Running Performance Metrics for Evolutionary Multi-Objective Optimization

Kalyanmoy Deb and Sachin Jain
Kanpur Genetic Algorithms Laboratory (KanGAL)
Indian Institute of Technology Kanpur
Kanpur, PIN 208016, India
{deb,jsachin}@iitk.ac.in

KanGAL Report No. 2002004

Abstract

It is now well established that more than one performance metrics are necessary for evaluating a multi-objective evolutionary algorithm (MOEA). Although there exist a number of performance metrics in the MOEA literature, most of them are applied to the final non-dominated set obtained by an MOEA to evaluate its performance. In this paper, we suggest a couple of *running* metrics – one for measuring the convergence to a reference set and other for measuring the diversity in population members at every generation of an MOEA run. Either using a known Pareto-optimal front or an agglomeration of generation-wise populations, the suggested metrics reveal important insights and interesting dynamics of the working of an MOEA or help provide a comparative evaluation of two or more MOEAs.

1 Introduction

Over the past couple of years or so, more emphasis in the area of multi-objective evolutionary algorithms (MOEAs) has been laid on the development of performance metrics. The primary reason for its lukewarm interest in the past is that unlike in the single-objective EA studies, where the performance metric is directly related to the objective function (both being scalar quantities), in multi-objective optimization the performance metric must assess a number of solutions each having a vector of objective values. This immediately demands the need of more than one performance metrics. A recent study (Zitzler et al., 2002) has shown that for an M-objective optimization problem, at least M performance metrics must be used. In the context of single-objective EAs, the record of a running performance metric with generation provided a plethora of information about the working of the EA on the problem at hand. It is not surprising that if appropriate performance measures of MOEA populations are also recorded and analyzed, salient information about their working would be imminent. In this paper, we discuss and suggest potential running performance metrics for multi-objective optimization.

Although a number of different performance metrics have been suggested, many are only applicable to two-objective problems and most importantly it is not obvious which of these performance metrics one may use in practice. Knowles and Corne (2002) compared most of these metrics based on their compatibility with outperformance relations between two sets of solutions and their monotonicity properties. The metrics are classified based on strong, weak and complete outperformance of one set to another. When one set outperformes another, it is also important to evaluate the extent of outperformance, instead of just determining whether one outperformes the other or not. The study also suggested the use of any of the three R-metrics introduced in Hansen and Jaszkiewicz (1998). The R-metrics measure the proportion of occasions one approximation set will be preferred by a pre-defined set of utility functions. Although such a metric is a pragmatic one, they may not be computationally efficient to be used as a running performance metric.

It is intuitive that the use of a set of metrics less than the number of objectives would mean a lose of a dimension and would immediately make the approach theoretically inaccurate. However, one of the ways to overcome the dimensionality problem practically is to use a *functionally* independent set of variables (Goldberg, 1993). By realizing that there are two main functional goals of MOEAs, efforts can be made in devising two metrics: (i) one for measuring the convergence of solutions to the Pareto-optimal front and (ii) the other for measuring the diversity of solutions. Such a set of two metrics will enable two or more solution sets to be compared among each other in terms of their functional achievements. Although the first metric may simply be devised based on the distance from a reference set, the second metric is not that straightforward to explain, particularly in the case of large number of objectives.

In the remainder of this paper, we briefly review the existing performance metrics. Thereafter, we suggest a metric each for evaluating two functionalities described above. Finally, the need of using running metrics in MOEAs is amply demonstrated by applying them on a number of two and three-objective problems.

2 Existing Performance Metrics

Veldhuizen (1999), in his dissertation, reported a number of performance metrics for multi-objective optimization. Later, the first author, in his book (Deb, 2001), classified existing performance metrics into three classes: (i) metrics for convergence, (ii) metrics for diversity estimation, and (iii) metrics for both convergence and diversity. Although advantages and disadvantages of each metric were qualitatively mentioned, Knowles and Corne (2002) analyzed most of those metrics on the basis of extent of outperformance relations between two sets of non-dominated solutions. This study suggested the use of any of the three performance metrics of Hansen and Jaszkiewicz (1998). We describe them briefly here.

The essential idea is that for any two approximate sets (A and B), these metrics use a number of utility functions and determine the expected number of occasions the solutions of one set is better than the other. These metrics, in other words, declare that set as the winner which will be the choice of most decision-makers. For example, the metric $R1$ is defined as follows:

$$R1(A,B,U,p) = \int_{u\in U} C(A,B,u)p(u)du, \tag{1}$$

where U is the set of all utility functions considered, $p(u)$ is the frequency of occurrence of the utility function u, and the outcome function C is defined as follows:

$$C(A,B,u) = \begin{cases} 1, & \text{if } u^*(A) > u^*(B), \\ 0.5, & \text{if } u^*(A) = u^*(B), \\ 0, & \text{if } u^*(A) < u^*(B). \end{cases} \tag{2}$$

Here, $u^*(A)$ is the maximum value of the utility function u on approximation set A and is denoted by $u^*(A) = \max_{z \in A}\{u(z)\}$ (z is an objective vector from A). It is clear from equation 1 that the $R1$-metric measures the expected proportion of occasions the set A is better than set B. The developers suggested to declare A to be the winner if $R1(A, B, U, p) > 0.5$ and B to be the winner if $R1(A, B, U, p) < 0.5$. If $R1(A, B, U, p) = 0.5$, superiority of either A or B cannot be concluded from the study. However, the difficulties with this metric are the requirement of a number of utility functions, their probabilities of occurrence, and a numerical integration technique to actually evaluate the metric. To avoid the non-associativity problem involved in comparing more than two approximation sets, developers have also suggested comparing each set A with respect to a reference set R, as follows: $R1_R(A, U, p) = R1(A, R, U, p)$. In order to reduce the bias from the chosen reference set R, developers suggested using more than one reference sets to conclude the superiority of one set over the other. To quantify the extent of superiority of one set over the other, developers have suggested two more metrics:

$$R2(A, B, U, p) = \int_{u \in U} \left(u^*(A) - u^*(B)\right) p(u) du, \tag{3}$$

$$R3(A, B, U, p) = \int_{u \in U} \frac{u^*(A) - u^*(B)}{u^*(A)} p(u) du. \tag{4}$$

The $R2$-metric measures the expected degree of superiority and the the $R3$-metric measures the expected proportion of superiority. These two metrics can also be modified to be used with a reference set R, but need to be used as $R2_R(A, U, p) = R2(R, A, U, p)$ and $R3_R(A, U, p) = R3(R, A, U, p)$, respectively. As pointed out earlier, these metric values largely depend on the chosen utility functions. For two-objective problems with a convex Pareto-optimal region, the following utility function $u_k \in U$ (where $k \in [0, 1]$) can be chosen:

$$u_k = kf_1 + (1 - k)f_2. \tag{5}$$

In general, however, weighted Tchebycheff metrics can be used as the family of utility functions.

