 +	1.079e−01	3.173e−02 ≈	4.853e−04
	10	6.880e−01 ≈	7.017e−01	1.067e−01 ≈	1.974e−03
	20	2.406e+00 +	2.600e+00	3.850e−01 ≈	1.602e−02
RTP1	5	2.522e−01 +	4.660e−01	6.994e−02 −	1.997e−04
	10	2.832e−01 +	4.705e−01	2.761e−01 ≈	4.096e−04
	20	1.989e−01 +	4.728e−01	5.772e−01 −	1.059e−03
RTP2	5	3.244e−01 +	4.476e−01	4.561e−02 ≈	2.373e−04
	10	2.816e−01 +	4.718e−01	2.732e−01 −	4.882e−04
	20	2.423e−01 +	4.743e−01	4.695e−01 −	1.338e−03
RTP3	5	5.462e−01 +	1.150e+00	1.514e−01 −	3.301e−02
	10	6.442e−01 +	1.243e+00	7.420e−01 −	4.937e−02
	20	5.473e−01 +	1.169e+00	1.399e+00 −	2.532e−02
RTP4	5	3.384e+00 −	9.724e−01	2.735e−01 ≈	8.451e−03
	10	4.684e+01 −	1.842e+01	2.898e+00 ≈	2.562e−02
	20	1.550e+02 −	9.487e+01	2.998e+01 −	6.250e−02
RTP5	5	2.471e−02 +	3.160e−01	6.757e−02 −	1.077e−03
	10	8.165e−02 +	3.326e−01	1.338e−01 −	3.555e−03
	20	6.226e−02 +	3.464e−01	2.887e−01 −	1.041e−02
$+/-/\approx$		14/3/1	-	0/11/7	-
Avg. Rank		1.167	1.833	2.000	1.000

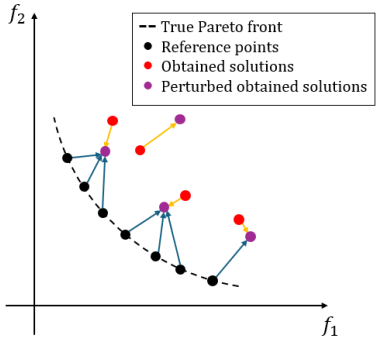

Fig. 6: Graphical example of overestimation of obtained solutions under variable perturbations.

appropriately indicate the sensitivity of solutions (i.e., the objective values do not vary significantly due to perturbations) in such cases. Thus, the IGD_{mean}-based performance assessment heavily depends on the specific perturbations applied, and this issue is difficult to mitigate simply by using a large number of perturbations (e.g., to 1000), due to the inherent characteristics of IGD calculation.

The overall results suggest a possible drawback of IGD_{mean} indicator and indicate that the IGD_{MEOF} indicator provides a more reasonable performance assessment, consistent with the ground truth results. Furthermore, the statistical analysis conducted on the IGD_{MEOF}-based results yields more consistent findings with the success counts observed for the algorithms.

4. ADDITIONAL EXPERIMENT

Although the previous section demonstrates the adequacy of the IGD_{MEOF} indicator, it has a critical limitation: IGD_{MEOF} in its basic form is only applicable to integrable objective functions, as IGD_{MEOF} is an integral-based indicator. One possible solution to overcome this limitation is to use a sampling-based variant of IGD_{MEOF} instead, thereby expanding the applicability of IGD_{MEOF}

Table 3: Average $\text{IGD}_{\text{MEOF-}T}$ values with $T = \{10, 50, 100, 1000\}$.

(a) $\text{IGD}_{\text{MEOF-10}}$ (b) $\text{IGD}_{\text{MEOF-50}}$ (c) $\text{IGD}_{\text{MEOF-100}}$ (d) $\text{IGD}_{\text{MEOF-1000}}$

