On the Performance Assessment and Comparison of Stochastic Multiobjective Optimizers

Carlos M. Fonseca and Peter J. Fleming

Department of Automatic Control and Systems Engineering
The University of Sheffield
Mappin Street, Sheffield S1 3JD, U.K.
C.Fonseca@shef.ac.uk, P.Fleming@shef.ac.uk

Abstract. This work proposes a quantitative, non-parametric interpretation of *statistical* performance of stochastic multiobjective optimizers, including, but not limited to, genetic algorithms. It is shown that, according to this interpretation, typical performance can be defined in terms analogous to the notion of median for ordinal data, as can other measures analogous to other quantiles.
Non-parametric statistical test procedures are then shown to be useful in deciding the relative performance of different multiobjective optimizers on a given problem. Illustrative experimental results are provided to support the discussion.

1 Introduction

The growing interest devoted to evolutionary algorithms for multiobjective optimization is well reflected by the increasing number of different approaches being proposed in the literature (see [1] for a review). Unfortunately, and although the power and usefulness of such techniques is recognized, a well-established approach for the quantitative characterization of the performance of multiobjective evolutionary algorithms, which could, in turn, enable their comparison also in quantitative terms, is still lacking. Arguments for and against the methods currently available have thus remained largely based on qualitative aspects of the evolutionary process, e.g. on whether or not the population tends to cover "well" the, often unknown, trade-off surface of a particular problem.

This work proposes a quantitative, non-parametric interpretation of *statistical* performance of stochastic multiobjective optimizers based on the notion of goal attainment. It then discusses how, and to what extent, the performance of different multiobjective approaches *on a given problem* can be compared based on the proposed interpretation.

2 Performance of a Stochastic Optimizer

The notion of performance of an optimizer inevitably involves the quality of the solutions it is able to produce and the amount of computation effort it requires

to arrive at those solutions. In some cases, it is possible to actually guarantee that the optimum will be found, or at least approached at a certain rate, as long as the problem itself satisfies certain conditions. More generally, if there is ultimately no such guarantee, then the performance of the algorithm may still be evaluated experimentally on selected problems. Depending on the algorithm, it then may or may not be possible to correctly extrapolate the results obtained to other problems in the same class as that considered.

In the case of stochastic optimizers, such as evolutionary algorithms and simulated annealing, and generally, in that of any optimizer for which no performance guarantee exists under the given experimental conditions (e.g. limited computation time, multiple local optima), performance characterization is often attempted by repeated, independent experimentation.

In the remainder of this work "performance" will simply refer to the quality of the *final* solutions produced by the optimizer under study on a given problem, regardless of the termination criterion chosen. Evolutionary algorithms, for example, are often applied under the condition of limited computation time (typically measured in terms of elapsed time, CPU time, or number of function evaluations).

2.1 The Single-Objective Case

In the single-objective case, where results from multiple runs of an optimizer can at least be ordered according to their quality (as measured by the objective function which defines the problem), the distribution of the quality of such results can usually be represented easily by means of histograms and/or empirical distribution functions, and summarized by measures such as the median in combination with other statistics of the same type (e.g. upper and lower quartiles, best and worst result, other quantiles).

Moreover, a whole range of (non-parametric) statistical test procedures exist [2] which enable a comparison of the results produced by different optimization methods on a given problem, provided that each method can be independently applied to that problem a number of times. Typically, such procedures make only very weak assumptions about the (unknown) distributions the properties of which they attempt to compare.

If, additionally, the quality of a solution can be represented in terms of a numeric figure, then other measures of location and dispersion can also be used, namely the mean and the standard deviation. Both of these measures are, however, tied to the particular scale in which the objective function is expressed. In contrast, a simple genetic algorithm with rank-based fitness assignment, for example, or an evolution strategy, are insensitive to that scale. Even if the optimizer does depend on the scaling of the objective function, as does simulated annealing and GAs with proportional selection, mean and standard deviation alone would seem insufficient to summarize the performance of such algorithms. Note that the distribution of the results of many runs has unknown shape and cannot generally be expected to approximate a normal distribution (e.g. it is bounded below, assuming minimization, and thus asymmetric).

2.2 The Multiobjective Case

In the multiobjective case, the outcome of an optimization run will generally consist of a varying number of non-dominated solutions. Informally, one would like to obtain a "diverse" sampling of the trade-off surface which was, simultaneously, as close to the real trade-off surface as possible. A possible sampling of a given trade-off surface is illustrated in Fig. 1, for the simultaneous minimization of two objectives.