For measuring diversity and convergence of obtained solutions, Zitzler's hyper-volume metric (also known as the $\mathcal{S}$-metric) can also be used. The metric computes the hyper-volume of the objective space dominated by an approximation set. Although a set with a good diversity of solutions would mean a larger hyper-volume metric value (hence better), the $\mathcal{S}$-metric value depends on the chosen reference point used for the hyper-volume calculation and demands normalization of the objectives before computing the metric.

All these authors seem to have made one point clear: the comparison of two non-dominated set of solutions is not a straightforward matter, because of the dimensionality involved in the sets. It is also generally agreed that one performance metric is not enough to judge the performance of an MOEA. Recently, Zitzler et al. (2002) have given a formal proof stating that for an M-objective problem, at least M performance metrics are needed to compare two or more set of solutions. Intuitively this makes sense, as otherwise this would suggest an inaccurate judgment made with a reduction in dimensionality. However, we argue that although mathematically incorrect, it may be possible to compare two or more approximation sets functionally, as is often followed in understanding behaviors of complex systems. In multi-objective optimization, there are two primary functionalities that an MOEA must achieve: (i) approaching the Pareto-optimal front and (ii) maintenance of a diverse set of solutions. Thus, for a comparison based on the attainment of each of the two functionalities of multi-objective optimization, it may be possible to define two performance metrics, one for measuring each functionality *exclusively*, even for $M > 2$ objectives.

2.1 Need for Running Metrics

In most MOEA studies, including those involving performance metrics, two or more MOEAs have been usually compared purely on the basis of what have been obtained at the end of a simulation run. Either a carefully updated archive or an EA population at the end of a simulation run is evaluated for this purpose. However, interesting information about how an MOEA arrives at the final population has practically never been discussed in an MOEA study. On the contrary, most single objective EA studies analyze such a generation-wise performance measure showing how the average or best fitness or some other performance metric is varying with generation. Such an information provides a lot of insights to the working of an algorithm and helps decide which problems are difficult or easy for an EA. The main reason for their absence in the MOEA literature is the cardinality of the necessary performance metrics to properly evaluate an MOEA and the complexities involved in computing them.

However, some recent studies (Abbass, 2002; Farhang-Mehr and Azarm, 2002; Lu and Yen, 2002) have just begun to show generation-wise dynamics of performance metrics. In this paper, we demonstrate the use of two such running performance metrics for MOEAs by applying them on a number of test problems and discuss some interesting outcomes.

3 Suggested Running Metrics

Particularly for designing running metrics, the following properties are worth considering:

1. The metric should take a value between zero and one in an absolute sense. Since the metric is to be compared generation-wise, an absolute scaling of a running metric between zero and one will allow to assess the change of the metric value from one generation to another.

2. The target (or desired) metric value (calculated for an ideally converged and diversified set of points) must be known.

3. The metric should provide a monotonic increase or decrease in its value, as the population gets improved or deteriorated slightly. This will also help in evaluating the extent of superiority of one approximation set with another.

4. The metric should be scalable to any number of objectives. Although this is not an absolutely necessary property, but if followed, it will certainly be convenient for evaluating scalability issues of MOEAs in terms of number of objectives.

5. The metric may be computationally inexpensive, although this is not a stringent condition to be followed.

Since two independent metrics are to be chosen, each for measuring one of the two functionalities of multi-objective optimization, each metric may preferably evaluate the corresponding functionality only, ignoring the other functionality. Since two functionalities can not be measured by one single metric, it is better to evaluate each functionality independently. On this account, some suggested metrics, such as the $D1_R$ metric (average distance of reference points from the approximation set) suggested by Czyzak and Jaszkiewicz (1998) for convergence measure or the $\mathcal{S}$-measure (used in Zitzler (1999)) for diversity measure cannot be adequately used. Because of computational expenses involved in computing the R-metrics, they may not be suitable candidates to be used as running metrics. In the following two subsections, we propose two different metrics for evaluating MOEAs by keeping in mind the above properties.

3.1 Metric for Convergence

We use a simple metric for evaluating convergence towards a reference set. A target set of points P^* (or reference set) can be either a set of Pareto-optimal points (if known) or the non-dominated set of points in a combined pool of all generation-wise populations obtained from an MOEA run. Thereafter, for a population $P^{(t)}$ at each generation, we compute the convergence metric in the following manner:

Step 1 Identify the non-dominated set $\mathcal{F}^{(t)}$ of $P^{(t)}$.

Step 2 From each point i in $\mathcal{F}^{(t)}$, calculate the smallest normalized Euclidean distance to P^* as follows:

$$d_i = \min_{j=1}^{|P^*|} \sqrt{\sum_{k=1}^{M} \left(\frac{f_k(i) - f_k(j)}{f_k^{\max} - f_k^{\min}} \right)^2}. \tag{6}$$

Here, $f_k^{\max}$ and $f_k^{\min}$ are the maximum and the minimum function values of k-th objective function in P^*.

Step 3 Calculate the convergence metric by averaging the normalized distance for all points in $\mathcal{F}^{(t)}$:

$$C(P^{(t)}) = \frac{\sum_{i=1}^{|\mathcal{F}^{(t)}|} d_i}{|\mathcal{F}^{(t)}|}. \tag{7}$$

In order to keep the convergence metric within $[0, 1]$, once the above metric values are calculated for all generations, we normalize the $C(P^{(t)})$ values by its maximum value (usually $C(P^{(0)})$): $\overline{C}(P^{(t)}) = C(P^{(t)})/C(P^{(0)})$.

3.2 Metric for Diversity

In terms of measuring the diversity of solutions, a recent study (Farhang-Mehr and Azarm, 2002) suggested an entropy-based technique. Each obtained solution is projected on a suitable hyper-plane[1]. Thereafter, a $(M-1)$-dimensional normally distributed entropy function is assigned with its mean being on each projected point and with a user-defined standard deviation. All such entropy functions for all projected points are added together and a normalized entropy function is calculated. If the projected points are well distributed on the hyper-plane and a suitable standard deviation of the entropy function is chosen, the resulting normalized entropy function will be a flat function, thereby causing a large value of the Shannon's entropy measure calculated using the

[1]A plane with a direction vector equal to the unit vector of the line joining the ideal point and the point with the worst individual objective values of the reference set is suggested in the original study.

normalized entropy function. On the other hand, if the resulting entropy function is peaky, meaning that points are crowded in some parts of the projected plane, the entropy measure will be small, thereby meaning a poor diversity among solutions. The idea of measuring diversity of solutions using a simulated entropy function is meaningful, but the approach has the following shortcomings:

1. The entropy measure largely depends on the chosen standard deviation, as the resulting distribution being peaky or flat depends largely on the variance of the normal entropy function used.

