Long-run Behaviour of Multi-fidelity Bayesian Optimisation

Gbetondji J-S Dovonon GBETONDJI@MATTERHORN.STUDIO
University College London
Matterhorn Studio

Jakob Zeitler JAKOB@MATTERHORN.STUDIO
Matterhorn Studio

Abstract

Multi-fidelity Bayesian Optimisation (MFBO) has been shown to generally converge faster than single-fidelity Bayesian Optimisation (SFBO) (Poloczek et al. (2017)). Inspired by recent benchmark papers, we are investigating the long-run behaviour of MFBO, based on observations in the literature that it might under-perform in certain scenarios (Mikkola et al. (2023), Eggensperger et al. (2021)). An under-performance of MBFO in the long-run could significantly undermine its application to many research tasks, especially when we are not able to identify when the under-performance begins. We create a simple benchmark study, showcase empirical results and discuss scenarios and possible reasons of under-performance.

Keywords: Multi-fidelity Bayesian Optimisation, Long-run behaviour

1. Introduction

The optimisation of costly-to-evaluate functions is a significant challenge in hyperparameter-optimisation in machine learning, material science, drug discovery and more (Kandasamy et al. (2020), Liang et al. (2021), Bellamy et al. (2022)). Bayesian Optimisation (BO, Frazier (2018)) has become an established method to tackle these challenges, also due to the fact that it does not require access to a gradient. As part of so called *grey-box* BO methods (Astudillo and Frazier (2021)) that take into account internal structure of an optimisation problem, *multi-fidelity* BO (MBFO) has emerged as popular method to utilise access to different information sources on the same optimisation problem (Huang et al. (2006)).

2. Review of MFBO

The core assumption of MFBO is that we have access to different auxiliary fidelities that inform us on our target fidelity. Further assumptions on the nature of those auxiliary fidelities and their relationship to the target fidelity will determine the type of algorithm to use. Poloczek et al. (2016) provide one of the earliest implementations of the MFBO process which has established itself as the predominant choice. They adapt the knowledge gradient acquisition function to the MFBO setting, and provide a new set of mean and covariate functions. Two noteworthy design choices here are the continuous fidelities and the use of a multi-output gaussian process to model both the objective and fidelity. Similar extensions have been made for maximum entropy search (Takeno et al. (2020)) and expected improvement (Irshad et al. (2023); Daulton et al. (2020)) acquisition functions. While

several methods come with attractive theoretical guarantees, when applied to domains where Bayesian optimization is the preferred choice, it is common to see heuristics. On benchmarks like HPOBench (Eggensperger et al. (2021)), heuristic-driven methods dominate (Awad et al. (2021); Cowen-Rivers et al. (2022)).

3. Problem and Solution

Multi-fidelity BO (MFBO) is generally accepted to outperform single-fidelity BO approaches (SFBO). Only recently, the literature has focused on possible failure-modes of MFBO, that could rank it below established SFBO performance. This is partially driven by a range of new and more reliable benchmarks that allow fair comparison of BO algorithms, such as the HPOBench (Eggensperger et al. (2021)) that utilise execution containers to standardise comparison across compute environments.

(Mikkola et al. (2023)) recently evaluated the impact of unreliable information sources on MFBO performance. As part of their investigation, they compared MFBO and SFBO on the Hartmann6D test function, see Figure 1. We can observe a *cross-over point* at a budget of about 25, where the SF-MES starts to outperform the multi-fidelity methods. The confidence intervals still overlap up to the final budget point of 80 and as such do not allow us to conclude that single-fidelity, on average, is outperforming multi-fidelity approaches in the long-run.

Evaluation plots created with the HPOBench provide an additional perspective on our observation of a cross-over point in 1. (Eggensperger et al. (2021)) show in Figure 2 the mean rank of single-fidelity (dotted) and multi-fidelity (not dotted) methods over increasing fraction of budget spent. While the multi-fidelity methods clearly rank higher for most of the budget spend, we can observe a long-term trend of single-fidelity methods steadily improving in rank and outperforming the other methods at the very right edge of the plot (marked by a red circle). The nature of the budget used in HPOBench, a finite set of training data points, does not allow increasing the budget unless the dataset itself is increased. We hypothesize that in settings where hyperparameter optimization can be done efficiently on much larger datasets that single-fidelity might continue to outperform, in line with our observations in Figure 1.

