Knowledge-Based Systems 300 (2024) 112186

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

Imprecise Bayesian optimization

Julian Rodemann [*], Thomas Augustin

Department of Statistics, Ludwig-Maximilians-Universität München (LMU), Ludwigstraße 33, Munich, 80539, Bavaria, Germany

ARTICLE INFO

MSC:
62C10
62C12
62F15
90C26
90C31

Keywords:
Bayesian optimization
Gaussian process
Imprecise probabilities
Robust optimization
Prior near-ignorance
Regret bounds

ABSTRACT

Bayesian optimization (BO) with Gaussian processes (GPs) surrogate models is widely used to optimize analytically unknown and expensive-to-evaluate functions. In this paper, we propose a robust version of BO grounded in the theory of imprecise probabilities: Prior-mean-RObust Bayesian Optimization (PROBO). Our method is motivated by an empirical and theoretical analysis of the GP prior specifications' effect on BO's convergence. A thorough simulation study finds the prior's mean parameters to have the highest influence on BO's convergence among all prior components. We thus turn to this part of the prior GP in more detail. In particular, we prove regret bounds for BO under misspecification of GP prior's mean parameters. We show that sublinear regret bounds become linear under GP misspecification but stay sublinear if the misspecification-induced error is bounded by the variance of the GP. In response to these empirical and theoretical findings, we introduce PROBO as a univariate generalization of BO that avoids prior mean parameter misspecification. This is achieved by explicitly accounting for prior GP mean imprecision via a prior near-ignorance model. We deploy our approach on graphene production, a real-world optimization problem in materials science, and observe PROBO to converge faster than classical BO.[1,2]

1. Introduction: Law of decreasing flexibility?

In a thought-provoking essay for the *New Yorker*, Jonathan Zittrain argues that aiming for "answers first, explanations later" has become ubiquitous in machine learning [2]. He describes the *modus operandi* in machine learning research as discovering what works without knowing why it works, and then putting "that insight to use immediately, assuming that the underlying mechanism will be figured out later" [2]. The so-acquired burden of unexplained phenomena is dubbed "intellectual debt". Unlike in medical and other scientific areas, Zittrain argues, such theory-free advances are an intrinsic part of "statistical-correlation engines" in machine learning. He paints a bleak picture of machine learning's future: With a growing number of unknown mechanisms in complex systems, "the number of tests required to uncover untoward interactions must scale exponentially" [2].

When Zittrain wrote his essay in 2019, this was indeed considered a painful subject for machine learning research. Interpretable machine learning and causality had long been considered research niches, but not anymore. Both fields are rapidly growing. However, we argue the lack of interpretation and causal understanding is not the mere cause for increasing the intellectual credit line in machine learning, but also the hidden assumptions upon which a myriad of models rely. While influential with regard to the model's predictions, many assumptions are hardly questioned, let alone empirically tested.

In this work, we demonstrate the influence of unquestioned assumptions using Bayesian optimization (BO), a popular stochastic derivative-free optimization method, especially for hyperparameter tuning of machine learning models. We will outline how to make BO more robust against changing these assumptions. This requires representing partial or no knowledge about the model specification. The framework of imprecise probabilities (IP) offers a way to do this for Gaussian processes (GPs), a functional regression approach essential to Bayesian optimization. The main idea behind BO is to approximate the unknown objective function with a GP, referred to as a surrogate model, and optimize a transformation of it (e.g., a linear combination of mean and variance prediction) as a cost-effective proxy for the typically expensive target function.

Using imprecise Gaussian processes, we will account for a set of GPs as surrogate models, making the optimizer more robust against misspecification. Although models are often specified arbitrarily in practice, as seen in popular libraries like `spearmint`, `BOTorch` (Python),

[*] Corresponding author.
E-mail address: j.rodemann@lmu.de (J. Rodemann).