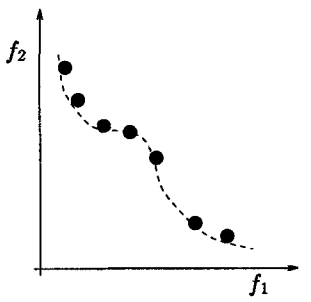

Fig. 1. Visualisation of non-dominated data in two dimensions. From [3].

Interpolating between the data to obtain a smooth representation of the trade-off surface is, however, not generally correct, firstly because there may be no guarantee that it will actually be smooth, and secondly because actual solutions corresponding to those intermediate objective vectors, even if they exist, are not known. At most, one could draw a boundary in objective space separating those points which are dominated by or equal to at least one of the data points, from those which no data point dominates or equals. Such a boundary (see Fig. 2) can also be seen as the locus of the family of tightest *goal vectors* which, given the data, are known to be attainable. It will be called an *attainment surface.*

An attainment surface actually combines information about the quality and the distribution of the corresponding individual non-dominated points across the trade-off surface. On the one hand, the closer to actually non-dominated the approximate solutions are, the closer to the real trade-off surface the attainment surface will be. On the other hand, any "holes" in the distribution of these solutions will have the opposite effect, i.e., will result in the corresponding region of the attainment surface being drawn away from the real trade-off surface, as Fig. 2 also illustrates. Thus, both types of information are expressed in terms of the *location* of the points which constitute the boundary, relative to the real trade-off surface. However, note that the way in which information is combined is independent of the scale in which objectives are expressed, as it relies solely on an ordinal relation, in this case, $\leq$.

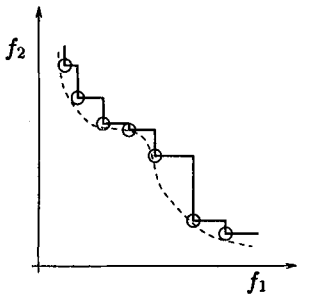

Fig. 2. The family of tightest goals known to be attainable as a result. From [3].

2.3 Multiple Runs

In order to gain some insight into how a certain multiobjective optimizer may typically perform on a given problem, multiple runs can of course be performed. Simply super-imposing non-dominated points obtained from various runs of a multiobjective optimizer does give an idea of how good individual points found in each run tend to be, but, unfortunately, information on how they tend to be distributed along the trade-off surface is largely lost, as Fig. 3 illustrates.

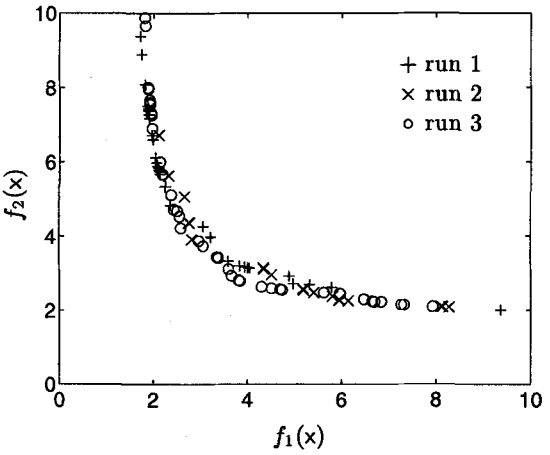

Fig. 3. The superposition of 3 sets of non-dominated points. From [3].

On the other hand, the superposition of the corresponding attainment surfaces preserves both kinds of information better. In Fig. 4, the uneven distribution of some of the sets of non-dominated solutions along the trade-off surface is more apparent. Nevertheless, for a large number of runs, even these plots rapidly become too dense and, therefore, difficult to interpret.

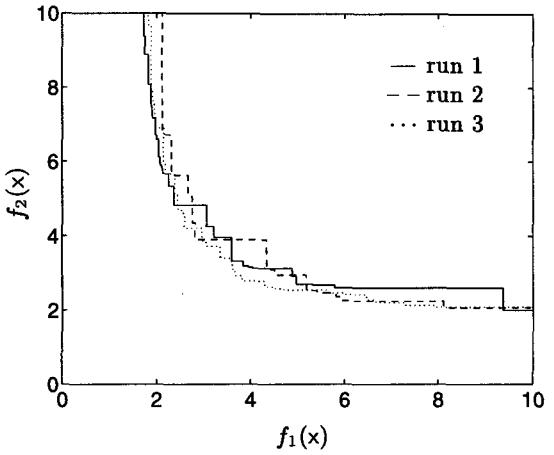

Fig. 4. The superposition of the 3 corresponding boundaries. From [3].

