5

MOEA Testing and Analysis

It doesn't matter how beautiful your theory is, it doesn't matter how smart you are. If it doesn't agree with experiment, it's wrong.

Richard P. Feynman

5.1 Introduction

Regarding the scientific method of experimentation, it is desirable to construct an accurate, reliable, consistent and non-arbitrary representation of multi-objective evolutionary algorithm (MOEA) architectures and performance over a variety of multi-objective optimization problems (MOPs). In particular, through the use of standard procedures and criteria, one should attempt to minimize the influence of bias or prejudice of the experimenter when testing a MOEA hypothesis. The design of each experiment must conform then to an accepted "standard" approach as reflected in any generic scientific method. When employing the scientific method, the detailed design of MOEA experiments can draw heavily from outlines presented by Barr et al. [93] and Jackson et al. [765]. These generic articles discuss computational experiment design for heuristic methods, providing guidelines for reporting results and ensuring their reproducibility. Specifically, they suggest that a well-designed experiment follows the following steps:

1. Define experimental goals;
2. Choose measures of performance - metrics;
3. Design and execute the experiment;
4. Analyze data and draw conclusions;
5. Report experimental results.

The scientific method as a more generic approach has four steps: observation[1], hypothesis, predict using hypothesis, and testing [1706, 911, 94]. Another very important experimental goal is determining how well the test problems and proposed metrics capture essential MOP and MOEA characteristics and performance. This chapter follows all these generic guideline concepts in developing experimental MOEA testing procedures. Such comparative experiments use appropriate MOP benchmarks or test suites as developed in Chapter 4.

The main goal of testing is usually to compare MOEA effectiveness over various chosen MOPs by measuring solution quality[2]. Once a meta-level testing process has been designed using the guidelines, specific MOPs and metrics must be selected. Observe that metrics usually fall into two performance categories: (1) **Efficiency** (measuring computational effort to obtain solutions, e.g., CPU time, number of evaluations/iterations - use of spatial and temporal resources), and (2) **Effectiveness** (measuring the *accuracy* and *convergence* of obtained solutions and the data interface to the environment). Effectiveness includes *Robustness* (measuring how well the code recovers from improper input), *Scalability* (measuring how large a class of problems the code can solve as related to the increasing problem dimension) and *Ease of use* (measuring the amount of effort required to use the software - user friendliness). Of course, there is always the trade-off between effectiveness (solution quality) and efficiency (execution time).

In order to study and analyze the dynamics of MOEA execution, generational population measurement of PF_{known} and P_{known} is necessary . In all cases, the associate measures provide qualitative date that is usually reduced to qualitative statements through metrics. Such metrics are usually based upon the concept of Pareto dominance as discussed in Chapter 1. Also, empirical stochastic distributions for effectiveness and efficiency can be evolved from these measurements using first-order (mean) and second-order (variance) and higher-order statistics. Using just the mean and variance values of course is explicitly assuming an underlying normal distribution. Note that the distribution form is characterized by estimating the cumulative distribution function.

Many MOEA researchers' *modus operandi* is an algorithm's comparison (possibly the researcher's own new and improved variant) against some other MOEA by analyzing results for specific MOP(s). Results are often "clearly" shown in visual graphical form indicating the new algorithm is more effective for selected MOPs. These empirical, relative experiments are of course incomplete as regarding robustness and general MOEA comparisons. The litera-

[1] Observation in this case is used as a foundation to model and approximate the real-world problem.

[2] Although MOEAs are classified as "stochastic multi-objective optimizers," the attainment of a MOP's true Pareto optimal front can be difficult and may even be impossible. Thus, the quantitative measurement of performance relies on PF_{known}, an attainment set by definition.