[1] **Open Science:** Implementations of PROBO and reproducible scripts for the experimental analysis as well as all reported data are available at: https://github.com/rodemann/imprecise-bayesian-optimization.

[2] Some parts of an earlier version of this work have been presented at the *Ninth International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making* (IUKM) and published in the corresponding proceedings [1], see Appendix D for details.

https://doi.org/10.1016/j.knosys.2024.112186
Received 22 December 2023; Received in revised form 12 June 2024; Accepted 21 June 2024
Available online 4 July 2024

and `mlr3MBO` (R), we will show that model choice greatly influences optimization performance. One (if not *the*) founding father of Bayesian optimization, Jonas Močkus, has proclaimed that "the development of some system of *a priori* distributions suitable for different classes of the function f is probably the most important problem in the application of [the] Bayesian approach to (...) global optimization" [3], cited after [4].

On the background of Zittrain's essay it shall be noted that BO is by far not the only stochastic derivative-free optimizer that heavily relies on probabilistic elements; be they advanced surrogate models or simple probability measures. Examples comprise, for instance, simulated annealing [5] or covariance matrix adaptive evolutionary search [6], see also Section 7.

We argue that the flexibility of the optimization path to capture global optima can be increased by relaxing the assumptions about the probabilistic elements by means of IP. Leaning on the famous quotation by Manski, "The credibility of inference decreases with the strength of the assumptions maintained" [7, page 1], it will be demonstrated for the example of BO that a relaxation of the assumptions can increase optimizers' modeling capacity and hence their performance, suggesting a "Law of Decreasing Flexibility":

> *The exploratory flexibility of Bayesian optimization decreases with the strength of the probabilistic prior assumptions maintained.*

As it will turn out in Section 5, the generality of IP models allow for more flexibility of BO through an additional exploratory dimension in the well-understood exploration–exploitation trade-off [8] in Bayesian optimization. This is in line with recent deliberations by [9], who suggest a decomposition of reducible (epistemic) uncertainty into "modeling uncertainty" and "approximation uncertainty", the latter relating to classical statistical estimation uncertainty. By exploring the domain of the to-be-optimized function, classical Bayesian optimization aims at the reduction of this latter approximation uncertainty. By explicitly accounting for modeling uncertainty by means of a prior-near ignorance model from IP, our extension of BO will also explore the function's domain to reduce this second type of reducible uncertainty. Somewhat counter-intuitive from a statistical perspective at first glance, weakening the modeling assumptions might help obtain better solutions.

This paper demonstrates both theoretically and empirically that weakening even a small part of the GP specification can improve BO's performance. We conduct a thorough simulation study of BO's behavior under different specifications for all GP prior components (mean functional form, mean function parameters, kernel functional form, kernel function parameters). We find the mean function parameters to be the most influential. This is why we focus on this prior component in more detail and leave the kernel untouched—contrary to recent work by [4,10–12].

We prove cumulative regret bounds for Bayesian optimization under misspecification of the GP's prior mean function parameters. Surprisingly, the regret bounds grow linearly in BO's iterations, as opposed to sublinearly when GP prior mean function parameters are correctly specified. However, if we bound the misspecification-induced error by the GP's variance, we can restore regret bounds that grow sublinearly in the iterations.

We further propose Prior-mean-RObust Bayesian Optimization (PROBO), which builds on imprecise Gaussian processes [13,14], see also [1]. PROBO accounts for a set of GP prior mean parameter specifications, making it more robust to model imprecision. This is incorporated by a novel acquisition function, the Generalized Lower Confidence Bound (GLCB). We apply our method to the problem of optimizing graphene production and observe it outperforms competing acquisition functions.

