IRIDIA

Université Libre de Bruxelles

Institut de Recherches Interdisciplinaires
et de Développements en Intelligence Artificielle

How to assess and report the performance of a stochastic algorithm on a benchmark problem:

Mean or ***best*** result on a number of runs?

Mauro Birattari and Marco Dorigo

IRIDIA – Technical Report Series

Technical Report No.
TR/IRIDIA/2005-007

April 2005
Last revision: May 2005

IRIDIA – Technical Report Series
ISSN 1781-3794

Published by:

IRIDIA, *Institut de Recherches Interdisciplinaires
et de Développements en Intelligence Artificielle*
Université Libre de Bruxelles
Av F. D. Roosevelt 50, CP 194/6
1050 Bruxelles, Belgium

Technical report number TR/IRIDIA/2005-007

Revision history:

TR/IRIDIA/2005-007.002	May 2005
TR/IRIDIA/2005-007.001	April 2005

The information provided is the sole responsibility of the authors and does not necessarily reflect the opinion of the members of IRIDIA. The authors take full responsability for any copyright breaches that may result from publication of this paper in the IRIDIA – Technical Report Series. IRIDIA is not responsible for any use that might be made of data appearing in this publication.

How to assess and report the performance of a stochastic algorithm on a benchmark problem:

Mean or *best* result on a number of runs?

Mauro Birattari mbiro@ulb.ac.be
Marco Dorigo mdorigo@ulb.ac.be

IRIDIA, Université Libre de Bruxelles, Brussels, Belgium

Last revision: May 2005

Notwithstanding the publication of a number of good methodological papers [1, 2, 3, 4], many research works dealing with stochastic optimization algorithms still propose unsatisfactory empirical assessments. It is undeniable that empirical analyses play a major role in the study of stochastic optimization algorithms, in particular of metaheuristics for which gaining an analytical insight appears rather problematic. For this reason, we think that elevating the quality of empirical studies is of paramount importance for the community: Much of the future of our research field will depend on the definition of high-quality experimental standards to be consistently followed in all works. With this short paper, we wish to address an apparently still open issue concerning how to measure and report the performance of stochastic algorithms.

In most research works, the performance of one or more stochastic algorithms on one or more benchmark problem instances is evaluated. In unfortunately far too many cases, authors perform a number of runs of an algorithm under analysis on each benchmark problem instance and then report, for each instance, the *best* result observed. We show in the following that this is an improper practice that leads to an *over-optimistic* assessment. We argue that a more meaningful assessment is obtained if the *mean* performance is reported. Moreover, we illustrate some convenient *non-parametric* alternatives, such as reporting the *median* performance or other *quantiles*.

We consider here a minimization problem.[1] Moreover, we discuss the case in which the measure of performance considered is the cost of the best solution found in a run of t seconds. The same considerations can be drawn for other measures such as, for example, the time needed by the algorithm under analysis to find the best solution or, alternatively, the time needed to find a solution of a given quality.

All authors seem to agree on the following statement:

Statement 1. *For evaluating the performance of a stochastic algorithm on a benchmark problem instance, one single run is not sufficient.*

This statement should be clarified: Indeed, the performance of a stochastic algorithm $\mathcal{A}$ is a *random variable* and its full description is provided by a probability distribution or, equivalently, by all its infinite moments. Nevertheless, it is customary to think that a sufficiently profitable description of the performance of a stochastic algorithm can be given by providing some *parameter* of the aforementioned distribution. Typically, for a number of theoretical and practical reasons, the *expectation*—that is, the first moment m_1 of the distribution—is considered. The widely accepted praxis consists in running the algorithm $\mathcal{A}$ under analysis for a number of times, say N,

[1] The modifications needed for handling a maximization problem are trivial and we do not feel like commenting any further on the issue.

and reporting then the *mean* performance observed. Formally, if c_i is the cost of the best solution found in run i, it is customary to report the *empirical mean*:

$$\hat{\mu}_N = \frac{1}{N}\sum_{i=1}^{N} c_i. \tag{1}$$

Some authors report also an estimate $\hat{\sigma}^2_N$ of the variance σ^2, which is strictly related to the second moment m_2 of the distribution:

$$\hat{\sigma}^2_N = \frac{1}{N-1}\sum_{i=1}^{N}(c_i - \hat{\mu}_N)^2. \tag{2}$$

Since the cost of the best solution found in a run is a random variable, the estimator $\hat{\mu}_N$ is a random variable as well.[2] It is well known from basic statistics[3] that, for any value of N, $\hat{\mu}_N$ is an *unbiased* estimator of the expectation μ, that is, $\mathrm{E}[\hat{\mu}_N] = \mu, \forall N$. Moreover, the variance of $\hat{\mu}_N$ is:

$$\mathrm{var}\,[\hat{\mu}_N] = \frac{\sigma^2}{N}, \tag{3}$$

where σ^2 is the variance of the cost of the best solution found in a run of the stochastic algorithm $\mathcal{A}$ under analysis.[4] The variance of the estimator decreases therefore with N. In some sense, this result can be taken as a formal justification of Statement 1. At the same time, Equation 3 urges to a less categorical reading of Statement 1. Indeed, it is not *strictly* true that one single run is not sufficient: The estimator based on one single run is indeed *unbiased* and has variance σ^2. Whether one run is *sufficient* or not depends solely on the magnitude of σ^2 and on the degree of uncertainty on the estimate that we are willing to tolerate. Considering more runs, say N, is indeed a good idea, since this effectively reduces the variance of the estimator.

Reporting $\hat{\mu}_N$ as a measure of performance of a stochastic algorithm $\mathcal{A}$ is particularly appropriate since it is an unbiased estimate of the *expectation* which, as the name itself suggests, is the cost we should expect to observe if we were running once our stochastic algorithm $\mathcal{A}$ on the instance under study. Reporting also $\hat{\sigma}^2_N$ is for sure beneficial since it brings further information on the distribution of the costs and, thanks to Equation 3, it gives a measure of the accuracy of the estimate of μ provided by $\hat{\mu}_N$. Reporting the pair $\hat{\mu}_N$ and $\hat{\sigma}^2_N$ is particularly appropriate if the costs are normally distributed. In this case, mean and variance univocally define the distribution. On the contrary, if the distribution is far from being normal and in particular if it is strongly asymmetric with a long tail, as it is often the case with stochastic algorithms, reporting (solely) these two quantities can be improper and even misleading.

The approach that we have outlined can be called *parametric statistics* since it aims at estimating some parameters of the distribution of the costs. An alternative approach that is adopted by some authors and that, in our opinion, should find a wider application goes under the name of *nonparametric statistics*. It consists in describing the distribution of the costs through its *quantiles*. A *quantile* is a value of a variable that divides the distribution into two parts: on the one hand, the values greater than the quantile value and on the other hand the values that are less. In particular, some authors report the *median* (50th percentile) of the distribution which is a value that divides the distribution in two equal parts: If we run N times the stochastic algorithm $\mathcal{A}$ under analysis we should expect to obtain $N/2$ times a cost that is larger than the median and $N/2$ times a cost that is less. Other quantiles that are often considered are the *first and third quartiles* (25th and 75th percentile, respectively) which split the distribution in 25%/75% and 75%/25%, respectively. For some specific applications, other quantiles could be used: for example, the 95th

[2] The same holds for $\hat{\sigma}^2_N$.

[3] See any textbook on probability and statistics as, for example, Papoulis [5].

[4] To be more precise, Equation 3 holds only under some hypotheses which are nevertheless satisfied in the typical experimental settings adopted when assessing the performance of a stochastic algorithm. In particular, the equation relies on the *independence* of the N runs considered.

or the 98th percentile effectively comprise *almost* the whole distribution. A nonparametric approach is particularly appropriate for describing *asymmetric* and long-tailed distributions as those that are typically encountered when dealing with the performance of a stochastic algorithm. The nonparametric approach is often adopted in graphical representations of experimental results: Commonly encountered *box-plots* and *histograms* are of nonparametric nature.

Both the parametric and the nonparametric approach, as outlined above, can be effectively adopted for assessing and reporting the performance of a stochastic algorithm. Unfortunately, way too often some authors report the *best* result obtained in a number of runs. Formally:

$$b_N = \min_i c_i. \tag{4}$$

This quantity is not of real interest. Indeed, it is just a particularly *over-optimistic* measure of the performance of the stochastic algorithm $\mathcal{A}$ under analysis. It should be noticed that the quantity b_N is not either a *good* estimator of the cost of the best solution that algorithm $\mathcal{A}$ can find. The empirical estimation of the minimum of a distribution is particularly problematic. It is indeed always biased, since all possible observations are by definition larger than or equal to the quantity to be estimated. Even worse, the uncertainty on the estimate does not nicely decrease with the size of the sample—as it does, for example, in the case of the estimation of the expected value. Indeed, irrespectively of the size of the sample, it is always possible that the observed quantity assumes an arbitrary small value which might emerge with very low probability: In virtue of its low probability, such a small value might fail to reveal itself even in large samples.

