Implementing the Cumulative Difference Plot in the IOHanalyzer

Etor Arza
BCAM - Basque Center for Applied Mathematics
Bilbao, Spain
earza@bcamath.org

Josu Ceberio
University of the Basque Country (UPV/EHU)
Donostia-San Sebastian, Spain
josu.ceberio@ehu.eus

Ekhiñe Irurozki
Telecom Paris
Paris, France
irurozki@telecom-paris.fr

Aritz Pérez
BCAM - Basque Center for Applied Mathematics
Bilbao, Spain
aperez@bcamath.org

ABSTRACT

The *IOHanalyzer* is a web-based framework that enables an easy visualization and comparison of the quality of stochastic optimization algorithms. *IOHanalyzer* offers several graphical and statistical tools analyze the results of such algorithms. In this work, we implement the cumulative difference plot in the *IOHanalyzer*. The cumulative difference plot [1] is a graphical approach that compares two samples through the first-order stochastic dominance. It improves upon other graphical approaches with the ability to distinguish between a small magnitude of difference and high uncertainty.

KEYWORDS

first order stochastic dominance, benchmarking, graphical statistics

ACM Reference Format:
Etor Arza, Josu Ceberio, Ekhiñe Irurozki, and Aritz Pérez. 2022. Implementing the Cumulative Difference Plot in the IOHanalyzer. In *Genetic and Evolutionary Computation Conference Companion (GECCO '22 Companion), July 9–13, 2022, Boston, MA, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3520304.3534050

1 INTRODUCTION

The *IOHprofiler* [2] is a benchmarking tool to compare the performance of iterative optimization algorithms. It has several modules, each with a different benchmark-related purpose. One of such modules is the *IOHanalyzer* [6]: a web-based interface for visualizing and statistically assessing the differences in performance of the algorithms.

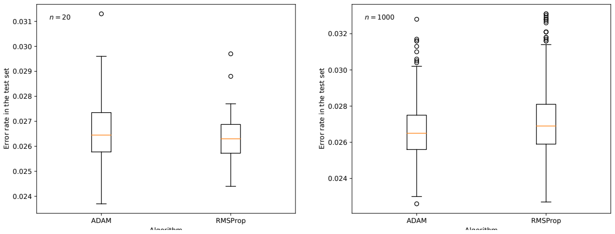
The *IOHanalyzer* proposes analyzing the data from different perspectives. The best well-known measures are probably the *expected value of the objective function with a fixed evaluation budget (quality)*, and the *expected number of function evaluations to obtain a target value of the objective function (evaluations).*

For the sake of simplicity, we now focus on comparing the *quality* of the algorithms, although the methodology discussed in the following is also applicable to the *evaluations*. Once we choose an *evaluation budget* and assume there is a single problem instance with several runs of the optimization algorithms, the task is simplified to comparing several random variables through their observed samples. *IOHanalyzer* offers two visualizations to compare these samples: the histogram and the box/violin plot.

1.1 Motivation

The histogram and the box/violin plot are two visualization tools designed to summarize samples. They are also two of the most used visualizations to compare samples, each with its limitations. To illustrate such limitations, let us look at an example. Let us assume that we are interested in comparing *ADAM* and *RMSProp*: two gradient descent-based algorithms, useful for training neural networks. Specifically, we compare them in an image classification task in the *MNIST* dataset. We use the term *observation* of the quality to refer to training the neural network in the train-set and measuring the accuracy in the test-set once. Since the parameters of the neural networks are randomly initialized, each observation can be different [3]. We save 1000 observations of the quality for each algorithm.

Figure 1 shows the box plots of 20 and 1000 observations. If we only consider 20 observations (Figure 1a), it would seem that *RMSProp* is slightly better algorithm than *ADAM* (*RMSProp* has lower median and slightly smaller outliers). However, with 1000 samples, we obtain the opposite result. The same applies to the histogram, as shown in Figure 2.



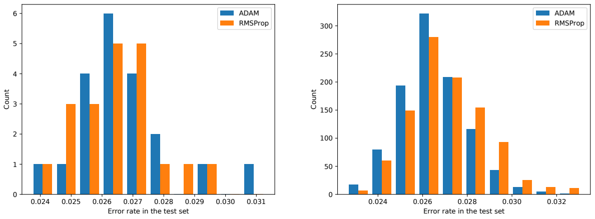
(a) 20 observations

(b) 1000 observations

Figure 1: box plots of the observations of the quality of *ADAM* and *RMSProp.*

GECCO '22 Companion, July 9–13, 2022, Boston, MA, USA

ACM ISBN 978-1-4503-9268-6/22/07.
https://doi.org/10.1145/3520304.3534050



(a) 20 observations

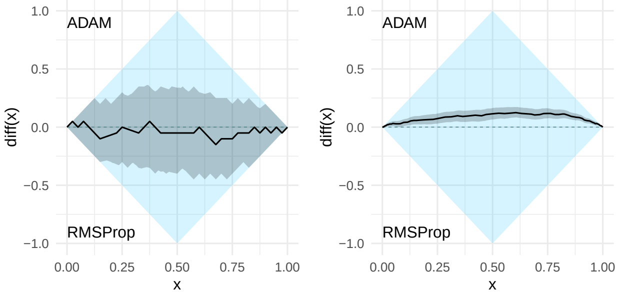
(b) 1000 observations

Figure 2: Histograms of the observations of the quality of *ADAM* and *RMSProp*.