To study our observations in Figure 1 and Figure 2, we devise our own simulation runs based on the BoTorch Hartmann6D tutorial where the Hartmann6D test function is augmented with a continuous fidelity choice (Balandat (2021)). Our results are shown in Figure 3. Plotting MFBO and SFBO in a single plot requires adjusting MFBO to the equidistant budget query points of SFBO. While SFBO with a query cost of 1 will query 100 samples with a budget of 100, MFBO will query more than 100 samples (i.e. low and high fidelity), most of the time not at the same budget point as SFBO. Hence, we 'normalise' our MFBO results by querying and reporting the high fidelity every time the MFBO run budget crosses over a SFBO budget point. This does not perfectly represent the MFBO decision making, but aligns the calculation of confidence intervals which leads to a more fair and interpretable comparison plot.

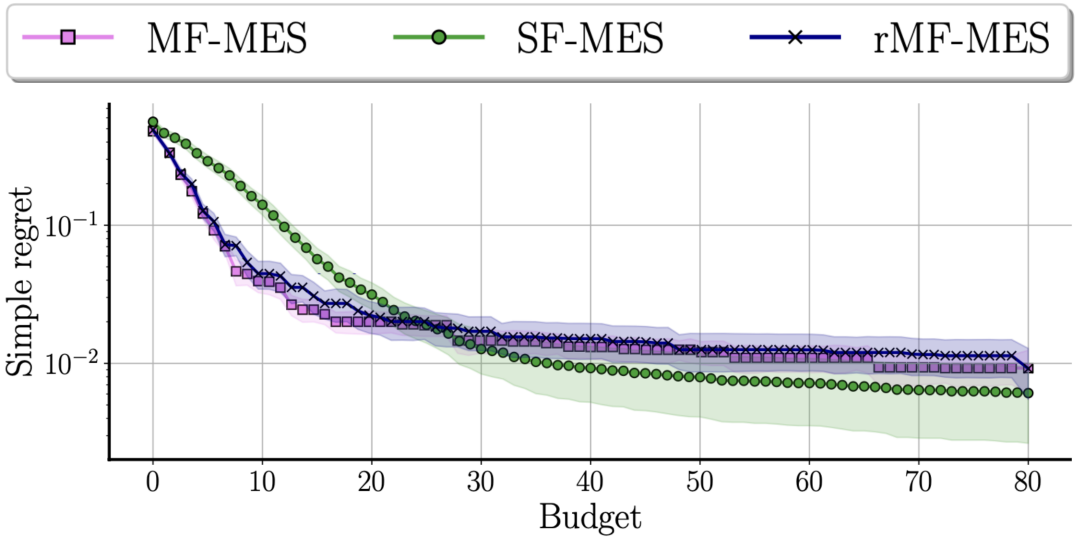

Figure 1: Possible long-run underperformance of MFBO as found in (Mikkola et al. (2023), Figure 1). The plot shows maximum-entropy search for single-fidelity (green, SF-MES), multi-fidelity (pink, MF-MES) and robust multi-fidelity (purple, rMF-MES, the method suggested by (Mikkola et al. (2023)) to handle unreliable information sources).