Problem	D	NSGA-II	NSGA-II-DTI	NSGA-II	NSGA-II-DTI	NSGA-II	NSGA-II-DTI	NSGA-II	NSGA-II-DTI
TP8	5	3.788e−02 ≈	3.804e−02	2.929e−02 −	2.131e−03	2.962e−02 −	6.400e−04	3.104e−02 ≈	4.928e−04
	10	8.120e−02 ≈	7.140e−02	9.612e−02 −	2.455e−03	9.879e−02 ≈	1.698e−03	1.044e−01 ≈	1.973e−03
	20	2.663e−01 −	8.702e−02	3.447e−01 −	1.439e−02	3.564e−01 ≈	1.557e−02	3.773e−01 ≈	1.605e−02
RTP1	5	7.069e−02 −	1.227e−02	7.013e−02 −	1.283e−03	6.997e−02 −	6.405e−04	6.983e−02 −	2.412e−04
	10	2.552e−01 −	8.184e−03	2.740e−01 −	6.674e−04	2.753e−01 −	4.859e−04	2.758e−01 ≈	4.169e−04
	20	4.992e−01 −	8.210e−03	5.688e−01 −	9.733e−04	5.738e−01 −	9.921e−04	5.765e−01 −	1.049e−03
RTP2	5	5.230e−02 −	1.685e−02	4.572e−02 −	1.959e−03	4.558e−02 −	1.013e−03	4.515e−02 −	3.235e−04
	10	2.636e−01 −	1.162e−02	2.722e−01 −	1.051e−03	2.724e−01 −	6.492e−04	2.720e−01 −	5.068e−04
	20	4.228e−01 −	1.164e−02	4.639e−01 −	1.290e−03	4.668e−01 −	1.251e−03	4.689e−01 −	1.329e−03
RTP3	5	2.391e−01 ≈	2.126e−01	2.362e−01 −	1.897e−01	2.331e−01 −	1.884e−01	2.350e−01 −	1.865e−01
	10	6.399e−01 −	2.154e−01	7.148e−01 −	2.005e−01	7.218e−01 −	1.995e−01	7.250e−01 −	1.986e−01
	20	1.056e+00 −	1.962e−01	1.302e+00 −	1.772e−01	1.316e+00 −	1.766e−01	1.336e+00 −	1.763e−01
RTP4	5	5.891e−01 ≈	4.872e−01	3.142e−01 −	4.949e−02	2.726e−01 −	2.839e−02	2.735e−01 ≈	1.168e−02
	10	2.635e+00 −	5.991e−01	2.849e+00 −	7.798e−02	2.871e+00 ≈	4.770e−02	2.852e+00 ≈	2.759e−02
	20	2.746e+01 −	8.120e−01	2.968e+01 −	1.307e−01	2.968e+01 −	8.671e−02	2.988e+01 −	6.167e−02
RTP5	5	6.718e−02 −	1.179e−02	6.740e−02 −	1.072e−03	6.765e−02 −	1.071e−03	6.750e−02 −	1.053e−03
	10	1.335e−01 −	7.424e−03	1.338e−01 −	3.419e−03	1.340e−01 −	3.454e−03	1.338e−01 −	3.562e−03
	20	2.906e−01 −	7.103e−03	2.887e−01 −	1.027e−02	2.889e−01 −	1.035e−02	2.887e−01 −	1.042e−02
+/−/≈		0/14/4	-	0/18/0	-	0/15/3	-	0/12/6	-
Avg. Rank		1.944	1.056	2.000	1.000	2.000	1.000	2.000	1.000

to black-box problems. In this section, we further validate the adequacy of the sampling-based IGD_{MEOF} indicator in terms of its approximation accuracy relative to the true IGD_{MEOF} values.

To begin with, an MEOF for the i-th original objective function is approximated by its sampling-based variant [15], denoted as $\hat{f}_i^{\mathit{eff}}(\boldsymbol{x}|T)$, given by;

$$f_i^{\mathit{eff}}(\boldsymbol{x}) \simeq \hat{f}_i^{\mathit{eff}}(\boldsymbol{x}|T) = \frac{1}{T}\sum_{t=1}^{T} f_i(\boldsymbol{x}^{(t)}), \tag{8}$$

where T is the sample size; and $\boldsymbol{x}^{(t)}$ is a t-th solution sampled from the neighborhood region of $B(\boldsymbol{x}|\delta)$. The sampling-based IGD_{MEOF} indicator, denoted as $\text{IGD}_{\text{MEOF-}T}$, uses the $\hat{f}_i^{\mathit{eff}}(\boldsymbol{x}|T)$ values instead of the true $f_i^{\mathit{eff}}(\boldsymbol{x})$ values.

