Reporting on Experiments in Evolutionary Computation

Mike Preuss

16. Januar 2007

Surprisingly, despite around 40 years of empirical tradition, in EC a standardized scheme for reporting experimental results never developed. The natural sciences, e.g. physics, possess such schemes as de-facto standards. Where does this difference originate from?

- As already stated, empiricism in the natural sciences has a very long tradition. Compared to computer science, the actual running time of experiments is rather long. Results are thus extremely valuable.
- In computer science, empiricism is a relatively recent phenomenon. Many important works date from the 1980s and 1990s (e.g. McGeoch [Cat86], Sacks et al. [SWMW89], and Barr et al. [BGK$^+$95]). In relation to the time needed to set up an experiment, its actual running time is often rather short. This entails a much more volatile character of the results.

Additionally, the impact of nondeterminism on the outcome of EC algorithms may have been underestimated in the past. The result of a comparison between two deterministic computer programs is sufficiently described by reporting its output values. As soon as any stochastic process is involved, much more data has to be provided and taken into account, including at least the recorded performance value samples or suitable statistics derived from them. For the sake of clarity and reproducibility, it is therefore the more important to properly report an experiments composition and outcome, the more it depends on subjective choices (e.g. of performance criteria) and the less exact its result are.

We argue that for scientific readers as well as for writers, a well-defined report structure is beneficial: As with the common overall publication structure (introduction, conclusions, etc.), a standard provides guidelines for readers, what to expect, and where. Writers are steadily reminded to describe the important details needed to understand and possibly replicate their experiments. They are also urged to separate the outcome of fairly objective observing from subjective reasoning. Therefore, we propose organizing the presentation of EC experiments into 7 parts, as follows.

ER-1: Research question
Briefly names the matter dealt with, the (possibly very general) objective, preferably in one sentence. This is used as the report's 'headline'.

ER-2: Preexperimental planning
Summarizes the first—possibly explorative—program runs, leading to task and setup (ER-3 and ER-4). Decisions on employed benchmark problems or performance measures shall be taken according to the data collected in preliminary runs. The report on preexperimental planning shall also include negative results, e.g. modifications to an algorithm that did not work, or a test problem that turned out to be too hard, if they provide new insight.

ER-3: Task
Concretizes the question in focus and states scientific claim and derived statistical hypotheses to test. Note that one scientific claim may require several, sometimes hundreds of statistical

hypotheses. In case of a purely explorative study, as with the first test of a new algorithm, statistical tests may be not applicable. Still, the task should be formulated as precise as possible.

ER-4: Setup
Specifies problem design and algorithm design, including the investigated algorithm, the controllable and the fixed parameters, and the chosen performance measuring. The information provided in this part should be sufficient to replicate an experiment.

ER-5: Results/Visualization
Gives raw or produced (filtered) data on the experimental outcome, additionally provides basic visualizations where meaningful.

ER-6: Observations
Describes exceptions from the expected, or unusual patterns noticed, without subjective assessment or explanation. As an example, it may be worthwile to look at parameter interactions. Additional visualizations may help to clarify what happens.

ER-7: Discussion
Decides about the hypotheses specified in part ER-3, and provides necessarily subjective interpretations for the recorded observations.

This scheme is tightly linked to the 12 steps of experimentation suggested in [BB06] and depicted in Tab. 1, but on a slightly more abstract level. The scientific claim and statistical hypothesis are treated together in part ER-3, and the SPO core (parameter tuning) procedure, much of which may be automated, is hidden in part ER-5. In our view, it is especially important to divide parts ER-6 and ER-7, to facilitate different conclusions drawn by others, based on the same

Tabelle 1: SPO as 12 step procedure given in [BB06] may be further partitioned into three phases on two conceptual levels. Phase I (green/light): Experiment construction, phase II (orange/dark): Tuning core, and phase III (purple/light): Result evaluation. Phases I and III build the basic methodological framework, and phase II is especially targetted at highly parameter-dependent optimization algorithms like EAs.

Step	Action
(S-1)	Preexperimental planning
(S-2)	Scientific claim
(S-3)	Statistical hypothesis
(S-4)	Specification of the (a) optimization problem (b) constraints (c) initialization method (d) termination method (e) algorithm (important factors) (f) initial experimental design (g) performance measure
(S-5)	Experimentation
(S-6)	Statistical modeling of data and prediction
(S-7)	Evaluation and visualization
(S-8)	Optimization
(S-9)	Termination: If the obtained solution is good enough, or the maximum number of iterations has been reached, go to step (S-11)
(S-10)	Design update and go to step (S-5)
(S-11)	Rejection/acceptance of the statistical hypothesis
(S-12)	Objective interpretation of the results from step (S-11)

results/observations. This distinction into three parts of increasing subjectiveness is similar to the suggestions of Barr et al. [BGK$^+$95] who distinguish between results, their analysis, and the conclusions drawn by the experimenter. We suggest to employ this report organization (or derive a better one) for all experiments in forthcoming work on Evolutionary Computation.

Acknowledgment

The research leading to this paper was supported by the DFG (Deutsche Forschungsgemeinschaft) as part of the collaborative research center “Computational Intelligence” (531) and by project grant no. 252441, “Mehrkriterielle Struktur- und Parameteroptimierung verfahrenstechnischer Prozesse mit evolutionären Algorithmen am Beispiel gewinnorientierter unscharfer destillativer Trennprozesse”.

Special thanks go to Thomas Bartz-Beielstein and Boris Naujoks, without whom I would never have thought about experiment reports, or would never have published these suggestions.

Literatur

[BB06] Thomas Bartz-Beielstein. *Experimental Research in Evolutionary Computation – The New Experimentalism.* Natural Computing Series. Springer, Berlin, 2006.

[BGK$^+$95] Richard S. Barr, Bruce L. Golden, James P. Kelly, Mauricio G.C. Resende, and William R. Stewart. Designing and reporting on computational experiments with heuristic methods. *Journal of Heuristics*, 1(1):9–32, 1995.

[Cat86] Catherine Cole McGeoch. *Experimental analysis of algorithms.* PhD thesis, Carnegie Mellon University, Pittsburgh, 1986.

[SWMW89] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn. Design and analysis of computer experiments. *Statistical Science*, 4(4):409–435, 1989.