The remainder of the paper is structured as follows. In Section 2, we formally introduce Bayesian optimization, Gaussian processes and acquisition functions. The section also discusses convergence and optimality of BO and summarizes related work. Section 3 conducts a Bayesian sensitivity analysis of classical Bayesian optimization with Gaussian processes. As this section finds the prior's mean parameters to be the most influential prior component, we theoretically analyze the latter's effect on BO's regret bounds in Section 4 and – as a consequence – introduce PROBO in Section 5. Section 6 describes detailed experimental results from benchmarking PROBO to classical BO on graphene production, an open problem in materials science. We conclude by a brief discussion of our method and an outlook to future work in Section 7.

2. Background

2.1. Bayesian optimization

Bayesian optimization (BO) is arguably one of the most popular methods for optimizing functions that are expensive to evaluate and do not have any analytical description ("black-box-functions"). Its applications range from engineering [15] to drug discovery [16], COVID-19 detection [17] and cybersecurity [18]. BO's main popularity, however, stems from machine learning, where it has become one of the predominant hyperparameter optimizers [19] after the seminal work of [20]. BO approximates the unknown target function through a surrogate model. In the case of all covariates being real-valued, Gaussian Process (GP) regression is the most popular surrogate model, while random forests are usually preferred for categorical and mixed covariate spaces. BO scalarizes the surrogate model's mean and standard error estimates through a so-called acquisition function,[3] that incorporates the trade-off between exploration (uncertainty reduction) and exploitation (mean optimization). The arguments of the acquisition function's minima are eventually proposed to be evaluated. Algorithm 1 describes the basic procedure of Bayesian optimization applied on a problem of the sort:

$$\min_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x}), \tag{1}$$

where we observe

$$\Psi : \mathcal{X} \to \mathbb{R}, \boldsymbol{x} \mapsto f(\boldsymbol{x}) + \epsilon, \tag{2}$$

with $\mathcal{X}$ a p-dimensional covariate[4] space and ϵ an *i.i.d.* zero-mean real-valued random variable. That is, we observe a noisy version $\Psi(\boldsymbol{x})$ of continuous $f(\boldsymbol{x})$. Here and henceforth, minimization is considered without loss of generality. Our theoretical analysis of BO under GP misspecification in Section 4 will require the assumption that f is sampled from an unknown Gaussian process.

Algorithm 1 Bayesian Optimization

```
1: create an initial design $D = \{(\boldsymbol{x}^{(i)}, \Psi^{(i)})\}_{i=1,\dots,n_{init}}$ of size $n_{init}$
2: while termination criterion is not fulfilled do
3:     train a surrogate model (SM) on data $D$
4:     propose $\boldsymbol{x}^{new}$ that optimizes the acquisition function $AF(SM(\boldsymbol{x}))$
5:     evaluate $\Psi$ on $\boldsymbol{x}^{new}$ and update $D \leftarrow D \cup (\boldsymbol{x}^{new}, \Psi(\boldsymbol{x}^{new}))$
6: end while
7: return $\arg\min_{\boldsymbol{x} \in D} \Psi(\boldsymbol{x})$ and respective $\Psi(\arg\min_{\boldsymbol{x} \in D} \Psi(\boldsymbol{x}))$
```

Notably, line 4 imposes a new optimization problem, sometimes referred to as "auxiliary optimization". Compared to $\Psi(\boldsymbol{x})$, however,

[3] Also referred to as infill criterion.

[4] The nomenclature in the literature is not consistent with regard to $\mathcal{X}$. This comes at no surprise, since $\mathcal{X}$ indeed is a servant of two masters. On the one hand, it is the "input" or "feature" of an optimization problem. On the other hand, it is a "covariate" of a surrogate model. As the latter is of particular interest in this work, we stick with the latter.