2.4 Statistical Interpretation

Plots such as that in Fig. 4 clearly divide the objective space into three main areas. The first area, located down and to the left of the attainment surfaces, is composed of all those objective (or goal) vectors which were never attained in any of the runs. As the number of runs increases, this area should approximate the real set of infeasible objective vectors better and better. The second area, located up and to the right of the attainment surfaces, is composed of those objective vectors which were attained in all of the runs. This may give some idea of worst-case performance for the algorithm, but this area must be expected to depend strongly on the number of runs considered. Finally, the third area, located within the boundaries, is composed of objective vectors which were attained in some of the runs, but not in others. This area can be divided further into sub-areas, according to the percentage of runs in which the corresponding objective vectors were attained.

This way of looking at the superposition of multiple attainment surfaces provides a basis on which a quantitative notion of "typical" performance can be built. In particular, one may now look to identify the family of goal vectors likely to be attained, *each on its own*, in exactly 50% of the runs (what will be called the 50%-attainment surface of the optimizer). In order to estimate this, consider the points of intersection of a set of attainment surfaces obtained experimentally with an *arbitrary* auxiliary straight line, diagonal to the axes and running in the direction of improvement in all objectives, as illustrated in Fig. 5.

These points can be seen to define a sample distribution which is essentially uni-dimensional and, thus, can be strictly ordered. The desired result can be obtained through the successive estimation of the median of all possible samples defined in this way. Note that this result is independent of the slope of the auxiliary lines used to produce it, as long as they remain diagonal to the axes

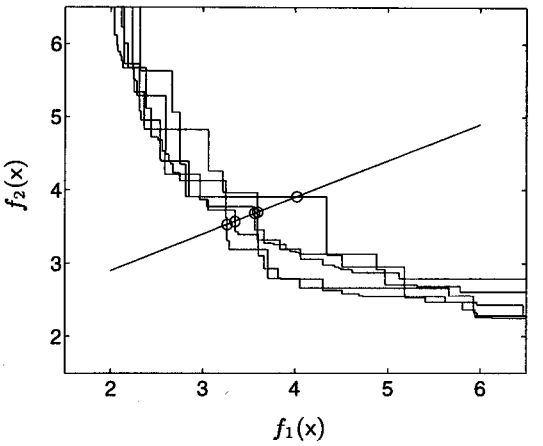

Fig. 5. The points of intersection with a line diagonal to the axes. From [3].

and running in the direction of improvement in all objectives.

Estimates for the 25% and 75% attainment surfaces could be produced exactly in the same way by estimating the lower and upper quartiles instead of the median. In Sect. 4, 25%, 50% and 75%-attainment surfaces estimated in this way from 21 independent runs of each of two different algorithms on a sample problem will be given, as well as the corresponding experimental lower and upper bounds.

3 Performance Comparison

The above considerations on how the performance of a multiobjective optimization approach may be described apply equally well when the performance of two or more different approaches is to be compared. In that case, the intersection of an auxiliary line with the experimental attainment surfaces will define not one, but two (or more) univariate-like samples, each sample corresponding to one approach. The properties of their underlying distributions can then be compared, one set of samples at a time, by means of standard non-parametric statistical test procedures. Provided that

1. the result of testing on a pair (or set) of samples can be associated with a point (or a region) in objective space, and that
2. the result associated with a point (or a region) in objective space is the same independently of the auxiliary line used to arrive at it,

it should be possible to identify those regions in objective space where a statistically significant difference between (the relevant aspects of) the localized performance of the methods under study can be found. As in the single-objective

case, the tests should not assume any similarity in shape of the distributions involved, since such similarities cannot in principle be expected.

Suitable test procedures include the median test, its extensions to other quantiles, and tests of the Kolmogorov-Smirnov type [2]. The result of the median test, for example, can be associated with the median of the combined samples (or grand median), as it only depends on how many observations from each sample there are to the left (or to the right) of the grand median. For each point on the grand 50%-attainment surface, this number is the same whatever the slope of the auxiliary line adopted to formulate the test.