2. Since a continuous normal entropy function is used, if a problem involves disconnected Pareto-optimal fronts, this method may be erroneous.

3. For problems where the Pareto-optimal front is a degenerated lesser-dimensional curve (such as DTLZ5 (Deb et al., 2001)), a sub-optimal front may produce a higher entropy measure than a set of Pareto-optimal points, thereby allocating a large entropy value to an inferior set.

4. A performance metric measuring the diversity of solutions cannot adequately measure the converging ability of an MOEA, and vice versa.

The diversity metric suggested next is similar in concept to the above metric, except that it attempts to alleviate most of the above difficulties. Moreover, this and the convergence metric suggested earlier together can be used to systematically evaluate MOEAs for both convergence and diversity properties. The suggested metric is also computationally fast.

The essential idea is that the obtained non-dominated points at each generation is projected on a suitable hyper-plane, thereby losing a dimension of the points. The plane is divided into a number of small grids (or $(M-1)$ dimensional boxes). Depending on whether each grid contains an obtained non-dominated point or not, a diversity metric is defined. If all grids are represented with at least one point, the best possible (with respect to the chosen number of grids) diversity measure is achieved. If some grids are not represented by a non-dominated point, the diversity is poor. The parameters required from the user are the direction cosine of the reference plane, the number of grids (G_i) in each of $(M-1)$ dimension, and the target (or reference) set of points P^*. Here is the procedure:

Step 1 From $P^{(t)}$, determine the set $\mathcal{F}^{(t)}$ which are non-dominated to P^*.

Step 2 For each grid indexed by $(i, j, \ldots)$, calculate following two arrays:

$$H(i,j,\ldots) = \begin{cases} 1, & \text{if the grid has a representative point in } P^*; \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

$$h(i,j,\ldots) = \begin{cases} 1, & \text{if } H(i,j,\ldots)=1 \text{ and the grid has a representative point in } \mathcal{F}^{(t)}; \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Step 3 Assign a value $m(h(i,j,\ldots))$ to each grid depending on its and its neighbor's $h()$. Similarly, calculate $m(H(i,j,\ldots))$ using $H()$ for reference points.

Step 4 Calculate the diversity metric by averaging the individual $m()$ values for $h()$ with respect to that for $H()$:

$$D(P^{(t)}) = \frac{\sum_{\substack{i,j,\ldots \\ H(i,j,\ldots)\neq 0}} m(h(i,j,\ldots))}{\sum_{\substack{i,j,\ldots \\ H(i,j,\ldots)\neq 0}} m(H(i,j,\ldots))}. \tag{10}$$

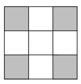
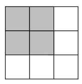
In the simple case, the *value function* $m()$ for a grid can be computed by using its $h()$ and two neighboring $h()$ dimension-wise. With a set of three consecutive binary $h()$ values, there are a total of 8 possibilities. Any value function may be assigned by keeping in mind the following:

- A `111` is the best distribution and a `000` is the worst.
- A `010` or a `101` means a periodic pattern with a good spread and may be valued more than a `110` or a `011`. For example, the above valuation will make an approximation set with 50% coverage of grids but having a wider spread (such as `1010101010`) better than another set having the same coverage but with a smaller spread (such as `1111100000`).
- A `110` or a `011` may be valued more than a `001` or a `100`, because of more covered grids.

Based on above observations, the following $m()$ values for $h()$ are suggested and used in this study:

$h(\ldots j-1\ldots)$	$h(\ldots j\ldots)$	$h(\ldots j+1\ldots)$	$m(h(\ldots j\ldots))$
0	0	0	0.00
0	0	1	0.50
1	0	0	0.50
0	1	1	0.67
1	1	0	0.67
0	1	0	0.75
1	0	1	0.75
1	1	1	1.00

Identical value are used for $H()$. In the current study, two or more dimensional hyper-planes are handled by calculating the above metric dimension-wise, whereas a higher-dimensional version of the above value function can also be carefully designed by considering a moving set of hyper-boxes. One such consideration for a two-dimensional set of 9 boxes is shown below:

 is better than 

Obviously, with more number of objectives, the value function will be difficult to define. It remains as an interesting future study to find if such higher-dimensional hyper-boxes are really necessary compared to the dimension-wise calculation of the proposed one-dimensional metric.

As an illustration to the above calculation procedure, Figure 1 shows a set of target points (marked as filled circles) and a set of population points (marked as shaded and open boxes) for a two-objective minimization problem. The points marked with shaded boxes are the non-dominated points with respect to the target points and are used for the diversity calculation (this is Step 1 of the procedure). The $f_2 = 0$ plane is used as the reference plane here and the complete range of f_1 values are divided into $G_1 = 10$ grids. In the next step, for each grid both $h()$ and $H()$ values are calculated. For the boundary grids (extreme grids and grids $(\ldots, j, \ldots)$ with $H(\ldots, j-1, \ldots) = 0$

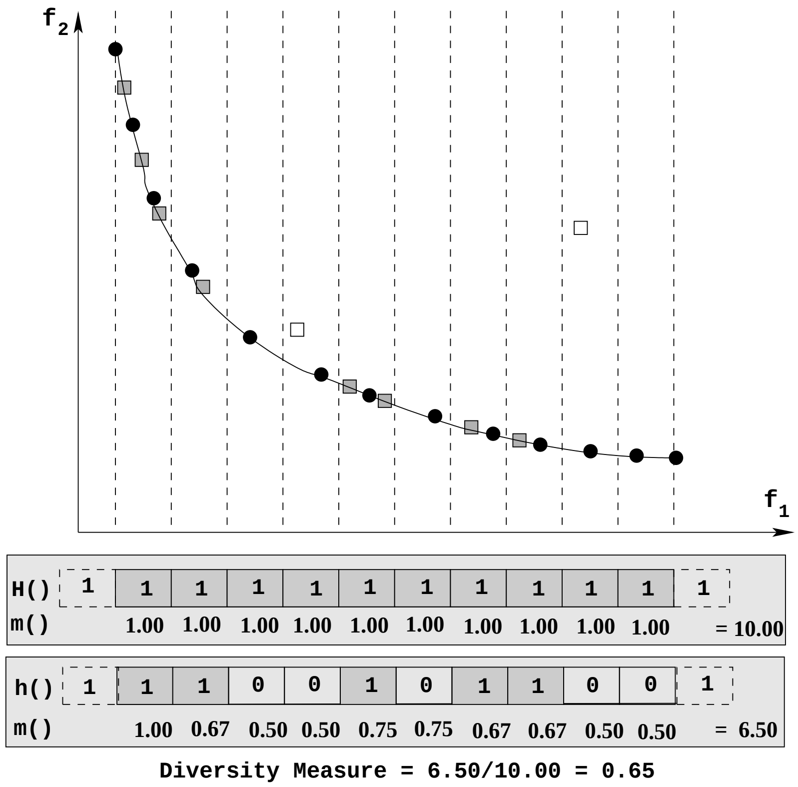


Figure 1: The diversity metric computation is illustrated.

