Adaptive Selection of Gaussian Process Model for Active Learning in Expensive Optimization

Jakub Repický [1,2], Zbyněk Pitra [1,3], and Martin Holeňa [1]

[1] Institute of Computer Science, Czech Academy of Sciences, repicky@cs.cas.cz
[2] Faculty of Mathematics and Physics, Charles University
[3] Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University

1 Introduction

Black-box optimization denotes the optimization of objective functions the values of which are only available through empirical measurements or experiments. Such optimization tasks are most often tackled with evolutionary algorithms and other kinds of metaheuristics methods (e. g., [12]), which need to evaluate the objective function in many points. This is a serious problem in situations when its evaluation is *expensive* with respect to some kind of resources, e.g., the cost of needed experiments.

A standard attempt to circumvent that problem is to evaluate the original objective function only in a small fraction of those points, and to evaluate a *surrogate model* of the original function [5] in the remaining points. Once a model has been trained, the success of the optimization in the remaining iterations depends on a *resource aware selection* of points in which the original function will be evaluated, which is a typical *active learning task*.

The surrogate model used in the reported research is a *Gaussian process* (GP) [11], which treats the values of an unknown function as jointly Gaussian random variables. The advantage of GP compared to other kinds of surrogate models is its capability of quantifying the uncertainty of prediction, by calculating the variance of the posterior distribution of function values.

2 Novel Approach to GP-Based Active Learning

So far, active learning of points in which the original objective function will be evaluated during a GP-based surrogate modelling nearly always used a fixed covariance function of the surrogate model and a fixed active learning algorithm (e.g., [9, 15]).

In the reported research, an *adaptive selection of the covariance function* according to the available data is investigated.

To this end, a pool of kernels of 8 various kinds has been established, including non-stationary kernels and composed kernels of depth one (Figure 1).

Adaptive selection of GP covariance functions is known from GP regression literature [3, 6, 8]. Differently to our approach, the number of available kinds of simple kernels is smaller; on the other hand, they can be recurrently composed to an arbitrary depth through algebraic operations. An additional difference concerns the criteria according to which the models are selected: Whereas in [6], only likelihood is used, and in [3, 8], only the Bayes information criterion (*BIC*) [13] was used, our research includes the investigation of suitability of different criteria for the selection of GP covariance functions in surrogate modelling. Since the maximum likelihood estimate is prone to favour overfitting models, we consider only criteria that account for model complexity. Apart from the BIC used in [3, 8], we consider also the Akaike information criterion (*AIC*) [1], the deviance information criterion (*DIC*) [14], and a criterion proposed by Watanabe in [16], called by him widely applicable information criterion (*WAIC*).

2.1 Algorithm of Active Learning

The adaptive selection of covariance function is based on GP-based Doubly trained surrogate covariance matrix adaptation evolutionary strategy (DTS-CMA-ES) [2, 10]. A GP model is built in each generation of the evolution strategy. When the model is trained, a fraction of the population maximizing the probability of improvement [7] is selected for evaluation with the expensive objective function. The model is retrained with the newly evaluated points and used to rank the whole population. The fraction of points for evaluation is adapted according to recent performance of the surrogate model, more precisely, to its ranking difference error in the last iteration. A novel contribution of the reported research is a selection of the covariance function according to one of the aforementioned criteria, taking place at the training phase.

3 Results and Conclusion

The best performing model selection in our experiments are those based on AIC and BIC, with no clear difference between them. Results for DIC and WAIC are not shown because we have not yet succeed to implement these.

On average, optimization results do not show an improvement on the DTS-CMA-ES. Nevertheless, a promising result has been obtained on the “Step ellipsoidal” function, characterized by plateaus lying on a quadratic structure, especially in multi-dimensional variants (Figure 2). The most frequently selected kernel for this function under both AIC and BIC has been the sum of the squared exponential and the quadratic kernel, which provides an intuitive interpretation of the function.

This result suggests that composite kernels can capture global features of the objective function landscape and provide better predictive performance, but extending the pool of covariances might be needed e. g., by composite kernels of arbitrary depth. This is challenging due the computational cost of searching through an open-ended space of kernel expressions. A possible solution might be found in a co-evolution of the kernel for the surrogate model and the candidate solution to the objective function.