On the other hand, the result of a test of the Kolmogorov-Smirnov type could be associated with the points in objective space where the Kolmogorov-Smirnov distance exceeds the critical value of the test statistic. Again, the result of the test would not depend on the slope of the auxiliary line adopted to formulate it.

4 Experimental Results

Two different multiobjective genetic algorithms were run for 100 generations, 21 times each, on the optimal regulator design problem described in [4]. Algorithm A did not include sharing or mating restriction, whereas Algorithm B included sharing and mating restriction in the phenotypic domain. In this case, the niche sizes were computed at each generation depending on the distribution of the population, as proposed in [5]. The results for each algorithm are presented in Figs. 6 and 7.

In both plots, the lower grey line indicates the best trade-off approximation known as a consequence of all the runs, while the upper grey line delimits the set of goal vectors attained in all the runs. The thick black line provides an estimate of the real 50%-attainment surface of each algorithm, and the thin black lines provide analogous estimates, but for 25% and 75%.

These plots can be seen to convey information about the location, dispersion, and even skewness of the distributions of the results in each region (considered in isolation) of the trade-off surface, much like the box plot used in statistics.

By looking at the 50%-attainment surface in each plot, Algorithm B would appear to produce better results more often than Algorithm A in the upper-left region of the trade-off surface. Differences in other regions of the trade-off surface are not clearly apparent. To assess in which regions of the trade-off surface the two methods may have significantly different 50%-attainment performance, the grand 50%-attainment surface may be computed, and a sufficiently large number of median tests performed along this surface, as suggested above. In practice, since the samples are finite, so is the number of *different* tests that may be performed.

Figure 8 presents test results, plotted on the grand 50%-attainment surface, for a 95% confidence level. The regions where no significant difference between the points of the 50%-attainment surfaces of the two algorithms could be found are indicated in light grey, whereas those regions where the points of the two surfaces were found to differ with the given confidence level are plotted in black.

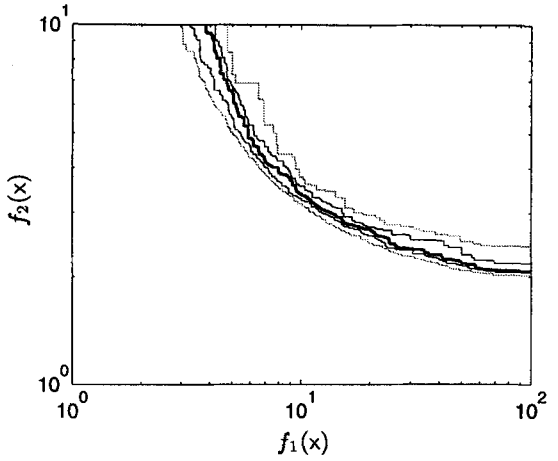

Fig. 6. Representation of the statistical performance of Algorithm A (21 runs).

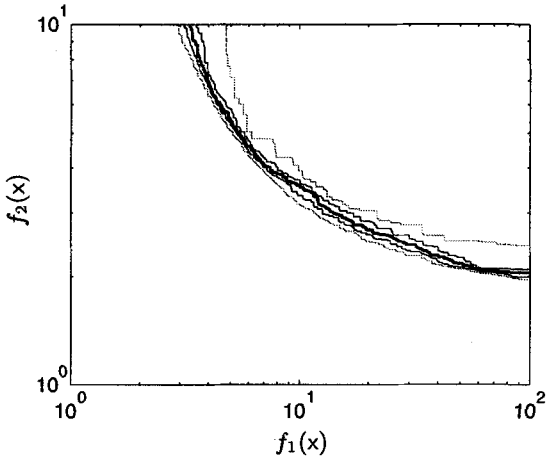

Fig. 7. Representation of the statistical performance of Algorithm B (21 runs).

The labels A and B indicate on which side of the grand 50%-attainment surface each algorithm has its own 50% attainment surface.

It is interesting to note that, although the addition of niche induction techniques in Algorithm B may be said to have improved performance in the extreme regions of the trade-off surface, this was at the expense of degradation in the middle region. This could be explained by the fact that forcing the population to spread more actually weakens its exploitative ability in those regions of the trade-off surface where it would otherwise be more concentrated. However, choosing a higher confidence level (99%) for the individual tests, this degradation is no longer statistically significant. The corresponding plot is shown in Fig. 9.