The box plot and the histogram show the uncertainty and the magnitude of the difference in the same way, and this is what causes the conclusions to be different. With both sample sizes, it seems that the difference is small, but with 20 observations, the uncertainty is too high for the result to be meaningful (an increase in the sample size provides a different result). These two plots have other limitations: for example, several different populations can have the same box plot [4].

2 CUMULATIVE DIFFERENCE PLOT

As an alternative that overcomes these limitations, Arza et al. [1] proposed the cumulative difference plot. In Figure 3, we show the cumulative difference plot for the same observations as in the previous figures. The proposed plot compares the samples from two algorithms through the first-order stochastic dominance [5]. In the left side of the x-axis, the best values that the algorithms produced are compared, while on the right side, the worst values are compared. For example, if the cumulative difference (the black curve) is positive in $x = 0.25$, this means that the top **quartile** of the quality of *ADAM* is better than the top quartile of the quality of *RMSProp*. Consequently, if the cumulative difference is positive in $(0, 1)$, the quality of *ADAM* **stochastically dominates** [5] *RMSProp* (all the quantiles are better).



(a) 20 observations

(b) 1000 observations

Figure 3: Cumulative difference plot [1] of the observations of the quality of *ADAM* and *RMSProp*.

Unlike the box plot and the histogram, the cumulative difference plot models both the magnitude and the uncertainty of the difference between the algorithms. The gray area around the cumulative difference is the 95% **confidence band** (estimated via bootstrap), and represents the uncertainty of the estimate. Moreover, the proportion of the blue square that is under the cumulative difference is an estimation of the **probability** that the observed quality of *ADAM* is better than that of *RMSProp*. In addition, the estimation of the **dominance rate** [1]—a measure between 0 and 1 correlated with the first-order stochastic dominance—is equal to the length of the x-axis in which the cumulative difference is positive.

With this plot, we reach a different conclusion than with the histogram and the box plot. With 20 observations (Figure 3a), the uncertainty of the estimate is high (the confidence band is wide), and we cannot conclude that one of the algorithms is better than the other. On the other hand, with 1000 observations (Figure 3b), we conclude that the quality of *ADAM* stochastically dominates the quality of *RSMProp*. In addition, we can also deduce that *the probability that an observation of ADAM is better than an observation of RSMProp* is slightly higher than 0.5.

3 CONCLUSION

The *IOHanalyzer* [6] is a web-based tool to visually and statistically compare the performance of stochastic optimization algorithms. In this work, we added the cumulative difference plot [1] to the *IOHanalyzer*. The cumulative difference plot overcomes some of the limitations of the box plot and the histogram, such as the ability to model both the uncertainty and the magnitude of the difference between two algorithms. As future work, it might be interesting to adapt the methodology to the second order stochastic dominance.

Acknowledgments

This work is partially supported by the data science and artificial intelligence chair for digitalized industry and services, Telecom Paris, Institut Polytechnique de Paris, and by the Basque Government through the BERC 2022-2025 and the ELKARTEK programs (KK-2020/00049, SIGZE, KK-2021/00065) the Research Groups 2019-2021 (IT1244-19) and by the Spanish Ministry of Economy and Competitiveness, through BCAM Severo Ochoa excellence accreditation SEV-2017-0718 and the research project PID2019-106453GA-I00/AEI/10.13039/501100011033.

REFERENCES

[1] Etor Arza, Josu Ceberio, Ekhiñe Irurozki, and Aritz Pérez. 2022. Comparing Two Samples through Stochastic Dominance: A Graphical Approach. *arXiv preprint arXiv:2203.07889* (2022).

[2] Carola Doerr, Hao Wang, Furong Ye, Sander van Rijn, and Thomas Bäck. 2018. IOHprofiler: A Benchmarking and Profiling Tool for Iterative Optimization Heuristics. *arXiv e-prints:1810.05281* (Oct. 2018).

[3] Xavier Glorot and Yoshua Bengio. 2010. Understanding the Difficulty of Training Deep Feedforward Neural Networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 9)*, Yee Whye Teh and Mike Titterington (Eds.). JMLR Workshop and Conference Proceedings, Chia Laguna Resort, Sardinia, Italy, 249–256.

[4] Justin Matejka and George Fitzmaurice. 2017. Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 1290–1294.

[5] James P. Quirk and Rubin Saposnik. 1962. Admissibility and Measurable Utility Functions. *The Review of Economic Studies* 29, 2 (Feb. 1962), 140–146.

[6] Hao Wang, Diederick Vermetten, Furong Ye, Carola Doerr, and Thomas Bäck. 2022. IOHanalyzer: Detailed Performance Analyses for Iterative Optimization Heuristics. *ACM Transactions on Evolutionary Learning and Optimization* 2, 1 (March 2022), 1–29.