or $H(\ldots, j+1, \ldots) = 0$), an imaginary neighboring grid with a $h()$ or $H()$ value of one is always assumed. In the figure, these grids are shown in dashed boxes. The $h()$ values are one in the first two grids, fifth, seventh and eighth grids. Notice that although more than one point may exist in a grid, the $H()$ or $h()$ value is still one for that grid. Based on a moving window containing three consecutive grids, the $m()$ values are computed in the figure and the diversity metric is calculated. To avoid the boundary effects (the effect of using the imaginary grids), we normalize the above metric as follows:

$$\overline{D}(P^{(t)}) = \frac{\sum_{\substack{i,j,\ldots \\ H(i,j,\ldots)\neq 0}} m(h(i,j,\ldots)) - \sum_{\substack{i,j,\ldots \\ H(i,j,\ldots)\neq 0}} m(\mathbf{0})}{\sum_{\substack{i,j,\ldots \\ H(i,j,\ldots)\neq 0}} m(H(i,j,\ldots)) - \sum_{\substack{i,j,\ldots \\ H(i,j,\ldots)\neq 0}} m(\mathbf{0})},$$

where $\mathbf{0}$ is a zero-valued array. A careful thought will reveal that the $H(i,j,\ldots) \neq 0$ consideration in computing the $\overline{D}(P^{(t)})$ term and the boundary grid adjustment suggested above allow a generic way to handle problems with disconnected Pareto-optimal fronts. The metric does not include the value function for a grid on which there exists no reference solution. We have included one such test problem (ZDT3) in our simulation studies.

If the Pareto-optimal front is not known (particularly to real-world problems), the target set may be determined in the following way. First, an MOEA is run for T generations and the generation-wise populations ($P^{(t)}$, $t = 0, 1, \ldots, T$) are stored. Thereafter, the non-dominated members $\mathcal{F}^{(t)}$ of each population are combined together and the target set is defined as the non-dominated set of the combined population: $P^* = \text{Non-dominated}(\cup_{t=0}^{T} \mathcal{F}^{(t)})$.

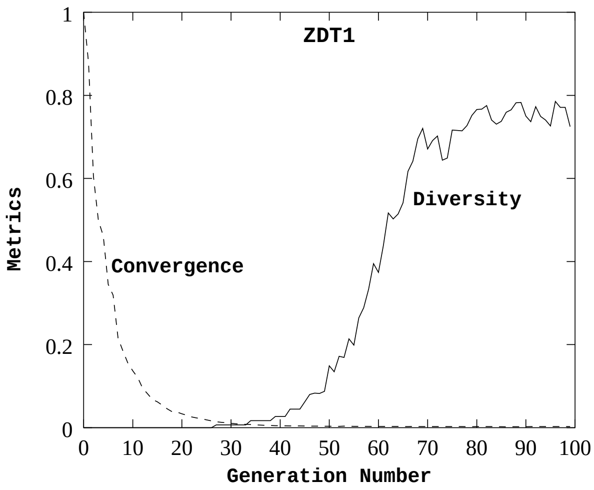


Figure 2: Two running metrics for ZDT1 using a set of Pareto-optimal points.

4 Simulation Results

In this section, we apply the above metrics on solutions obtained using NSGA-II (Deb et al., 2000) on a number of test problems suggested in the literature and on a real-world two-objective gearbox design problem. In each case, the metrics are applied on a typical run of NSGA-II. However, it is advised to apply the metrics on multiple runs of an MOEA and to report average metric values. In all cases with NSGA-II, we have used the SBX recombination operator (Deb, 2001) with $p_c = 1.0$ and $\eta_c = 10$, and the polynomial mutation operator with $p_m = 1/n$ and $\eta_m = 20$.

First four test problems are two-objective test problems chosen from Zitzler, Deb and Thiele (2000). Since they are test problems and the exact location of the Pareto-optimal front is known in each case, we apply the suggested metrics using a set of 200 uniformly set of points on the Pareto-optimal front (in the objective space).

Figure 2 shows the two metrics on simulation results obtained using NSGA-II using $N = 100$ population size running till 100 generations. In this and other problems of this study, we set number of grids in each dimension as $G_i = N^{1/(M-1)}$. Thus, for two-objective problems, we have used the number of grids equal to the population size, as if each population member is expected on each grid. We also use $f_2 = 0$ plane to project the points, unless otherwise stated. The figure shows that the convergence metric quickly moves to zero, thereby implying that NSGA-II solutions starting from a random set of solutions quickly approach the Pareto-optimal front. A value of zero of the convergence metric implies that all non-dominated solutions match the chosen Pareto-optimal points. After about 30 generations, the NSGA-II population comes very close to the Pareto-optimal front. Similarly, the diversity metric shows that till about the first 26 generations, no solution non-dominated to the Pareto-optimal set was found. But from this generation onwards, NSGA-II finds more and more points non-dominated with the chosen Pareto-optimal points. The choice of points are such that the diversity metric increases exponentially till about 70 generations, after which the diversity remains more or less the same. Although the obtained points are very close the chosen Pareto-optimal points, diversity metric oscillates near a stable value, a matter which is inherent to

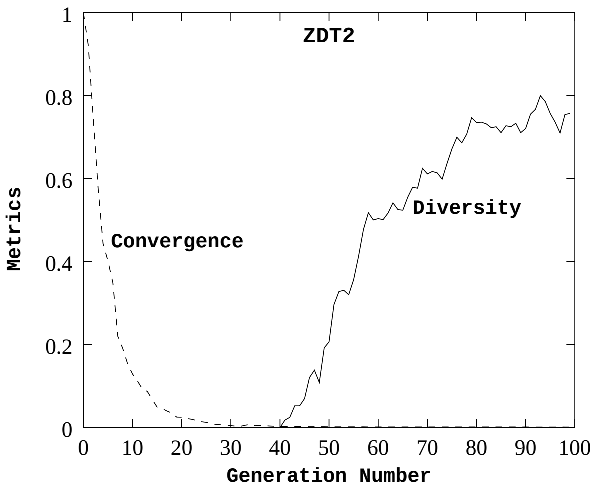


Figure 3: Two running metrics for ZDT2 using a set of Pareto-optimal points.