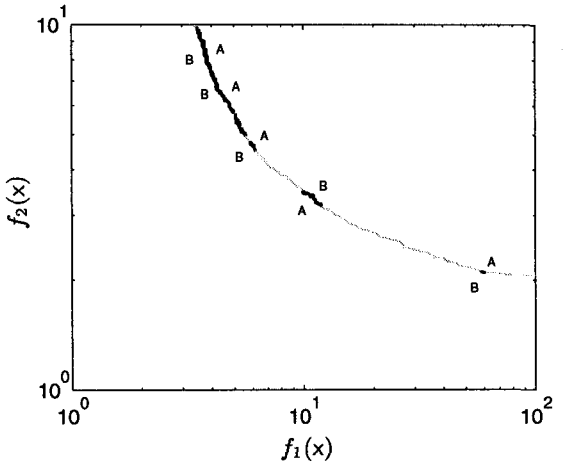

Fig. 8. Comparison of the 50%-attainment surfaces of algorithms A and B (95% confidence).

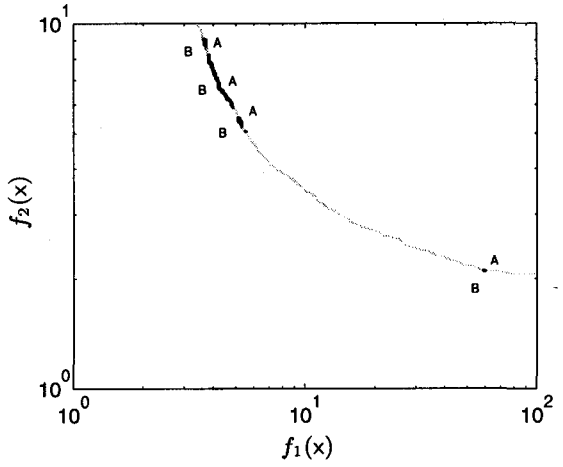

Fig. 9. Comparison of the 50%-attainment surfaces of algorithms A and B (99% confidence).

5 Concluding Remarks

The proposed interpretation of the statistical performance of a multiobjective optimizer in terms of the probability of the optimizer attaining arbitrary goals provides the grounds on which a much needed assessment and comparison of the performance of current and future multiobjective evolutionary approaches can be based. Rather than attempting such a comparison here, a simple example was given in order to illustrate the proposed methodology.

From the initial results presented, such a comparative study would most likely fail to elect an overall "best" algorithm, even for the same optimization problem. Instead, it might find a number of "non-dominated" approaches, in the same spirit of multiobjective optimization itself. On what basis a single, overall "best" approach could be selected, and with what significance level, are questions which are still difficult to answer, especially due to the multiple testing involved and to the likely dependence between tests performed close to each other in objective space.

Perhaps the main limitation of the present results is that the notion of probability of attainment applies only to individual points, separately. How to estimate similar surfaces for given probabilities of attaining the whole surface *simultaneously*, if at all possible, remains unclear.

Finally, although all examples given here involve two objectives only, the considerations made are clearly valid for any number of objectives. The development of software and visualisation tools to support more than two objectives is the subject of further work.

Acknowledgement The authors wish to acknowledge the support of the UK Engineering and Physical Sciences Research Council under Grant GR/J70857.

References

1. C. M. Fonseca and P. J. Fleming, "An overview of evolutionary algorithms in multiobjective optimization," *Evolutionary Computation*, vol. 3, pp. 1–16, Spring 1995.
2. W. J. Conover, ed., *Practical Nonparametric Statistics*. New York: Wiley, 1971.
3. C. M. Fonseca, *Multiobjective Genetic Algorithms with Application to Control Engineering Problems*. PhD thesis, University of Sheffield, 1995.
4. C. M. Fonseca and P. J. Fleming, "Multiobjective optimal controller design with genetic algorithms," in *Proc. IEE Control'94 International Conference*, vol. 1, (Warwick, U.K.), pp. 745–749, 1994.
5. C. M. Fonseca and P. J. Fleming, "Multiobjective genetic algorithms made easy: Selection, sharing and mating restriction," in *First IEE/IEEE International Conference on Genetic Algorithms in Engineering Systems: Innovations and Applications*, (Sheffield, UK), pp. 45–52, The Institution of Electrical Engineers, 1995.