$AF(SM(\boldsymbol{x}))$ is analytically traceable. It is a deterministic transformation of the surrogate model's mean and standard error predictions, which are given by line 3. Thus, evaluations are cheap and optima can be retrieved through naive algorithms, such as grid search, random search or the slightly more advanced focus search[5], all of which simply evaluate a huge number of points that lie dense in $\mathcal{X}$. Various termination criteria are conceivable with a pre-specified number of iterations being one of the most popular choices.[6]

2.2. Gaussian processes

As stated above, Gaussian Process (GP) regressions are the most common surrogate models in Bayesian optimization for continuous covariates. The main idea of functional regression based on GPs is to specify a Gaussian process *a priori* (a GP prior distribution), then observe data and eventually receive a posterior distribution over functions, from which inference is drawn, usually by mean and variance prediction. In more general terms, a GP is a stochastic process, i.e. a set of random variables, any finite collection of which has a joint normal distribution.

Definition 1 (*Gaussian Process Regression*)**.** A function $f(\boldsymbol{x})$ is said to be generated by a *Gaussian process* $\mathcal{GP}\left(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}')\right)$ if for any finite vector of data $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$, the associated vector of function values $\boldsymbol{f} = (f(\boldsymbol{x}_1), \dots, f(\boldsymbol{x}_n))$ has a multivariate Gaussian distribution: $\boldsymbol{f} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = m(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ is a mean vector and $\boldsymbol{\Sigma} = k((\boldsymbol{x}_1, \dots, \boldsymbol{x}_n), (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)')$ a covariance matrix.

Hence, Gaussian processes are fully specified by a mean function $m(\boldsymbol{x}) = \mathbb{E}[f(\boldsymbol{x})]$ and a kernel[7] $k(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}\left[\left(f(\boldsymbol{x}) - \mathbb{E}[f(\boldsymbol{x})]\right)\left(f(\boldsymbol{x}') - \mathbb{E}[f(\boldsymbol{x}')]\right)\right]$ such that $f(\boldsymbol{x}) \sim \mathcal{GP}\left(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}')\right)$, see e.g. [22, page 13]. The mean function gives the trend of the functions drawn from the GP and can be regarded as the best (constant, linear, quadratic, cubic etc.) approximation of the GP functions. The kernel gives the covariance between any two function values and thus, broadly speaking, determines the function's smoothness and periodicity. Any polynomial function can serve as mean function. Any finitely positive semi-definite function (Definition 2) is a kernel function of a GP evaluated on a (finite) input vector.

Definition 2 (*Finitely Positive Semi-Definite Functions*)**.** A function $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is finitely positive semi-definite if it is symmetric ($\forall \boldsymbol{x}, \boldsymbol{z} \in \mathcal{X} : f(\boldsymbol{x}, \boldsymbol{z}) = f(\boldsymbol{z}, \boldsymbol{x})$) and the matrix $\boldsymbol{K}$ formed by applying f to any finite subset of $\mathcal{X}$ is positive semi-definite, i.e. for its quadratic form it holds $\boldsymbol{x}'\boldsymbol{K}\boldsymbol{x} \geq 0 \; \forall \boldsymbol{x} \in \mathcal{X}$.

A kernel is said to be isotropic if it is a function of the distance $\|\boldsymbol{x} - \boldsymbol{x}'\|$, conditioned on a norm, mostly the L2-Norm. Popular isotropic kernel families are listed in Appendix C.

Conclusively, both mean and kernel function consist of a functional form and parameters, both of which has to be specified beforehand. The effect of these four components on the BO will be assessed in Section 3. For the theoretical analysis in Section 4, we also need a popular representation of kernel functions: Reproducing kernel Hilbert spaces, which are defined as follows, where positive definite kernels serve as reproducing kernels, see [23,24] for instance.