NSGA-II and other MOEAs and was also discussed for NSGA-II and SPEA2 in an earlier study (Laumanns et al., 2001).

It is interesting to note that the diversity measure suggested here uses only solutions which are non-dominated with the reference set. This may mean to suggest that convergence properties are somehow in-built into this diversity measure, but this is not quite true. Notice how the diversity measure is consistently equal to zero till about 26 generations, until which time no solutions non-dominated to the reference set is found. On the other hand, the convergence metric during these early generations shows how the population has been consistently approaching the reference set. Thus, a combination of these two metrics provide a clear insight into the working of an MOEA.

Next, we choose the test problem ZDT2. This problem has a non-convex Pareto-optimal front. Figure 3 shows the two running metrics with the same GA parameters as before. The convergence metric shows in about 30 generations, NSGA-II population moves very close to the Pareto-optimal front. It is also interesting to note that NSGA-II takes about 10 more generations to find a solution non-dominated with the reference set in ZDT2 than that required in ZDT1. Once some solutions very close to the Pareto-optimal front are found, more and more points near the front are discovered and the diversity among solutions increases rapidly, indicating that NSGA-II finds solutions close to the Pareto-optimal front with a good diversity among them.

The test problem ZDT3 has a disjointed set of Pareto-optimal fronts. NSGA-II with identical parameter setting are applied to this problem for 100 generations and the corresponding running metrics are shown in Figure 4. The figure shows a similar behavior as that in ZDT1. In this problem, we have also computed the diversity metric by projecting the points on an inclined line, as suggested in the footnote in section 3.2. The corresponding diversity metric is plotted with a solid line. The figure shows that although there is some difference in magnitude of the two diversity metric computations (projected on $f_2 = 0$ plane and projected on the inclined line), their behaviors are very similar.

Next, we choose the test problem ZDT6, which has a non-uniform density of solutions along the Pareto-optimal front. This time, the NSGA-II is run for 300 generations. Figure 5 shows that

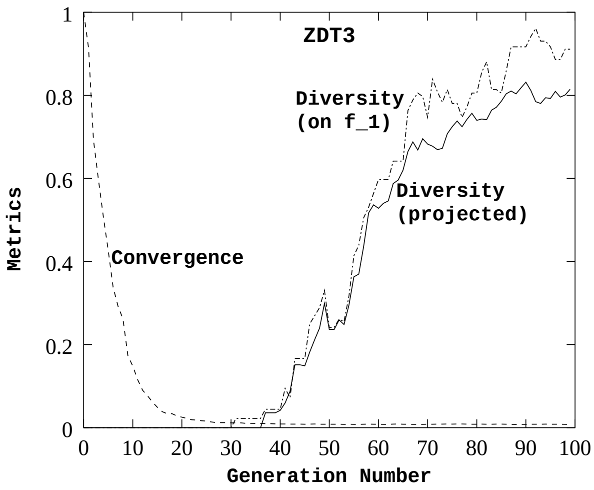


Figure 4: Two running metrics for ZDT3 using a set of Pareto-optimal points calculated using f_1 axis and projected on an inclined plane.

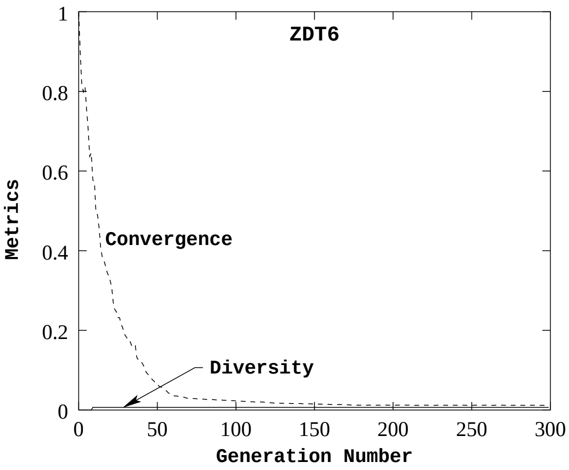


Figure 5: Two running metrics for ZDT6 using a set of Pareto-optimal points.

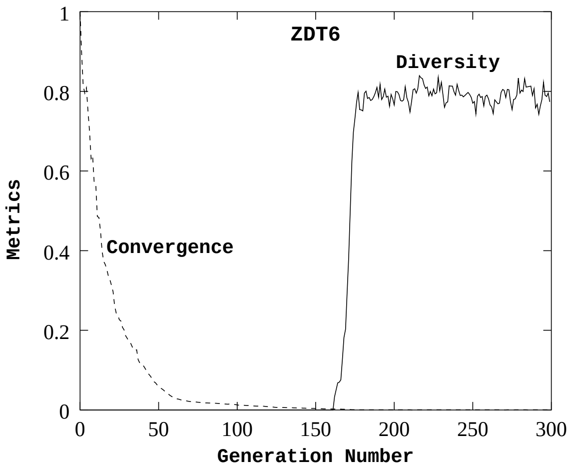


Figure 6: Two running metrics for ZDT6 using an agglomeration of generation-wise populations.

the population steadily approaches the Pareto-optimal front, but always maintains a finite distance from the true Pareto-optimal front. Since almost all obtained solutions are not non-dominated with the chosen Pareto-optimal points, the diversity metric is calculated with only a few solutions, thereby having a very small value. For problems which are difficult to solve to Pareto-optimality, we may use an agglomeration of populations (discussed in the last paragraph of section 3.2), instead of Pareto-optimal points, to compute the above metrics. Figure 6 shows that NSGA-II population reaches its final non-dominated front after around 165 generations. After this generation, the population quickly converges to the final non-dominated front and the population diversity resorts into a noisy steady-state value. This study shows that when it is difficult to obtain solutions on

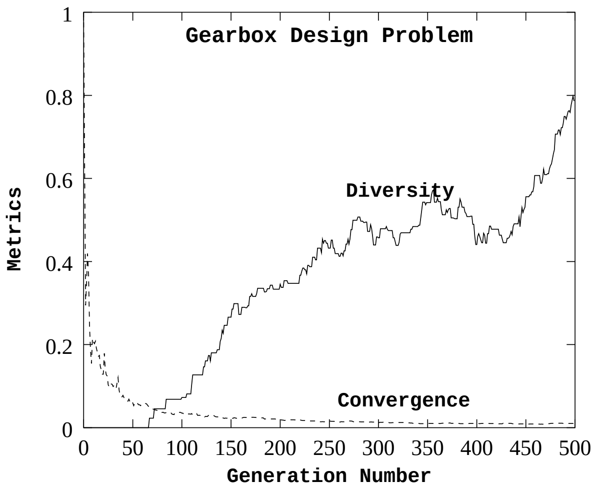


Figure 7: Two running metrics showing the convergence and diversity in solutions of NSGA-II.

the true Pareto-optimal front, it is better to use the population-agglomeration technique. This technique can also be used for comparing two or more MOEAs.

