Carola DOERR
CNRS researcher (HDR)
Sorbonne Université, Paris, France
carola.doerr@lip6.fr

Czech Technical University in Prague
Faculty of Nuclear Sciences and Physical Engineering
Prague, Czech Republic

Subject: Review of the PhD thesis *Surrogate Modeling and Landscape Analysis for Evolutionary Black-box Optimization,* submitted by Zbyněk PITRA

Paris, May 20, 2022

Dear colleagues,

It is with great pleasure that I write this review of the PhD thesis submitted by Zbyněk PITRA on *Surrogate Modeling and Landscape Analysis for Evolutionary Black-box Optimization* to express my strong support for the admission of Zbyněk Pitra to defend this thesis.

SHORT SUMMARY

The thesis investigates efficient methods for optimizing numerical black-box problems with small budgets. Such approaches are extensively needed in situations in which the comparison of possible solutions require expensive evaluations, e.g., via computer simulations or physical experiments. The key contributions of this thesis can be categorized into two main subjects:

1. Two surrogate-assisted variants of the Covariance Matrix Adaptation Evolution Strategy (CMA-ES), a state-of-the-art algorithm for derivative-free numerical black-box optimization, are suggested and empirically evaluated on the BBOB function suite of the COCO benchmark environment as well as on a set of simulation-based problem instances modelling submerged wave energy converters.
   The suggested algorithms extend previous surrogate-assisted CMA-ES variants, which were studied in the literature since the early 2000s. An impressive number of different combinations of surrogate models, algorithmic components, and hyper-parameter settings is studied in this thesis.
2. Since neither one of the algorithm instances investigated in the context of the above-mentioned comparisons dominates all others on all tested problem instances, the PhD candidate set out to investigate the relationship between problem characteristics, measured by numerical features, and the suitability of different surrogate models to predict the quality of solution candidates. In this context, a new set of features is suggested, the use of state variables of the optimization algorithm itself.

CHRONOLOGICAL SUMMARY

Chapter 1 motivates the research, summarizes the key findings of the thesis, and presents the thesis structure.

Chapter 2 provides necessary background on black-box optimization, evolutionary computation, surrogate modelling, other black-box optimization algorithms, exploratory fitness landscape analysis, benchmarking, and performance evaluation. The chapter also puts the thesis into context, by summarizing related work.

Chapter 3 is only half a page, and summarizes the key objectives of the thesis: the development of efficient surrogate-assisted CMA-ES variants, and the investigation of how fitness landscape features correlate with the accuracy of surrogate models.

Chapter 4 is the heart of the thesis. It presents the surrogate-assisted CMA-ES algorithms developed by the PhD candidate and his co-authors. It is shown that surrogate models can greatly improve the performance of CMA-ES, especially when the budget is relatively small. The two core algorithms, S-CMA-ES and DTS-CMA-ES, differ in the frequency at which the surrogate models are used: A generational approach is used for S-CMA-ES; that is, whole generations are evaluated using the true objective function, before alternating again with evaluations based on the surrogate model. The Doubly Trained Surrogate CMA-ES (DTS-CMA-ES), in contrast, evaluates only a certain fraction of the solution candidates suggested by CMA-ES via the original objective function – namely those that seem most valuable according to some criterion that is computed on top of the model-based evaluations. After their evaluation, a second surrogate, now taking into account the newly evaluated points, hence the name *doubly trained*.
The results suggest that the DTS-CMA-ES variant with Gaussian Process models are particularly useful for the optimization of multimodal problems with weak global structure, whereas its self-adaptive version provides very promising results for globally decreasing multimodal functions. Another conclusion that the PhD candidate draws from his results is that the selection of training sets via *k* nearest neighbors performed best. Several selection strategies are tested throughout this thesis. Regarding the Gaussian Process models, it is concluded that original metric regression models outperform ordinal models. Finally, Gaussian Process models are shown to provide more accurate predictions than the models based on random forests. However, at the same time, it is found that Random Forest models reduce the number of fitness evaluations required by CMA-ES to hit a given quality threshold more drastically than Gaussian Processes.
All in all, we learn from this chapter that a hybridization of CMA-ES (and probably also other black-box optimization algorithms) with surrogate modeling is a highly promising approach to reduce the number of objective function evaluations needed to identify high-quality solutions for expensive-to-evaluate optimization problems. We also learn that there are many different ways of building surrogate-assisted algorithms, and that there are no clearly identifiable general recommendations that can be made about the best setup.