Definition 3 (*Reproducing Kernel Hilbert Space*)**.** Let $\mathcal{X}$ be a nonempty set and k be a positive definite kernel on $\mathcal{X}$. A Hilbert space $\mathcal{H}_k$ of functions on $\mathcal{X}$ equipped with an inner-product $\langle \cdot, \cdot \rangle_{\mathcal{H}_k}$ is called a reproducing kernel Hilbert space (RKHS) with reproducing kernel k, if $\forall \boldsymbol{x} \in \mathcal{X}$: $k(\cdot, \boldsymbol{x}) \in \mathcal{H}_k$ and $\forall \boldsymbol{x} \in \mathcal{X} \; \forall f \in \mathcal{H}_k$:

$$f(\boldsymbol{x}) = \langle f, k(\cdot, \boldsymbol{x}) \rangle_{\mathcal{H}_k}.$$

GP's popularity is mainly due to the fact that its posterior distribution has a closed-form expression: For a noisy sample $\boldsymbol{y}_T = \left[y_1 \dots y_T\right]'$ at $\left\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_T\right\}$, $y_t = f\left(\boldsymbol{x}_t\right) + \epsilon_t$ with $\epsilon_t \overset{i.i.d.}{\sim} N\left(0, \sigma^2\right)$ Gaussian noise, and a zero mean prior $\mathcal{GP}\left(0, k(\boldsymbol{x}, \boldsymbol{x}')\right)$, the posterior over $\boldsymbol{f}$ is a GP distribution again, with mean $\mu_T(\boldsymbol{x})$, covariance $k_T\left(\boldsymbol{x}, \boldsymbol{x}'\right)$ and variance $\sigma_T^2(\boldsymbol{x})$:

$$\begin{aligned} \mu_T(\boldsymbol{x}) &= \boldsymbol{k}_T(\boldsymbol{x})'\left(\boldsymbol{K}_T + \sigma^2 \boldsymbol{I}\right)^{-1} \boldsymbol{y}_T, \\ k_T\left(\boldsymbol{x}, \boldsymbol{x}'\right) &= k_\theta\left(\boldsymbol{x}, \boldsymbol{x}'\right) - \boldsymbol{k}_T(\boldsymbol{x})'\left(\boldsymbol{K}_T + \sigma^2 \boldsymbol{I}\right)^{-1} \boldsymbol{k}_T\left(\boldsymbol{x}'\right), \\ \sigma_T^2(\boldsymbol{x}) &= k_T(\boldsymbol{x}, \boldsymbol{x}), \end{aligned} \tag{3}$$

where $\boldsymbol{k}_T(\boldsymbol{x}) = \left[k_\theta\left(\boldsymbol{x}_1, \boldsymbol{x}\right) \dots k_\theta\left(\boldsymbol{x}_T, \boldsymbol{x}\right)\right]'$ and $\boldsymbol{K}_T$ is the positive definite kernel matrix applied on $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_T\}$, see [22], for instance.

2.3. Acquisition functions

There exist several acquisition functions, among which expected improvement and lower confidence bound are the most popular. Their Definitions 4 and 5 are based on [25–27]. We start with the most fundamental criterion of selecting points, the probability of improvement. Therefore, let $\psi(\boldsymbol{x})$ be the surrogate model and Ψ_{min} the incumbent minimal function value. The probability of improvement (PI) of $\boldsymbol{x}$ is

$$PI(\boldsymbol{x}) = \mathbb{P}(\psi(\boldsymbol{x}) < \Psi_{min}), \tag{4}$$

where the probability measure $\mathbb{P}$ is with respect to $\psi(\boldsymbol{x})$. When using a Gaussian process as surrogate model, as assumed in what follows, the PI can be simplified. For each finite vector of function values $\Psi(\boldsymbol{x})$ we assume $\Psi(\boldsymbol{x}) \sim \mathcal{N}(\mu(\boldsymbol{x}), \mathrm{Var}(\boldsymbol{x}))$, where $\mu(\boldsymbol{x})$ is the mean function of Ψ at $\boldsymbol{x}$ and $\mathrm{Var}(\boldsymbol{x})$ is the variance function at $\boldsymbol{x}$. For our surrogate model $\psi(\boldsymbol{x})$ it is $\psi(\boldsymbol{x}) \sim \mathcal{N}\left(\hat{\mu}(\boldsymbol{x}), \widehat{\mathrm{Var}}(\boldsymbol{x})\right)$, where $\hat{\mu}(\boldsymbol{x}), \widehat{\mathrm{Var}}(\boldsymbol{x})$ are estimates from the posterior GP, see Definition 1. Since the variance function is typically estimated by the variance of the mean prediction function $\hat{\mu}(\boldsymbol{x})$, we write $\widehat{\mathrm{Var}}(\boldsymbol{x}) = \mathrm{Var}(\hat{\mu}(\boldsymbol{x}))$.[8] This allows standardization of $\psi(\boldsymbol{x})$ and Ψ_{min} in $PI(\boldsymbol{x})$ as follows:

$$\mathbb{P}(\psi(\boldsymbol{x}) < \Psi_{min}) = \mathbb{P}\left(\frac{\psi(\boldsymbol{x}) - \hat{\mu}(\boldsymbol{x})}{\sqrt{\mathrm{Var}(\hat{\mu}(\boldsymbol{x}))}} < \frac{\Psi_{min} - \hat{\mu}(\boldsymbol{x})}{\sqrt{\mathrm{Var}(\hat{\mu}(\boldsymbol{x}))}}\right) = \boldsymbol{\Phi}\left(\frac{\Psi_{min} - \hat{\mu}(\boldsymbol{x})}{\sqrt{\mathrm{Var}(\hat{\mu}(\boldsymbol{x}))}}\right). \tag{5}$$

As convention dictates, $\boldsymbol{\Phi}$ denotes the standard normal distribution function. Since $\boldsymbol{\Phi}$, Ψ_{min}, $\hat{\mu}(\boldsymbol{x})$ and $\sqrt{\mathrm{Var}(\hat{\mu}(\boldsymbol{x}))}$ are given in line 4 of algorithm 1, it can be seen that $PI(\boldsymbol{x})$ is indeed computationally cheap to evaluate. It requires nothing but a simple function call with given arguments. Also note that the probability of improvement is 0 for already visited points, as for such points $\sqrt{\mathrm{Var}(\hat{\mu}(\boldsymbol{x}))} \to 0$ and $\Psi_{min} - \hat{\mu}(\boldsymbol{x}) \leq 0$, thus

$$\boldsymbol{\Phi}\left(\frac{\Psi_{min} - \hat{\mu}(\boldsymbol{x})}{\sqrt{\mathrm{Var}(\hat{\mu}(\boldsymbol{x}))}}\right) \to \boldsymbol{\Phi}(-\infty) = 0. \tag{6}$$

With the same line of reasoning it follows that the probability of improvement $PI(\boldsymbol{x})$ for $\{\boldsymbol{x} : \Psi_{min} - \hat{\mu}(\boldsymbol{x}) \leq 0\}$ (counter-intuitively)

[5] Focus search shrinks the search space and applies random search, see [21, page 7].

[6] BO's computational complexity depends on the SM. In case of GPs, it is $\mathcal{O}(n^3)$ due to the required inversion of the covariance matrix, where n is total number of target function evaluations.

[7] Also called covariance function or kernel function.

[8] GPs have the convenient property of intrinsically estimating the posterior variance of the prediction function $\hat{\mu}(\boldsymbol{x})$. In case of deploying random forests as surrogate models, additional bootstrap or jackknife-after-bootstrap is needed for variance estimation. For instance, jackknife-after-bootstrap is used in `mlrMBO` [21].

decreases with $\sqrt{\mathrm{Var}(\hat{\mu}(\boldsymbol{x}))}$. This makes the PI a very exploitative acquisition function. For a detailed theoretical analysis of the acquisition functions' impact on the trade-off between exploration and exploitation, see [8]. The most widely used acquisition function is the Expected Improvement (EI), which is closely related to PI.