Acknowledgment

The research reported in this paper has been supported by the SVV project number 260 453 and the Czech Science Foundation (GAČR) grant 17-01251.

References

1. Akaike, H.: Information Theory and an Extension of the Maximum Likelihood Principle, pp. 199–213. Springer New York (1973)
2. Bajer, L.: Model-based evolutionary optimization methods. Ph.D. thesis, Faculty of Mathematics and Physics, Charles University in Prague, Prague (2018)
3. Duvenaud, D., Lloyd, J., Grosse, R., Tenenbaum, J., Zoubin, G.: Structure discovery in nonparametric regression through compositional kernel search. In: Dasgupta, S., McAllester, D. (eds.) Proceedings of the 30th International Conference on Machine Learning. vol. 28, pp. 1166–1174. PMLR, Atlanta, Georgia, USA (Jun 2013)
4. Hansen, N., Finck, S., Ros, R., Auger, A.: Real-parameter Black-Box Optimization Benchmarking 2012: Experimental setup. Tech. rep., INRIA (2012)
5. Jin, Y.: Surrogate-assisted evolutionary computation: Recent advances and future challenges. Swarm and Evolutionary Computation **1**(2), 61–70 (Jun 2011)
6. Kronberger, G., Kommenda, M.: Evolution of covariance functions for gaussian process regression using genetic programming. In: Moreno-Díaz, R., Pichler, F., Quesada-Arencibia, A. (eds.) Computer Aided Systems Theory - EUROCAST 2013. pp. 308–315. Springer Berlin Heidelberg, Berlin, Heidelberg (2013)
7. Kushner, H.J.: A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. J. Basic Eng. **86**(1), 97–106 (1964)
8. Lloyd, J.R., Duvenaud, D., Grosse, R., Tenenbaum, J.B., Ghahramani, Z.: Automatic construction and natural-language description of nonparametric regression models. CoRR **abs/1402.4304** (Apr 2014)
9. Lu, J., Li, B., Jin, Y.: An evolution strategy assisted by an ensemble of local Gaussian process models. In: Proceeding of the fifteenth annual conference on Genetic and evolutionary computation conference - GECCO '13. ACM (2013)
10. Pitra, Z., Bajer, L., Holeňa, M.: Doubly trained evolution control for the surrogate CMA-ES. In: Parallel Problem Solving from Nature – PPSN XIV, pp. 59–68. Springer International Publishing (2016)
11. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. Adaptative computation and machine learning series, MIT Press (2006)
12. Schaefer, R.: Foundations of Global Genetic Optimization. Springer Publishing Company, Incorporated, 1st edn. (2007)
13. Schwarz, G.: Estimating the dimension of a model. The Annals of Statistics **6**(2), 461–464 (1978)
14. Spiegelhalter, D., Best, N., Carlin, B., Van Der Linde, A.: Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society. Series B: Statistical Methodology **64**(4), 583–616 (12 2002)
15. Volz, V., Rudolph, G., Naujoks, B.: Investigating Uncertainty Propagation in Surrogate-assisted Evolutionary Algorithms. In: Proceedings of the Genetic and Evolutionary Computation Conference. pp. 881–888. GECCO '17, ACM, New York (2017)
16. Watanabe, S.: Algebraic Geometry and Statistical Learning Theory. Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press (2009)

