A Note on Performance Profiles for Benchmarking Software

NICHOLAS GOULD and JENNIFER SCOTT, Rutherford Appleton Laboratory

In recent years, performance profiles have become a popular and widely used tool for benchmarking and evaluating the performance of several solvers when run on a large test set. Here we use data from a real application as well as a simple artificial example to illustrate that caution should be exercised when trying to interpret performance profiles to assess the relative performance of the solvers.

Categories and Subject Descriptors: G.4 [**Mathematical Software**]: Algorithm design and analysis

General Terms: Algorithms, Performance

Additional Key Words and Phrases: Performance profiles, benchmarking, testing

ACM Reference Format:
Nicholas Gould and Jennifer Scott. 2016. A note on performance profiles for benchmarking software. ACM Trans. Math. Softw. 43, 2, Article 15 (August 2016), 5 pages.
DOI: http://dx.doi.org/10.1145/2950048

1. INTRODUCTION TO PERFORMANCE PROFILES

The quantities of data that results from benchmarking mathematical software (such as optimization packages or sparse linear solvers) with large problem sets have naturally led to researchers developing tools to analyse the data. A popular and widely used tool is the performance profile, which was proposed by Dolan and Moré [2002] as a means of providing objective information when benchmarking optimization software. Since their introduction, performance profiles have been used in many studies; as of May 2016, there were more than 1,750 citations of the original paper [Dolan and Moré 2002] listed on Google Scholar.

Benchmark results are generated by running a solver on a set $\mathcal{T}$ of problems and recording the information of interest (which might include, for example, the computation time, the number of function evaluations, the number of iterations or the memory used). Let $\mathcal{S}$ represent the set of solvers that are to be compared. Suppose that a given solver $i \in \mathcal{S}$ reports a statistic $s_{ij} \geq 0$ when run on example j from the test set $\mathcal{T}$, and that the smaller this statistic the better the solver is considered to be. For $j \in \mathcal{T}$, let $\hat{s}_j = \min\{s_{ij} : i \in \mathcal{S}\}$, and define $r_{ij} = s_{ij}/\hat{s}_j$ to be the *performance ratio*.[1] Then for $f \geq 1$ and each $i \in \mathcal{S}$, define

$$k(r_{ij}, f) = \begin{cases} 1 & \text{if } r_{ij} \leq f \\ 0 & \text{otherwise.} \end{cases}$$

[1] If a solver $i \in \mathcal{S}$ fails to solve problem j, $r_{ij} = \infty$.

This work was supported by EPSRC grant EP/I013067/1.

Authors' addresses: N. I. M. Gould and J. A. Scott, Scientific Computing Department, STFC Rutherford Appleton Laboratory, Harwell Campus, Oxfordshire, OX11 0QX, UK; emails: nick.gould@stfc.ac.uk, jennifer.scott@stfc.ac.uk.

DOI: http://dx.doi.org/10.1145/2950048

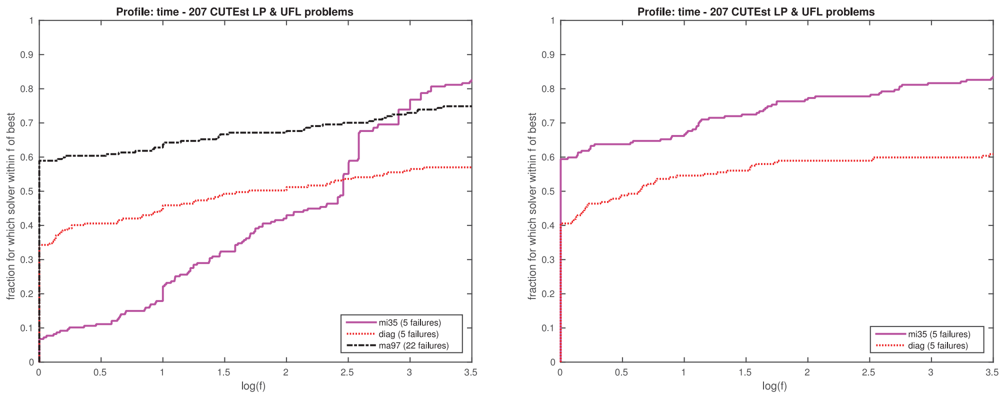

Fig. 1. Time performance profiles for a real test case for $\mathcal{S} = \{$ma97, mi35, diag$\}$ for $\tilde{\mathcal{S}} = \{$mi35, diag$\}$.