Definition 4 (*Expected Improvement*)**.** Let $\psi(\boldsymbol{x})$ be the surrogate model and Ψ_{min} the incumbent minimal function value. The expected improvement of $\boldsymbol{x}$ is

$$EI(\boldsymbol{x}) = \mathbb{E}(\max\{\Psi_{min} - \psi(\boldsymbol{x}), 0\}).$$

This time, the improvement is bounded from below. Uncertainty estimates only enter if mean estimates imply real improvement. This prohibits the negative effect of increasing uncertainty for $\{\boldsymbol{x} : \Psi_{min} - \hat{\mu}(\boldsymbol{x}) \leq 0\}$ and, thus, enforces exploration. EI was proposed by [3, Pages 1-2], disguised as a utility function in a decision problem that captures the expected deviation from the extremum. It follows from this formulation that a point proposed according to expected improvement is Bayes-optimal in a given iteration. This early definition of BO with EI is very close to the modern formulation in Definition 4 and algorithm 1. However, it lacked the idea of surrogate modeling and thus the simplifications that come with Gaussian processes (GPs). Namely, we can express $EI(\boldsymbol{x})$ in this case in closed form in a similar manner to Eq. (5):

$$EI(\boldsymbol{x}) = (\Psi_{min} - \hat{\mu}(\boldsymbol{x}))\, \Phi\left(\frac{\Psi_{min} - \hat{\mu}(\boldsymbol{x})}{\sqrt{\mathrm{Var}(\hat{\mu}(\boldsymbol{x}))}}\right) + \sqrt{\mathrm{Var}(\hat{\mu}(\boldsymbol{x}))}\; \phi\left(\frac{\Psi_{min} - \hat{\mu}(\boldsymbol{x})}{\sqrt{\mathrm{Var}(\hat{\mu}(\boldsymbol{x}))}}\right), \tag{7}$$

which can be derived by partial integration from Definition 4, and where $\phi(\cdot)$ denotes the standard normal density function. Note that EI equals 0 for points that have already been visited, just like in case of PI. What is more, it can be seen that EI is a weighted sum of (standardized) mean and standard error estimates, thus explicitly balancing exploitation and exploration. While this follows naturally from the expected deviation from the extremum and the GP assumptions in case of $EI(\boldsymbol{x})$, the same trade-off can also be gracelessly encoded by a direct weighted sum of $\hat{\mu}(\boldsymbol{x})$ and $\sqrt{\mathrm{Var}(\boldsymbol{x})}$ with weight τ_t. The acquisition function Lower Confidence Bound (LCB) does the latter.

Definition 5 (*Lower Confidence Bound*)**.** Let $\hat{\mu}(\boldsymbol{x})$ and $\sqrt{\mathrm{Var}(\hat{\mu}(\boldsymbol{x}))}$ be the mean and standard error prediction functions of the surrogate model. The upper/lower[9] confidence[10] bound of $\boldsymbol{x}$ is

$$LCB(\boldsymbol{x}) = -\hat{\mu}(\boldsymbol{x}) + \tau_t \cdot \sqrt{\mathrm{Var}(\hat{\mu}(\boldsymbol{x}))}.$$

The LCB was initially proposed by [27]. Unlike in the case of EI and PI, the user can manually guide the exploration–exploitation trade-off by setting τ_t. Notably, τ_t can also be scheduled, e.g. decreased over time [29]. The idea is to explore $\mathcal{X}$ first, then exploit selected regions in detail later.

2.4. Related work

While there exists a vast amount of literature dealing with Bayesian optimization, merely a small fraction of it is explicitly concerned with robustness, not to mention model imprecision and robustness towards misspecification of the surrogate model.[11]

2.4.1. Robust Bayesian optimization

In a recent work [34], Makarova et al. address the issue of overfitting in tuning hyperparameters of machine learning models by BO. As parameters are typically optimized with regard to the training error, the (unknown) test error can increase with BO iterations while the training error (of the best incumbent configuration) still monotonically decreases. The authors show that cross-validation can mitigate this, but comes at high computational cost. As an alternative, they propose a regret-based stopping criterion loosely inspired by early stopping, a popular regularization technique in deep learning.