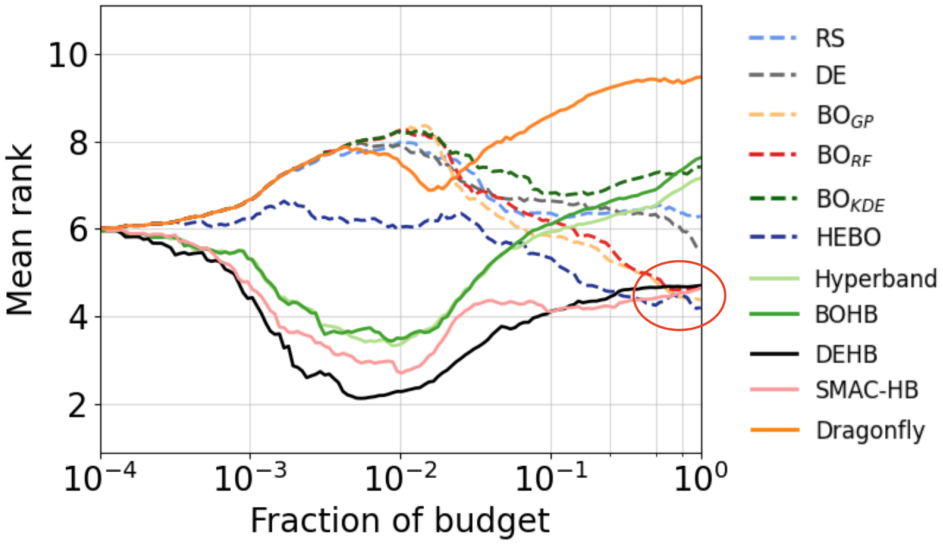

Figure 2: Possible long-run underperformance of MFBO as found in (Eggensperger et al. (2021), Figure 4). The plot shows the *mean rank* of single-fidelity (dotted) and multi-fidelity (not dotted) methods over increasing *fraction of budget spent*.