We conducted additional experiments with $\text{IGD}_{\text{MEOF-}T}$ with different sample sizes $T = \{10, 50, 100, 1000\}$. The experimental settings were the same as in Section 3.1. Table 3 shows the results of $\text{IGD}_{\text{MEOF-}T}$ values with different sample sizes, where the best value for each problem instance is highlighted. Similar to the previous section, the results of the Wilcoxon rank sum test were summarized as follows; the three symbols, "$+$", "$-$", and "$\approx$" indicate that NSGA-II is statistically better than, worse than or competitive with NSGA-II-DTI, respectively. Table 4 shows the mean squared error (MSE) between the estimated values $\text{IGD}_{\text{MEOF-}T}$ and their true IGD_{MEOF} values. Moreover, Table 5 summarizes the average runtime over 21 trials required to complete one trial using $\text{IGD}_{\text{MEOF-}T}$ with different sample sizes.

From the table, $\text{IGD}_{\text{MEOF-}T}$ with $T = 10$ incorrectly evaluated as NSGA-II performed better than NSGA-II-DTI for RTP1 with $D = 5$. This suggests that the sampling-based IGD_{MEOF} indicator with few samples may under/over estimate the algorithmic performances. As the value of T further increased to 50, 100, and 1000, this inconsistency was resolved, and the MSE values gradually improved. Moreover, the counts of "$-$" and "$+$" were become closer to those of the IGD_{MEOF} as the T values increased (see Table 2). Thus, our experimental

Table 4: The average MSE values between the estimated values $\text{IGD}_{\text{MEOF-}T}$ and their true IGD_{MEOF} values.

D	T			
	10	50	100	1000
5	4.342e−02	3.542e−03	3.050e−03	2.844e−03
10	2.440e−01	6.853e−03	4.055e−03	4.414e−03
20	2.540e+00	8.208e−02	2.276e−02	5.670e−03

Table 5: The average runtime [s] of $\text{IGD}_{\text{MEOF-}T}$ with $T = \{10, 50, 100, 1000\}$ on TP and RTP problems (300 generations, 21 trials).

D	T			
	10	50	100	1000
5	1.611	2.978	5.652	40.04
10	1.671	3.441	6.157	70.47
20	1.766	3.995	7.048	87.79

results confirm the obvious fact that larger samples sizes yield more accurate approximations for the true IGD_{MEOF} values; however, a drawback of using larger samples sizes is to require more computational time (see Table 5). Despite this obvious trend, our experimental results suggest that IGD_{MEOF} with the sample size of 50 may be acceptable to briefly identify the superiority of algorithms while reducing the computational time.

5. CONCLUSION

This paper conducted a comparative study of the performance indicators for robust multi-objective optimization. We compared the two performance indicators, IGD_{mean} and IGD_{MEOF}, to investigate which indicator provides a more reasonable assessment of algorithms in the context of robust multi-objective optimization. Experimental results showed that the IGD_{mean} indicator may mislead to inaccurate performance assessment owing to its strong dependency of perturbations and the nature of the IGD calculation. In contrast, the IGD_{MEOF} indica-

tor provides a more reasonable performance assessment, consistent with the ground truth comparison results. We further showed that the sampling-based IGD_{MEOF} indicator with at least 50 samples can serve as an alternative to the original IGD_{MEOF}, thereby expanding its applicability to black-box problems.

Based on these observations, we recommend using the IGD_{MEOF} indicator or its sampling-based variant to more appropriately assess algorithm performance in robust multi-objective optimization. We hope that our findings will contribute to further advancements in the field.

ACKNOWLEDGEMENT

The presented work was supported by Support Center for Advanced Telecommunications Technology Research.

REFERENCES

[1] Johan Andersson. A survey of multiobjective optimization in engineering design. *Department of Mechanical Engineering, Linktjping University. Sweden*, page 38, 2000.

[2] Zhenan He, Gary G Yen, and Zhang Yi. Robust Multiobjective Optimization via Evolutionary Algorithms. *IEEE Transactions on Evolutionary Computation*, 23(2):316–330, 2018.