Next, we use the population-agglomeration technique to evaluate the performance of NSGA-II to a two-objective gearbox design problem discussed elsewhere (Deb and Jain, 2002). In this problem, the Pareto-optimal front is not known. With 100 grid points on f_1, the metrics are plotted in Figure 7 for NSGA-II run using 100 population members. The figure shows that the convergence near the obtained non-dominated front is fast and the first point near this front was found after about 65 generations. It is interesting to note that at the final generation the diversity metric value is less than one. This means that NSGA-II does not contain all 100 solutions used as a reference set at the end of the simulation run. Some better distributed solutions are lost during the run.

The final two test problems are three-objective test problems borrowed from Deb et al. (2001). The problem DTLZ2 has a spherical Pareto-optimal surface. Here, we use 100 population members and NSGA-II and SPEA2 (Zitzler, Laumanns, and Thiele, 2001) are run for 300 generations. For computing the metrics, we have chosen 10×10 or 100 grids in the interval $[0,1]$ of f_1 and f_2 axes. The $f_3 = 0$ plane is chosen to project the points. Figure 8 shows the corresponding metrics, as they vary with the generation number. The convergence metric values for both MOEAs suggest that not all non-dominated points even at the final generation have reached close to the chosen Pareto-optimal points. The diversity metric values show that both NSGA-II and SPEA2 have found a solution non-dominated to the chosen Pareto-optimal points at around 20th generation. Thereafter, more and more points with increasing diversity among them have been found. Thereafter, like in ZDT problems, the population maintains a particular level of diversity with some variations. However, the diversity of solutions obtained by SPEA2 is much better than that of NSGA-II in this problem. In order to investigate the diversity obtained in each case, we have plotted the final obtained non-dominated solutions in Figures 9 and 10 for NSGA-II and SPEA2, respectively. It is clear from these two plots that SPEA2 has obtained a better spread than NSGA-II. The clustering operator of SPEA2 is more computationally expensive than the crowding operator used in NSGA-II, but SPEA2 is better able to distribute points. Figure 8 reveals the latter fact quantitatively by showing

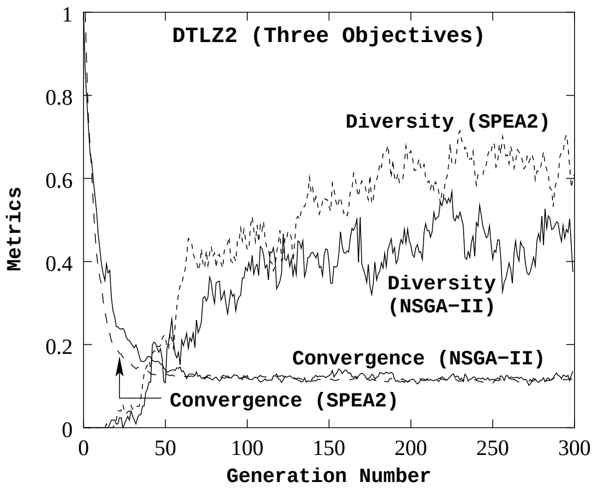


Figure 8: Two running metrics for DTLZ2 using a set of Pareto-optimal points for NSGA-II and SPEA2.

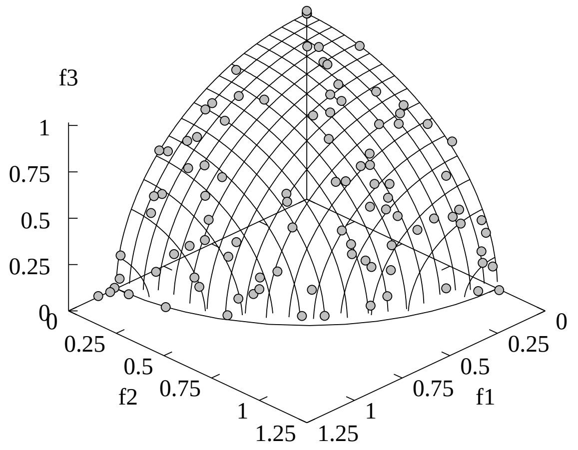


Figure 9: Non-dominated points obtained using NSGA-II.

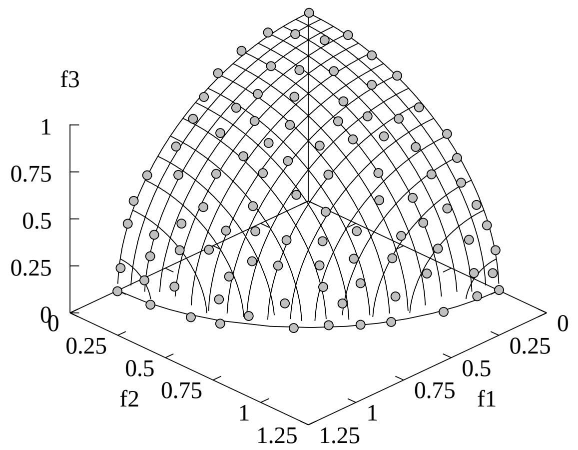


Figure 10: Non-dominated points obtained using SPEA2.

SPEA2 having a better diversity metric value than NSGA-II not only at the final generation, but in the entire run of the two algorithms; however the convergence levels of both algorithms are more or less the same. The fluctuations in the diversity metric by both methods arise due to the dynamic loss and gain of critical points on the non-dominated solutions, once they come close to the reference set. In any case, this study shows how the two suggested metrics can bring out the differences in the working of two MOEAs quantitatively over the entire run of the algorithms.

Our next problem is DTLZ5, which has a Pareto-optimal curve, instead of a three-dimensional Pareto-optimal surface. Here, we run NSGA-II for 200 generations, whereas all other NSGA-II parameters and parameters for computing the metrics are the same as in DTLZ2. Figure 11 shows the convergence and diversity metrics of a single run of NSGA-II. The convergence metric depicts a good convergence of NSGA-II on the Pareto-optimal front. Since the number of grids involving the

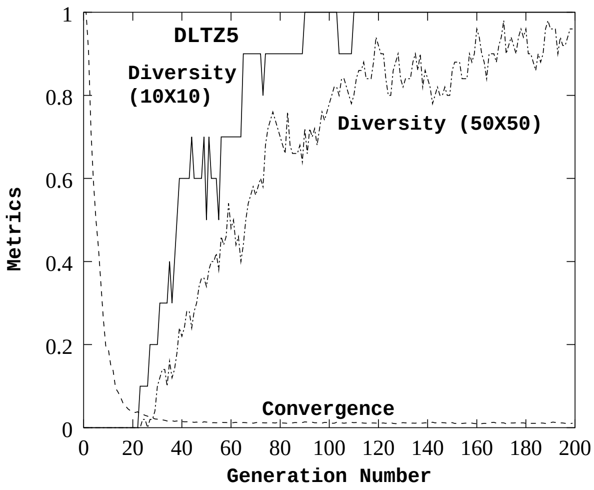


Figure 11: Two running metrics for DTLZ5 using a set of Pareto-optimal points.