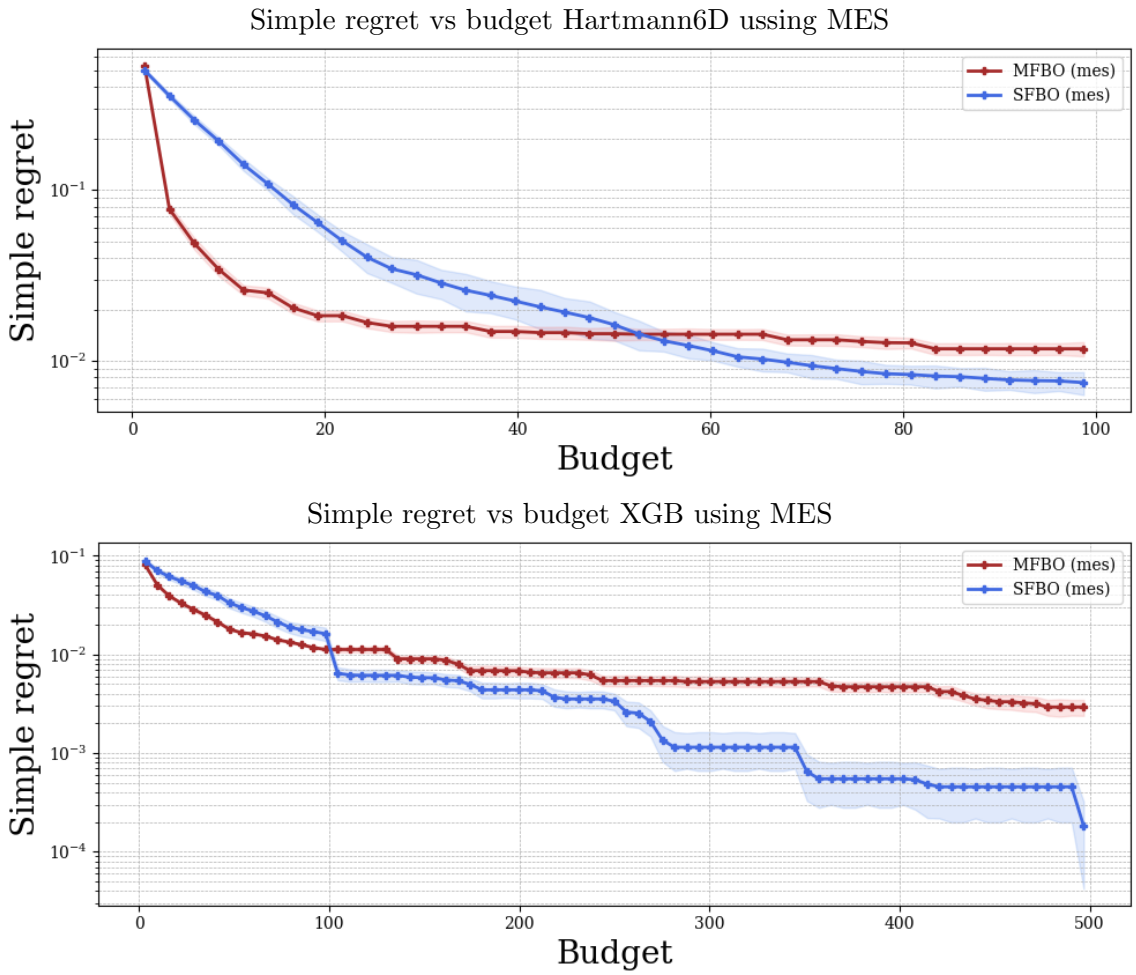

Figure 3: Our results of 100 trails each for SFBO and MFBO, at budget 100 for Hartmann6D and 500 for XGB, plotted with log-transformation on the y-axis. On both benchmarks, SFBO eventually outperforms MFBO. The crossing point is different with SFBO overtaking MFBO around a budget of 50 on Hartmann6D but 100 on XGB.

4. Discussion

We are able to reproduce the observations made in previous papers, as discussed in Figure 1 and 2. We observe the majority of low-fidelity queries occur at the very beginning of the run. MFBO seems to build a 'warm-start' set of low-fidelity observations which then informs a further run dominated by high-fidelity queries. That is why we see such strong performance of MFBO compared to SFBO in the initial budget range of 0 to 15, which is also usually observered in the MFBO literature.

Our initial results suggest both theoretical and practical next steps. It is important to collect more empirical results on different test functions and BO surrogates and acquisition functions, as our Figures only represent a small set of BO setups such that we cannot conclude that long-run underperformance of MFBO also occurs in other MFBO settings. More importantly, theoretical investigation needs to uncover the reason why we observere MFBO underperformance in Figure 3. We consider the following scenarios:

- **Lack of standardisation:** For a fair comparison, we will need to establish agreed standards and implementation strategies. HPOBench (Eggensperger et al. (2021)) attempts to establish such standard and partially succeed, also showcasing the significant impact a lack of standardisation can have. In Figure 4 in their paper, the upper row shows the non-standardise benchmark run, where single-fidelity methods have no clear trend. The bottom row instead, standardised to the benchmark, shows a clear trend of single-fidelity methods possibly outperforming multi-fidelity methods if given enough budget (i.e. the central question of our paper).

- **Application variety:** The problem structure, whether real-world or test function, presumably has an impact on MFBO performance. Indeed, (Mikkola et al. (2023)) study that exact question and in their Figure 1 showcase how unreliable fidelities render MFBO inferior to SFBO. Our empirical study, although trying to replicate their study of informative auxiliary information sources, might not be the best representation of problem spaces that are challenging to MFBO, and so we hope to expand our study beyond into a wider variety of test functions.

- **No-free-lunch theorem:** Given the significant outperformance of MFBO in the short-term over SFBO, it is reasonable to assume on an intuitive level, that long-term MFBO will suffer from inefficiencies it traded for superior efficiency in the short-term. We hope to explore this idea on a theoretical basis as a next step.

- **Compounding errors:** Considering that lower fidelities are "noisier" than higher fidelities, it is possible that the errors in measurements when using the lowest fidelities accumulates, leading the optimization process to get stuck in a local minima in the long-run.

5. Conclusion

We studied long-run behaviour of multi-fidelity Bayesian Optimisation (MBFO), observing in the literature a possible under-performance compared to single-fidelity Bayesian optimisation (SFBO). Our own empirical studies provide further evidence on these observations.

With a multitude of MFBO algorithms available [1], it is important to evaluate their limitations for the best outcome in applications of adaptive experimentation. We discussed a few possible scenarios and hope to expand our empirical studies to characterise the long-run behaviour of MFBO for a variety of applications beyond the test functions studied so far.

