Accounting for Imprecision of Model Specification in Bayesian Optimization *

Julian Rodemann J.RODEMANN@CAMPUS.LMU.DE
Thomas Augustin THOMAS.AUGUSTIN@STAT.UNI-MUENCHEN.DE
Institut für Statistik, Ludwig-Maximilians-Universität München (LMU), Munich, Germany

Motivation Bayesian optimization (BO) is a state-of-the-art method for optimizing[1] functions that are expensive to evaluate and lack analytical description. Its applications range from engineering to hyperparameter tuning in machine learning. BO approximates the objective function through a surrogate model, which in the case of real-valued covariates is a Gaussian process (GP). It then scalarizes the GP's mean and standard error estimates to an acquisition function (AF), that incorporates the trade-off between exploration (variance reduction) and exploitation (mean optimization). The arguments of the AF's maxima are eventually proposed to be evaluated. The question arises how sensitive BO's convergence is towards the prior of the GP, which is often specified arbitrarily due to the lack of knowledge about the function's nature.

Sensitivity Analysis As expensive functions imply few data, it comes at no surprise that the GP's predictions inside BO heavily depend on priors. Our results from optimizing 50 randomly selected benchmark functions suggest this translates to BO's convergence. It is especially sensitive towards the functional form of the kernel and the mean function's parameters.

Prior-Robust Bayesian Optimization In response to the latter, we propose two modifications of BO. Both rely on imprecise Gaussian processes [1]. They comprise a set of GP near ignorance priors $\left\{GP\left(Mh, k_\theta(x,x') + \frac{1+M}{c}\right) : h = \pm 1, M \geq 0\right\}$, given a base kernel $k_\theta(x,x')$ and a degree of imprecision c. This results in a set of posteriors, of which the upper and lower mean estimates $\underline{\mu}(x)_c$, $\overline{\mu}(x)_c$ can be derived analytically. In order to render BO more robust towards misspecification of the (constant) prior mean parameter, we deploy several $\underline{\mu}(x)_c$, $\overline{\mu}(x)_c$ for varying c as surrogate models either in parallel (1) or simultaneously (2). Approach (1) returns $2S+1$ optima for S imprecise surrogate models and the precise model. The $2S$ additionally proposed optima can be regarded as a hedge against getting stuck in a local optima in the regular process due to prior misspecification. Additionally, the budget of function evaluations can be divided into batches, after each of which the parallelization is broken and the worst performing surrogate models are thrown away. The remaining budget is then distributed among the remaining models to speed up convergence. Modification (2) explicitly accounts for the surrogate model's imprecision, i.e. the uncertainty with regard to the surrogate model's prior during the optimization procedure. For this purpose, we generalize the AF lower confidence bound $LCB(\mathbf{x}) = -\mu(\mathbf{x}) + \tau \cdot \sqrt{var(\mu(\mathbf{x}))}$ as follows: $GLCB(\mathbf{x}) = -\mu(\mathbf{x}) + \tau \cdot \sqrt{var(\mu(\mathbf{x}))} + \rho \cdot (\overline{\mu}(\mathbf{x})_c - \underline{\mu}(\mathbf{x})_c)$, where $\overline{\mu}(\mathbf{x})_c - \underline{\mu}(\mathbf{x})_c$ simplifies to an expression only dependent on the kernel vector, the kernel matrix and c. Hence, the surrogate models $\underline{\mu}(x)_c$ and $\overline{\mu}(x)_c$ do not have to be implemented and GLCB can be plugged into standard BO as an alternative AF without further ado and much computational cost. Just like LCB, the generalized LCB balances optimization of $\mu(x)$ and reduction of uncertainty with regard to the model's prediction variation $\sqrt{var(\mu(\mathbf{x}))}$ through τ. What is more, GLCB aims at reducing model imprecision caused by the prior specification, controllable by ρ. This would allow returning optima that are robust not only towards classical prediction uncertainty but also towards imprecision of the specified model.

Preliminary Results We compare the performance of prior-robust BO to classical BO with a budget of 90 evaluations and an initial design of size $n = 10$ on a univariate target function generated from a data set [2] describing the quality (to be maximized) of experimentally produced graphene, an allotrope of carbon. We find that parallel Prior-Robust BO (1) fails to beat classical BO, while the simultaneous approach using GLCB (2) systematically finds better configurations than LCB. Moreover, GLCB surpasses the popular AF expected improvement (EI) in late iterations (see attachment).

References

[1] Francesca Mangili. A prior near-ignorance Gaussian process model for nonparametric regression. In *ISIPTA '15: Proc. 9th Int. Symposium on Imprecise Probability: Theories and Applications*, pages 187–196, 2015.

[2] H. Wahab, V. Jain, A. S. Tyrrell, M. A. Seas, L. Kotthoff, and P. A. Johnson. Machine-learning-assisted fabrication: Bayesian optimization of laser-induced graphene patterning using in-situ raman analysis. *Carbon*, 167:609–619, 2020.

* *Acknowledgments* Julian Rodemann would like to thank the scholarship program of Evangelisches Studienwerk Villigst for the support of his studies.

1. Without loss of generality, consider minimization hereinafter, except for the data set in section Preliminary Results.