Motivated by the results presented in Chapter 4, Chapter 5 investigates the relationship between so-called fitness landscape features (i.e., numerical representations of the problem instances that aim at capturing some of their key characteristics) and the accuracy of the surrogate models. It is found that the feature description can largely influence the model accuracy. However, it is not straightforward to use the features to discriminate between the model accuracy of Gaussian Processes and Random Forrest models, respectively. A new set of features is also introduced in this chapter. Unlike the classically used exploratory landscape features, the PhD candidate suggests to use the state variables of the CMA-ES and their evolution to numerically characterize the optimization process. That is, instead of only looking at the problem instance itself, this approach suggests to exploit also information about the algorithmic optimization behavior for a more suitable selection of surrogate models. This is a very interesting idea which has already found applications in follow-up works.

Finally, Chapter 6 summarized the thesis and its key findings, in a more detailed fashion than Chapter 1. A number of promising directions for future work are also provided. This part also motivates again the long-term vision of the research presented in this thesis: an adaptive selection of the surrogate models and their precise settings, the optimal interplay between model-based evaluations and those using the original objective function, etc. – ideally using some form of meta-learning on top of the landscape and state variable features. The hybridization of the techniques investigated in this thesis with other base algorithms or other CMA-ES configurations is also mentioned.

ASSESSMENT

- **Choice of research topic and timeliness:** The thesis addresses an important problem in numerical black-box optimization, the low-budget case. This setting is very timely. With the increased accuracy of simulation models in various domains such as biology, mechanical engineering, and many others, we face more and more optimization scenarios in which the evaluation of a single solution candidate is computationally expensive. Optimization approaches that recommend high-quality solutions after only few such evaluation are therefore at high demand. The DTS-CMA-ES makes an important contribution to this domain, by exploiting recent advances in surrogate modelling with the state-of-the-art optimization algorithm CMA-ES. The empirical comparisons offered in this thesis convincingly demonstrate the benefits of DTS-CMA-ES in several optimization scenarios. The PhD candidate is very honest about the (inevitable) complementarity of different surrogate techniques. The investigation of landscape-aware selection and configuration techniques is a very natural and highly promising approach to exploit this complementarity. The thesis provides an impressive number of experiments, all available open source, and adheres to good scientific practices.

- **Contributions:** The contributions are original and innovative. These results presented in this thesis have been published in leading conferences in the evolutionary computation and learning domain, including the *ACM Genetic and Evolutionary Computation Conference* (GECCO), *Parallel Problem Solving from Nature* (PPSN), *Learning and Intelligent Optimization* (LION), in the *Evolutionary Computation* journal (MIT press), and at the *Workshop on Interactive Adaptive Learning*, co-located with the *European Conference on Machine Learning* (ECML).
  The thesis convincingly shows the great potential of the hybridization between model-based optimization and more classical evolution strategies and other derivative-free optimization approaches. The large number of experiments as well as the discussions provide a good starting point for future research. When reaching Chapter 5 about the relationship between features and model accuracy, one may be left with the feeling that this part is somewhat incomplete, as the natural continuation – the dynamic use of the suggested feature sets for model selection – is not further investigated. But this it is very natural for a good PhD thesis to leave directions for future work, and this thesis presents a large number of them. The key ideas underlying the Doubly Trained Surrogate CMA-ES as well as the idea to use state variables as features to describe the problem instance are novel, and carry a great potential for future investigations.

- **Presentation:** The thesis is exceptionally well written, with a nice flow of arguments, a good story line, and a detailed justification for the various choices that the PhD candidate faced during his research. Minor exceptions to this overall very positive reading experience are the part about Bayesian Optimization in Section 2.4. Since the algorithms mentioned here are components of the surrogate-based solvers presented in 2.3.3, it would have been better to swap the order of presentation. Another possibly weak spot is the summary of key goals in Section 3, which lacks detail. For example, it is not specified for which optimization problems (which applications, which budgets, which dimensions) surrogate-assisted CMA-ES variants shall be developed. For the fitness landscape analysis, it is not specified how the "suitability" of the different models is evaluated, nor what the long-term vision of this research is. Here, one could have mentioned fitness-aware algorithm configuration and selection as a goal. A common critique of research in this domain is an overemphasis of "academic" benchmark problems, such as the BBOB functions that are used in this thesis for algorithm comparison. While the thesis also presents results for a benchmark collection that aims at simulating a real-world problem, the choice of the BBOB suite could have benefitted from a more explicit motivation. Finally, I somewhat regret the non-explicit mention of what happens with the second surrogate model in the Doubly Trained Surrogate CMA-ES in the short summary of this algorithm in Section 4.4. Given that this is a key contribution of the thesis, it would have been preferable to have a more complete description of the algorithm in this part.

RECOMMENDATION

Based on the presented thesis, I have no doubts that Zbyněk PITRA is able to perform rigorous scientific work and that he will be able to defend his thesis successfully. I therefore strongly suggest that he is admitted to defend his work.

Please note that I have shared a scan with minor comments and typos directly with the PhD candidate.

I remain at your disposal for any further information and thank you for the opportunity to provide my assessment for this very interesting PhD thesis, which I very much enjoyed reading.

Sincerely

Carola DOERR