Pareto-optimal front is few in this problem compared to the total number of grids being considered in the diversity computation (due to the reduced dimensionality of the Pareto-optimal front), the diversity metric changes in finite steps. The attainment of a diversity metric value of unity suggests the perfect diversity being achieved with 10 grids in each objective. When 50 grids on each objective (f_1 and f_2) axis are chosen, the Pareto-optimal front is represented by more grids. The diversity metric for this case is also shown in the figure. A hyper-plane with more grids require more obtained solutions to attain a good diversity measure. It is not surprising that the diversity metric values degrade in this case from before. This illustrates the sensitivity of the suggested diversity metric on the chosen number of grids.

5 Discussions and Future Studies

The need of developing good running metrics and their use in MOEA studies is now clear from the above computer simulations. Although two metrics are used in this study, more metrics measuring further detailed properties of an approximation set can also be used and are a matter of immediate research interests.

Although the suggested template-based diversity metric is a way to measure the extent of diversity in an approximation set, there are at least a couple of difficulties with this metric: (i) the metric value depends on the chosen template (the value function $m()$) and (ii) as mentioned earlier it is difficult to assign a value function for higher dimensions, although a dimension-wise application has reasonably worked well in this paper. However, the overall strategy of projecting an approximation set to a suitable grid-ed hyper-plane and then applying a diversity metric remains a good approach for finding the diversity of an approximation set. Instead of the suggested template-based diversity metric, the following metrics can also be used:

1. A *grid-count* diversity metric counting the number of occupied grids in the projected plane can simply be used. To normalize the metric, the count can be divided by the total number

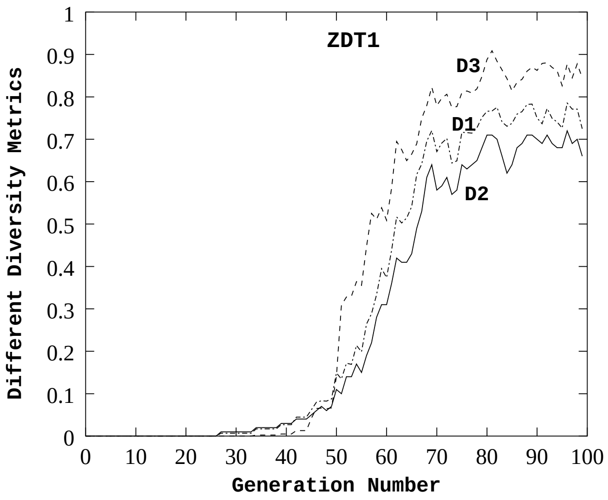


Figure 12: Three diversity running metrics on ZDT1 are shown. D1 is the metric defined in section 3.2, D2 is the grid-count metric and D3 is the variance-metric V.

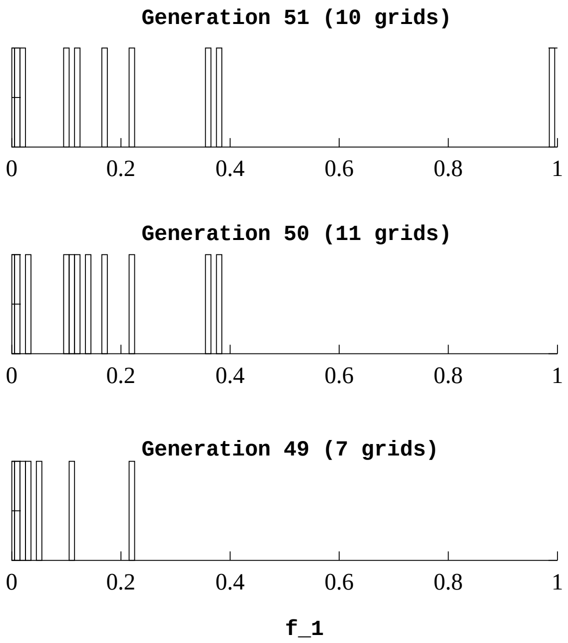


Figure 13: The grids occupied by NSGA-II at generations 49, 50, and 51 are shown for ZDT1.

of occupied grids obtained using the reference set. Although this metric does not tell the sparseness of occupied grids, the sheer number of them may be an important factor for comparing two or more MOEAs. Figure 12 plots this normalized grid-count metric (shown as D2) on the populations obtained using NSGA-II in problem ZDT1. The plot shows that NSGA-II solutions share about 70% grids with the reference set at the end of the run.

2. A variance in inter-member distances can be another measure of diversity of solutions on a projected plane. The calculation procedure is as follows. For a population $P^{(t)}$, first the non-dominated solutions $\mathcal{F}^{(t)}$ (with respect to P^*) are found and are projected to a chosen hyper-plane. Each occupied grid on the hyper-plane is represented by its center point $\vec{x}^{(i)}$. First, the centroid $\vec{g}$ of all occupied grids is calculated. Thereafter, the vector $\vec{v}^{(i)}$ from the grid $\vec{x}^{(i)}$ to the centroid $\vec{g}$ is computed. Then, a diversity metric V is calculated as follows:

$$V(P^{(t)}) = \sum_{j=1}^{M-1} \sqrt{\sum_{i=1}^{|\mathcal{F}^{(t)}|} \left(v_j^{(i)}\right)^2}. \tag{11}$$

For each dimension j, the term inside the square-root is the sum of the square of the projections of all solutions from the centroid along the j-th dimension. Thus, if the solutions are sparsely placed, the above metric will produce a large value. Since the projections are squared and added, the positive and negative projections do not cancel each other.