A Appendix

A.1 Covariance Functions

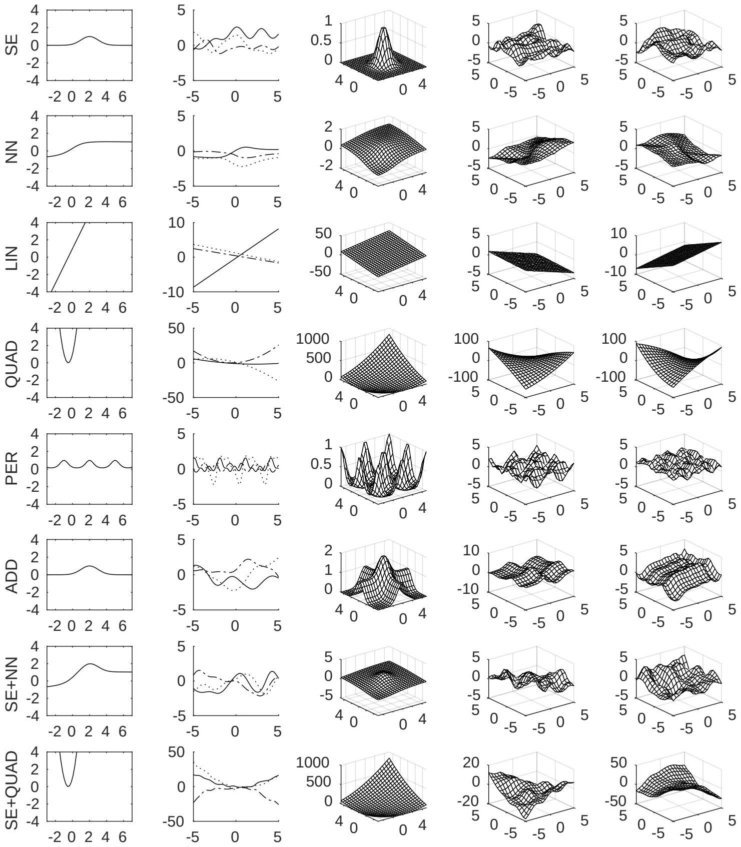

Fig. 1. Rows: Covariance functions. SE: Squared exponential. NN: neural network. LIN: linear. QUAD: quadratic. PER: periodic. ADD: additive. SE+NN: sum of squared exponential and neural network. SE+QUAD: sum of squared exponential and quadratic. Columns 1–2: The covariance function on $\mathbb{R}$ centered at point 2 (Col. 1) and three independent samples from the GP (Col. 2). Columns 3–5: The covariance function on $\mathbb{R}^2$ centered at $[2\ 2]^T$ (Col. 3) and two independent samples from the GP (Col. 4 and 5).

A.2 Experimental Setup

The ongoing experiments are performed within the framework Comparing continuous optimisers [4], in particular on the 24 benchmark functions forming its noiseless testbed. Each function is defined everywhere on $\mathbb{R}^D$ and has its global optimum in $[-5,5]^D$ for all dimensionalities $D \geq 2$. For every function and two dimensionalities, $2D$ and $10D$, 15 independent trials of each algorithm are conducted on multiple function instances, defined by transformations (translations, rotations and shifts) of both the search space and f-values. A trial terminates when the optimum is reached within a small tolerance or when a given budget of function evaluations, $250D$ in our case, is exhausted.

A.3 Experimental Results

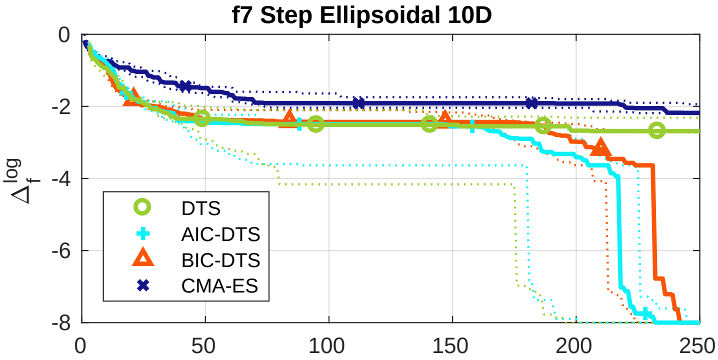

Fig. 2. Medians (solid) and 1st/3rd quartiles (dotted) of the distances to the optima against the number of function evaluations in $10D$ for all satisfactorily implemented algorithms. CMA-ES: Covariance matrix adaptation evolution strategy. DTS-CMA-ES: Doubly trained surrogate-CMA-ES. AIC-DTS, BIC-DTS: DTS-CMA-ES with the adaptive covariance selection according to AIC and BIC, respectively. The medians and quartiles were calculated from 15 independent runs on different function instances. Distances to optima are shown in the $\log_{10}$ scale.