[3] Kalyanmoy Deb and Himanshu Gupta. Introducing Robustness in Multi-Objective Optimization. *Evolutionary computation*, 14(4):463–494, 2006.

[4] CA Coello Coello. Evolutionary Multi-Objective Optimization: A Historical View of the Field. *IEEE computational intelligence magazine*, 1(1):28–36, 2006.

[5] Shubhkirti Sharma and Vijay Kumar. A Comprehensive Review on Multi-objective Optimization Techniques: Past, Present and Future. *Archives of Computational Methods in Engineering*, 29(7):5605–5633, 2022.

[6] Lie Meng Pang, Hisao Ishibuchi, Yang Nan, and Cheng Gong. Reliability of Indicator-Based Comparison Results of Evolutionary Multi-objective Algorithms. In *International Conference on Parallel Problem Solving from Nature*, pages 285–298. Springer, 2024.

[7] Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and TAMT Meyarivan. A Fast and Elitist Multiobjective Genetic Algorithm: NSGA-II. *IEEE transactions on evolutionary computation*, 6(2):182–197, 2002.

[8] Qingfu Zhang and Hui Li. MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition. *IEEE Transactions on evolutionary computation*, 11(6):712–731, 2007.

[9] Eckart Zitzler and Simon Künzli. Indicator-Based Selection in Multiobjective Search. In *International conference on parallel problem solving from nature*, pages 832–842. Springer, 2004.

[10] Hisao Ishibuchi, Lie Meng Pang, and Ke Shang. Difficulties in Fair Performance Comparison of Multi-Objective Evolutionary Algorithms. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pages 937–957, 2022.

[11] Eckart Zitzler and Lothar Thiele. Multiobjective Evolutionary Algorithms: A Comparative Case Study and the Strength Pareto Approach. *IEEE transactions on Evolutionary Computation*, 3(4):257–271, 1999.

[12] Carlos A Coello Coello and Margarita Reyes Sierra. A Study of the Parallelization of a Coevolutionary Multi-Objective Evolutionary Algorithm. In *MICAI 2004: Advances in Artificial Intelligence: Third Mexican International Conference on Artificial Intelligence, Mexico City, Mexico, April 26-30, 2004. Proceedings 3*, pages 688–697. Springer, 2004.

[13] Hisao Ishibuchi, Ryo Imada, Naoki Masuyama, and Yusuke Nojima. Comparison of Hypervolume, IGD and IGD^+ from the Viewpoint of Optimal Distributions of Solutions. In *Evolutionary Multi-Criterion Optimization: 10th International Conference, EMO 2019, East Lansing, MI, USA, March 10-13, 2019, Proceedings 10*, pages 332–345. Springer, 2019.

[14] Wenxiang Jiang, Kai Gao, Shuwei Zhu, and Lihong Xu. A novel robust multi-objective evolutionary optimization algorithm based on surviving rate. *Complex & Intelligent Systems*, 11(4):1–25, 2025.

[15] Yuxiang Shui, Hui Li, Jianyong Sun, and Qingfu Zhang. Approximating robust Pareto fronts by the MEOF-based multiobjective evolutionary algorithm with two-level surrogate models. *Information Sciences*, 657:119946, 2024.

[16] Zhenan He, Gary G Yen, and Jiancheng Lv. Evolutionary Multi-objective Optimization with Robustness Enhancement. *IEEE Transactions on Evolutionary Computation*, 24(3):494–507, 2019.

[17] Ye Tian, Ran Cheng, Xingyi Zhang, and Yaochu Jin. PlatEMO: A MATLAB Platform for Evolutionary Multi-Objective Optimization [Educational Forum]. *IEEE Computational Intelligence Magazine*, 12(4):73–87, 2017.

[18] Eckart Zitzler, Kalyanmoy Deb, and Lothar Thiele. Comparison of Multiobjective Evolutionary Algorithms: Empirical Results. *Evolutionary computation*, 8(2):173–195, 2000.

[19] Hui Li and Qingfu Zhang. Multiobjective Optimization Problems With Complicated Pareto Sets, MOEA/D and NSGA-II. *IEEE transactions on evolutionary computation*, 13(2):284–302, 2008.