In order to normalize the metric, the above value can be divided by the metric value of the reference set: $\overline{V} = V(P^{(t)})/V(P^*)$. Figure 12 also shows a plot of this metric (shown as D3) on ZDT1. The comparison of these three plots shows the similarity in their variations with generations. For ZDT1, it is clear that the undulations in the diversity metrics D1 or D3 are in tune with the undulations in the number of occupied grids (D2). An interesting difference among these three metrics is clear during the transition of the population from generation 49 through 51:

Generation number	Number of occupied grids	Diversity Metrics		
		D1	D2	D3
49	7	8.754/100	7/100	0.192/2.887
50	11	14.899/100	11/100	0.405/2.887
51	10	13.468/100	10/100	0.891/2.887

Figure 13 also shows the occupied grids at generations 49, 50, and 51. A total of 100 grids are assumed for $f_1 \in [0, 1]$. The grid-count (D2) metric for these generations does not reveal the fact that a better diversity of solutions is obtained in generation 51, although the grid-count is less in generation 51 compared to that in generation 50. Because of the use of the three-grid template in D1, the difference in diversity obtained from generation 50 to 51 is not apparent by this metric either. But the D3 metric shows a big jump in this metric value during this transition, meaning that a set with a much better diversity among them is obtained. However, the real difference among these metrics may be apparent in solving higher-dimensional or more complex problems and the choice of one metric over the other may also depend on the computational complexity associated with each metric.

3. The entropy-based metric suggested by Farhang-Mehr and Azarm (2002) can be modified to handle disconnected Pareto-optimal sets. The *holes* or discontinuities in the projected plane can be eliminated and all individual regions of Pareto-optimal sets can be placed next to each other, so that in the modified plane, there is no hole or discontinuity. Although the choice of the standard deviation of the normal distribution still remains as an important factor, other distributions may also be tried.

A detail study applying these metrics to the chosen test problems and comparing them with existing metrics such as the R-metrics or the $\mathcal{S}$-metric also remains an important future task.

6 Conclusions

In order to investigate the generation-wise dynamics of an MOEA, two metrics are suggested in this paper. The first metric is a distance measure of a population from a reference set and is used to measure the convergence ability of an MOEA. The second metric uses a local template based evaluation technique to estimate the diversity of one set compared to a reference set. Both metrics are normalized so as to have their values bounded between zero and one. On six test problems and one real-world engineering design problem, the application of these two metrics have shown interesting properties of NSGA-II. The non-dominated solutions continually approach the reference set in all cases, meaning thereby that the domination-based selection operator of NSGA-II is capable of bringing the population near the reference set without getting stuck in any intermediate level. Once some points near the reference set are discovered, the search operators and the diversity-preserving mechanism of NSGA-II are able to find more and more solutions which are non-dominated to the reference set and which have increasingly large diversity among them. In one test problem, NSGA-II and SPEA2 populations are compared generation-wise, revealing superior performance of SPEA2 in maintaining a better diversity among non-dominated solutions.

For problems with unknown Pareto-optimal front, this study has also suggested a way to construct a reference set based on an agglomeration of generation-wise non-dominated populations. Such a technique can also be used to compute other performance metrics requiring a reference set.

It is also important to highlight that the computational complexity of MOEAs is another important matter and must be taken into consideration while comparing two or more MOEAs. Nevertheless, the results of this study has shown the importance of using running metrics for multi-objective evolutionary computation and recommend more such studies in the near future.

Acknowledgement

Authors thank Marco Laumanns for sharing the SPEA2 results on test problem DTLZ2.

References

Abbass, H. A. (2002). The self-adaptive Pareto differential evolution algorithm. *Proceedings of the World Congress on Computational Intelligence*, pp. 831–836.

Czyzak, P. and Jaszkiewicz, A. (1998). Pareto simulated annealing – a metaheuristic for multiobjective combinatorial optimization. *Multi-Criteria Decision Analysis, 7*, 34–47.

Deb, K., Thiele, L., Laumanns, M., and Zitzler, E. (2001). Scalable test problems for evolutionary multi-objective optimization. TIK Report No. 112. Zürich, Switzerland: Computer Engineering and Networks Lab (TIK), Swiss Federal Institute of Technology (ETH).

Deb, K. (2001). *Multi-objective optimization using evolutionary algorithms.* Chichester, UK: Wiley.

Deb, K., Agarwal, S., Pratap, A., and Meyarivan, T. (2000). A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization.: NSGA-II. *Proceedings of the Parallel Problem Solving from Nature*, pp. 849–858. Springer.

Deb, K. and Jain, S. (2002). Multi-speed gearbox design using multi-objective evolutionary algorithms. KanGAL Report No. 2002001. Kanpur, India: Department of Mechanical Engineering, IIT Kanpur.

Farhang-Mehr, A. and Azarm, S. (2002). Diversity Assessment of Pareto-optimal solution sets: An entropy approach. *Proceedings of the World Congress on Computational Intelligence*, pp. 723–728.

Goldberg, D. E. (1993). Making genetic algorithms fly: A lesson from Wright brothers. *Advanced Technology for Developers, 2*, 1–8.

Hansen, M. P. and Jaszkiewicz, A. (1998). Evaluating the quality of approximations to the non-dominated set. Technical Report IMM-REP-1998-7. Institute of Mathematical Modeling, Technical University of Denmark.

Knowles, J. and Corne, D. (2002). On metrics for computing non-dominated sets. *Proceedings of the World Congress on Computational Intelligence*, pp. 711–716.

Laumanns, M., Thiele, L., Deb, K., and Zitzler, E. (2001). On the convergence and diversity preservation properties of multi-objective evolutionary algorithms. Technical Report 108. Zürich, Switzerland: Computer Engineering and Networks Lab (TIK), Swiss Federal Institute of Technology (ETH).

Lu, H, and Yen, G. (2002). Rank-density based multiobjective genetic search. *Proceedings of the World Congress on Computational Intelligence*, pp. 944–949.

Van Veldhuizen, D. A. (1999). *Multiobjective evolutionary algorithms: Classifications, analyses, and new innovations.* Doctoral dissertation. Dayton, Ohio: Air Force Institute of Technology.

Zitzler, E. (1999). *Evolutionary algorithms for multiobjective optimization: Methods and applications.* Doctoral dissertation. Zürich, Switzerland: Swiss Federal Institute of Technology (ETH) (Dissertation ETH No. 13398).

Zitzler, E., Deb, K., and Thiele, L. (2000). Comparison of multiobjective evolutionary algorithms: Empirical results. *Evolutionary Computation Journal, 8*(2), 125–148.

Zitzler, E., Laumanns, M., Thiele, L., Fonseca, C., and Fonseca, V. G. (2002). Why quality assessment of multiobjective optimizers is difficult? Unpublished document.

Zitzler, E., Laumanns, M., Thiele, L. (2001) SPEA2: Improving the Strength Pareto Evolutionary Algorithm. TIK Report No. 103. Zürich, Switzerland: Computer Engineering and Networks Lab (TIK), Swiss Federal Institute of Technology (ETH).