Acknowledgments

Thank you to the workshop reviewers for their detailed feedback that allowed us to improve and clarify our work. We are also grateful for the support for this work from the UKRI Innovate UK Transformative Technologies Grant 2023 Series.

References

Raul Astudillo and Peter I Frazier. Thinking inside the box: A tutorial on grey-box bayesian optimization. In *2021 Winter Simulation Conference (WSC)*, pages 1–15. IEEE, 2021.

Noor Awad, Neeratyoy Mallik, and Frank Hutter. Dehb: Evolutionary hyperband for scalable, robust and efficient hyperparameter optimization, 2021.

Balandat. Continuous multi-fidelity bo in botorch with knowledge gradient, 2021. URL `https://botorch.org/tutorials/multi_fidelity_bo`.

Hugo Bellamy, Abbi Abdel Rehim, Oghenejokpeme I Orhobor, and Ross King. Batched bayesian optimization for drug design in noisy environments. *Journal of Chemical Information and Modeling*, 62(17):3970–3981, 2022.

Alexander I. Cowen-Rivers, Wenlong Lyu, Rasul Tutunov, Zhi Wang, Antoine Grosnit, Ryan Rhys Griffiths, Alexandre Max Maraval, Hao Jianye, Jun Wang, Jan Peters, and Haitham Bou Ammar. Hebo pushing the limits of sample-efficient hyperparameter optimisation, 2022.

Samuel Daulton, Maximilian Balandat, and Eytan Bakshy. Differentiable expected hypervolume improvement for parallel multi-objective bayesian optimization, 2020.

Katharina Eggensperger, Philipp Müller, Neeratyoy Mallik, Matthias Feurer, René Sass, Aaron Klein, Noor Awad, Marius Lindauer, and Frank Hutter. Hpobench: A collection of reproducible multi-fidelity benchmark problems for hpo. *arXiv preprint arXiv:2109.06716*, 2021.

Peter I. Frazier. A tutorial on bayesian optimization, 2018.

Deng Huang, Theodore T Allen, William I Notz, and R Allen Miller. Sequential kriging optimization using multiple-fidelity evaluations. *Structural and Multidisciplinary Optimization*, 32:369–382, 2006.

Faran Irshad, Stefan Karsch, and Andreas Döpp. Leveraging trust for joint multi-objective and multi-fidelity optimization, 2023.

1. e.g. see Matterhorn Studio's OptStore for an overview: `https://matterhorn.studio`

Kirthevasan Kandasamy, Karun Raju Vysyaraju, Willie Neiswanger, Biswajit Paria, Christopher R Collins, Jeff Schneider, Barnabas Poczos, and Eric P Xing. Tuning hyperparameters without grad students: Scalable and robust bayesian optimisation with dragonfly. *The Journal of Machine Learning Research*, 21(1):3098–3124, 2020.

Qiaohao Liang, Aldair E Gongora, Zekun Ren, Armi Tiihonen, Zhe Liu, Shijing Sun, James R Deneault, Daniil Bash, Flore Mekki-Berrada, Saif A Khan, et al. Benchmarking the performance of bayesian optimization across multiple experimental materials science domains. *npj Computational Materials*, 7(1):188, 2021.

Petrus Mikkola, Julien Martinelli, Louis Filstroff, and Samuel Kaski. Multi-fidelity bayesian optimization with unreliable information sources. In *International Conference on Artificial Intelligence and Statistics*, pages 7425–7454. PMLR, 2023.

Matthias Poloczek, Jialei Wang, and Peter I. Frazier. Multi-information source optimization, 2016.

Matthias Poloczek, Jialei Wang, and Peter Frazier. Multi-information source optimization. *Advances in neural information processing systems*, 30, 2017.

Shion Takeno, Hitoshi Fukuoka, Yuhki Tsukada, Toshiyuki Koyama, Motoki Shiga, Ichiro Takeuchi, and Masayuki Karasuyama. Multi-fidelity bayesian optimization with max-value entropy search and its parallelization, 2020.