Some authors, see for example Eiben and Jelasity [4], justify the use of b_N by saying that in a real-world application, if they were to find a good solution to the given problem instance, they would run their algorithm $\mathcal{A}$ for a number of times, say N, and then they would return the *best* solution found in these N runs. Although this seemingly reasonable argumentation contains some elements of truth, it is nevertheless faulty on a number of grounds. Some closer inspection is needed in order to disentangle the undeniable facts from the somehow dubious conclusions: On the one hand, it is perfectly legitimate to run $\mathcal{A}$ for N times and to use the best result found; in this sense, Eiben and Jelasity are right when they maintain that a proper research methodology should take this widely adopted practice into account. On the other hand, the argumentation cannot be used for justifying the use of b_N as a measure of the performance of the algorithm $\mathcal{A}$. In the following, we accept the fact that one might wish to run $\mathcal{A}$ for N times for then selecting the best result and we highlight two main issues.

First of all, it should be recognized that, in this case, we are not discussing $\mathcal{A}$ but rather another algorithm, call it $\mathcal{A}^N$, which consists in *random restarting* $\mathcal{A}$ for N times. This entails two obligations. On the one hand, we should be clear from the beginning that we are interested in the $\mathcal{A}^N$ algorithm and not in $\mathcal{A}$. On the other hand, if we are indeed interested in $\mathcal{A}^N$ we should provide a proper assessment of it. In particular, it should be recognized that by reporting the quantity b_N we are considering a *single* run of $\mathcal{A}^N$. Since $\mathcal{A}^N$ is itself a stochastic algorithm, this contrasts with the general practice defined by Statement 1. An appropriate experimental methodology would study the *expected* performance of $\mathcal{A}^N$ and would operate so that the variance of the estimator is properly reduced. To this aim, one should average over say M runs of $\mathcal{A}^N$. This implies running $N * M$ times the underlying algorithm $\mathcal{A}$. As an alternative, a *nonparametric* analysis of the kind outlined above could be profitably adopted. Also in this case a number of say M runs of $\mathcal{A}^N$ are needed.

The second issue we wish to discuss is of a more subtle, if not provocative nature: The claim that we might be interested in $\mathcal{A}^N$ rather than in $\mathcal{A}$ sounds particularly suspicious when it comes from researchers working on metaheuristics. A metaheuristic is typically understood [6, 7, 8] as a general-purpose method for guiding an underlying optimization algorithm, such as a problem-specific heuristic or a local search. In this sense, *random restart*, which consists in performing a number of say N independent runs of the underlying algorithm, can be seen as the most trivial metaheuristic: the *null*-metaheuristic. Indeed, many well-designed empirical analysis of metaheuristics include random restart as a performance yardstick: Failing to improve over random restart is to be considered as a major failure for a metaheuristic. In the light of these considerations, it should sound strange that a researcher working in the metaheuristics field adopts a random restart strategy when he could have recourse to a more advance metaheuristic. This sounds somehow like betraying the fundamental principles of the research on metaheuristic.

Acknowledgments. The authors acknowledge support from the ANTS project, an *Action de Recherche Concertée* funded by the Scientific Research Directorate of the French Community of Belgium. Marco Dorigo acknowledges support from the Belgian FNRS, of which he is a Research Director.

References

[1] R. S. Barr, B. L. Golden, J. P. Kelly, M. G. C. Resende, and W. R. Stewart. Designing and reporting computational experiments with heuristic methods. *Journal of Heuristics*, 1(1):9–32, 1995.

[2] J. N. Hooker. Testing heuristics: We have it all wrong. *Journal of Heuristics*, 1(1):33–42, 1995.

[3] R. R. Rardin and R. Uzsoy. Experimental evaluation of heuristic optimization algorithms: A tutorial. *Journal of Heuristics*, 7(2):261–304, 2001.

[4] A. E. Eiben and M. Jelasity. A critical note on experimental research methodology in EC. In *Proceedings of the 2002 Congress on Evolutionary Computation (CEC'2002)*, pages 582–587. IEEE Publications, 2002.

[5] A. Papoulis. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, New York, NY, USA, third edition, 1991.

[6] T. G. Stützle. *Local Search Algorithms for Combinatorial Problems – Analysis, Algorithms, and New Applications*. PhD thesis, Technische Universität Darmstadt, Darmstadt, Germany, 1999.

[7] F. Glover and G. Kochenberger, editors. *Handbook of Metaheuristics*. Kluwer Academic Publisher, Norwell, MA, USA, 2002.

[8] H. H. Hoos and T. Stützle. *Stochastic Local Search. Foundations and Applications*. Morgan Kaufmann, San Francisco, CA, USA, 2004.