The *performance profile* of solver i is given by the function

$$p_i(f) = \frac{\sum_{j \in \mathcal{T}} k(r_{ij}, f)}{|\mathcal{T}|}, \quad f \geq 1,$$

where $|\mathcal{T}|$ denotes the cardinality of $\mathcal{T}$. Thus, $p_i(f)$ is the probability for solver $i \in \mathcal{S}$ that a performance ratio r_{ij} for each $j \in \mathcal{T}$ is within a factor f of the best possible ratio. In particular, $p_i(1)$ gives the fraction of the examples in $\mathcal{T}$ for which solver i is the winner (i.e., the best according to the statistic s_{ij}), while $p_i^* := \lim_{f\to\infty} p_i(f)$ gives the fraction of $|\mathcal{T}|$ for which solver i is successful. If we are just interested in the number of wins on $\mathcal{T}$, we need only compare the values of $p_i(1)$ for all the solvers $i \in \mathcal{S}$, but if we are interested in solvers with a high probability of success on the set $\mathcal{T}$, we should choose those for which p_i^* is largest.

As many researchers have found, for a selected test set, performance profiles provide a very useful and convenient means of assessing the performance of a solver relative to the best solver on each example from that set. When commenting on a performance profile presented in their paper, Dolan and Moré state that it “gives a clear indication of the relative performance of each solver” (see also Moré and Wild [2009]), and they go on to say that “performance profiles provide an estimate of the expected performance difference between solvers.” Data from a practical study of solvers applied to a large test set and a simple artificial example will show that using performance profiles to assess the relative performance of the solvers should be undertaken with a degree of caution.

2. EXAMPLE

We recently carried out a study to assess the performance of a number of sparse solvers (here denoted as diag, mi35, and ma97) on a set $\mathcal{T}$ of 207 linear least squares problems; details may be found in Gould and Scott [2015a, 2015b]. In particular, solution times for each solver were recorded and the performance measure s_{ij} was taken to be the time for solver i on problem j. One of the time performance profiles we obtained during the preliminary stages of our study is given in Figure 1(a). Here and elsewhere, log denotes logarithm to the base 2. The set of solvers is $\mathcal{S} = \{$ma97, mi35, diag$\}$. From this figure, it is clear that while the solver ma97 has the most failures (22 failures compared to 5 failures for solvers mi35 and diag), it has the highest number of wins (it is the fastest on 59% of the problems), and over our chosen range of f, it dominates the other solvers, while the solver diag wins a respectable 34% of the time. Solver mi35 has the lowest number of wins, and if we are only interested in solvers that are within a factor 5 of

Table I. Performance of Three Solvers on a Test Set $\mathcal{T}$ of Five Problems; Here, the Smaller the Statistic, the Better the Solver Performance

Problem	Solver A	Solver B	Solver C
1	2	1.5	1
2	1	1.2	2
3	1	4	2
4	1	5	20
5	2	5	20

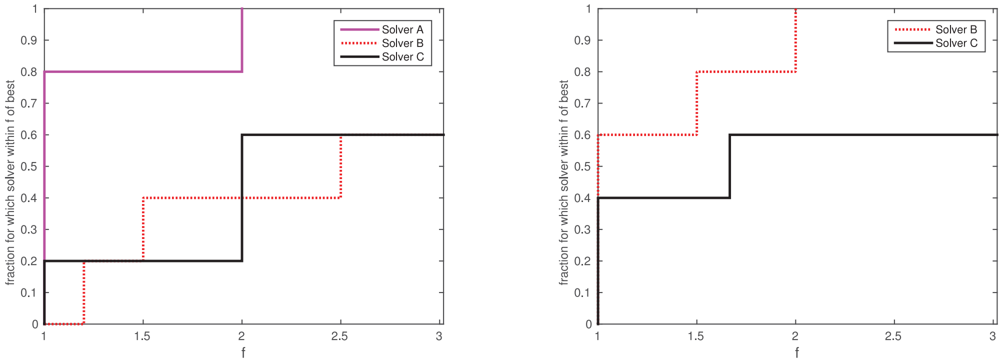

Fig. 2. Performance profiles for our artificial test case for $\mathcal{S}_1$ = {Solver A, Solver B, Solver C} and $\mathcal{S}_2$ = {Solver B, Solver C}.

the best, then it is tempting to conclude that, as the curve for solver mi35 lies below the other curves for $f \in [1, 5]$, it is the worst solver in $\mathcal{S}$ on the set $\mathcal{T}$ (see, e.g., Higham [2009] where a similar conclusion is drawn). However, if we remove solver ma97 and redraw the performance profiles for $\tilde{\mathcal{S}}$ = {mi35, diag}, we obtain Figure 1(b). We see that solver mi35 is the better solver in $\tilde{\mathcal{S}}$ for $f \in [1, 10]$.

This apparent change in fortunes can be seen clearly using the artificial sample data for five test problems and three solvers given in Table I and the corresponding performance profiles given in Figure 2. With $\mathcal{S}_1$ = {Solver A, Solver B, Solver C}, Solver A is the best on 80% of the problems in the test set, Solver B is not the winner on any, and if we are interested in having a solver that can solve at least 60% of the test problems with the greatest efficiency, then Solver A or C should be chosen. However, if $\mathcal{S}_2$ = {Solver B, Solver C} (i.e., Solver A is removed), Solver B, which was the second best solver in $\mathcal{S}_1$ on 60% of the test set, is the best solver in $\mathcal{S}_2$. In Figure 1(a), it is not apparent that solver mi35 is the second best solver in $\mathcal{S}$ for the set $\mathcal{T}$ on the interval $f \in [1, 5]$.

3. CONCLUSIONS

When comparing two solvers on a given test set, performance profiles give a clear measure of which is the better solver for a selected range of f. But as the examples above illustrate, if performance profiles are used to compare more than two solvers (and Dolan and Moré state that "performance profiles are most useful in comparing several solvers"), we can determine which solver has the highest probability $p_i(f)$ of being within a factor f of the best solver for f in a chosen interval, but we cannot necessarily assess the performance of one solver relative to another that is not the best. In some situations, being able to rank (or partially rank) the solvers may be important. For example, a user may not have access to the best solver and so may want to know which is second (or perhaps third) best. To rank the solvers for a chosen range

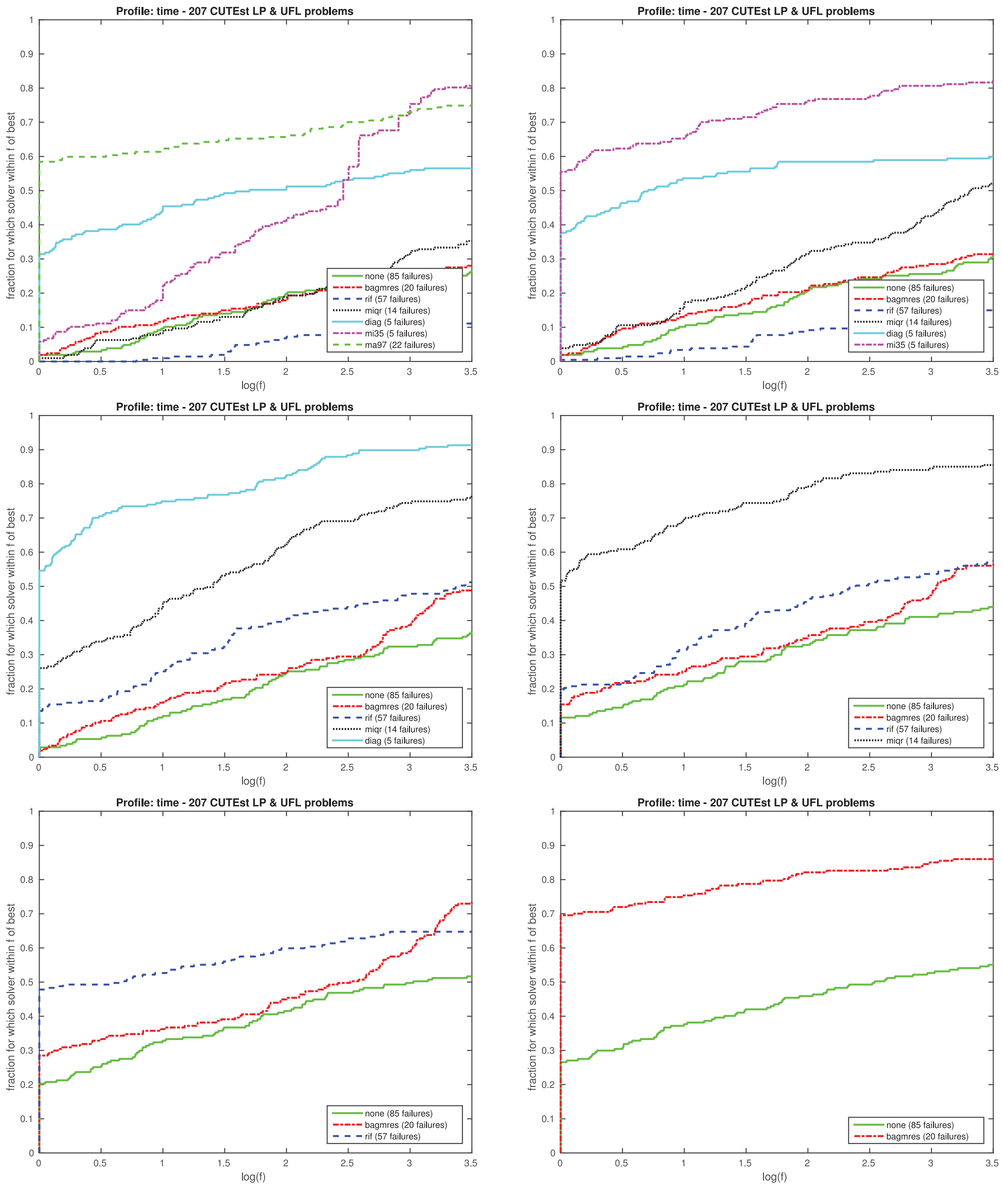

Fig. 3. A sequence of time performance profiles for the real test case from Section 2 in which the “best” solver is removed from the set $\mathcal{S}$ of solvers until only two remain.

$[1, f]$, an obvious approach is to produce a series of performance profiles, excluding the best solver over the range from successive profiles until only two remain. We illustrate this in Figure 3, again using real data from our least squares study but now with a larger set $\mathcal{S}$ of solvers. Notice how that, as before, removing the “leading” solver ma97 from $\mathcal{S}$ exposes mi35 as the runner up, and further removals illustrate that solver rif is higher in the performance hierarchy than the initial profiles might suggest.

A switch in the expected ordering may indicate the test set contains a large number of problems for which each solver performs in a consistent way and further examination of the test set and how it was selected may be advisable. However, our experience has been that, even without such a subset apparently present within the test set, switches can occur. We conclude that, while performance profiles are a powerful tool for benchmarking a solver relative to the best solver, as Dolan and Moré point out, "performance profiles must be used with care." Finally, we observe that elsewhere in the literature, limitations of performance profiles have been noted and other tools for comparing performances have been proposed (see, e.g., Moré and Wild [2009]).

ACKNOWLEDGMENTS

Many thanks to our colleagues in the Numerical Analysis Group at the Rutherford Appleton Laboratory for discussions on our findings and commenting on a draft of this note. Our thanks also to three anonymous referees.

REFERENCES

E. D. Dolan and J. J. Moré. 2002. Benchmarking optimization software with performance profiles. *Mathematical Programming* 91, 2 (2002), 201–213.

N. I. M. Gould and J. A. Scott. 2015a. *The State-of-the-Art of Preconditioners for Sparse Linear Least Squares Problems*. Technical Report RAL-P-2015-10. Rutherford Appleton Laboratory.

N. I. M. Gould and J. A. Scott. 2015b. *The State-of-the-Art of Preconditioners for Sparse Linear Least Squares Problems: The Complete Results*. Technical Report RAL-TR-2015-09. Rutherford Appleton Laboratory.

N. J. Higham. 2009. The scaling and squaring method for the matrix exponential revisited. *SIAM Review* 51, 4 (2009), 747–767.

J. J. Moré and S. M. Wild. 2009. Benchmarking derivative-free optimization algorithms. *SIAM Journal on Optimization* 20, 1 (2009), 172–191.

Received June 2015; revised May 2016; accepted May 2016