In statistics, quantile regression is a well-known alternative to mean regression. It is more robust against outliers in the response measurements than the standard linear model. [35] deploy quantile GP regression in BO. [36] show that Student-t processes are more flexible than Gaussian processes as prior over functions in a functional regression setting. They verify by simulation studies that Student-t processes are superior to Gaussian ones as surrogate models in Bayesian optimization on a wide range of problems. [37] propose a modification of Bayesian optimization that is robust towards distributional shifts of covariates, i.e., situations where the training data is sampled from a different distribution than the test data. [38] take a similar approach for the special case of Bayesian quadrature optimization, where the expectation of an expensive black-box integrand taken over a known probability distribution is maximized. [39] use stochastic policies (proposals) for data acquisition to handle input noise. They thus claim to render BO with regard to noisy covariates in a parallel optimization setting. With similar motivation for multi-criteria problems, [40] propose robust multi-objective Bayesian optimization under input noise. [41] introduce adversarially robust BO (ARBO) method suited to auto-tuning problems with time-invariant uncertainties that cannot be accounted for by small-scale noise term. Notably, using deep neural networks instead of Gaussian processes as surrogate models, which has gained popularity in recent years, has also been motivated by robustness arguments initially, see [42] for one of the earliest works on neural networks as surrogates in Bayesian optimization.

Optimizing more than one objective simultaneously can also be considered a form of robust extension of classical, single-objective Bayesian optimization. If the objectives are understood as different metrics for one and the same *latent* construct, the optimization will be more robust towards the choice of the latter's operationalization. Multi-objective Bayesian optimization (MOBO) has become a cornerstone technique for such scenarios. BO's founding father Jonas Močkus had already thought about multi-objective extensions of BO, see [43]. Later and more practical works were mainly inspired by multi-objective evolutionary algorithms, see [44,45], for instance. In this spirit, efficient multi-objective extensions of BO is proposed by [46,47]. These pioneering works have been extended by [48,49], who contributed to the theoretical understanding and practical implementation of acquisition functions in MOBO. As already mentioned above, [50] introduce an advanced framework that integrates deep learning with MOBO, enhancing its capability to handle high-dimensional data and complex objective landscapes. Furthermore, the work by [51] on predictive entropy search for multi-objective optimization addresses scenarios with expensive function evaluations, while scenarios with specifiable preferences over the objectives are dealt with by [52]. Finally, [53] should be mentioned who integrate MOBO with continuous evolutionary algorithms.

[9] The literature is not consistent with regard to this terminology.

[10] We are aware that in the context of Bayesian surrogate models such as GPs, *credible* confidence bound would be the more appropriate wording, see e.g. [28]. However, as the surrogate model can be any statistical model in general, we abstain from sticking to the specific terminology of Bayesian inference.

[11] The well-established field of robust optimization [30] deals with imprecise linear programming, where an analytical description of the target function – unlike in case of BO – exists. Further note that robustifying Gaussian processes is a vivid line of research itself, see [31–33] for instance, detached from its role in Bayesian optimization.

2.4.2. Surrogate model imprecision in Bayesian optimization

The specification of the surrogate model has been subject to research mainly from the perspective of how to incorporate expert knowledge in Bayesian optimization [54,55]. A recent simulation study has shed light on the importance of surrogate model specification in Bayesian optimization; the analysis by [56] mainly targets the effect of surrogate model calibration on BO performance. They find that well-calibrated models tend to perform better, i.e., achieve lower regrets. However, this correlation between BO performance and calibration is shown to diminish when controlling for the type of surrogate model, thus demonstrating that model choice is more relevant to BO performance than calibration within a certain model class. Apart from Bayesian optimization, there exist detailed empirical studies that analyze the impact of prior mean function and kernel on the posterior GP for a variety of real-world data sets, see [57] for a pioneering example. They typically show a strong dependence of posterior inference on the prior in case of small n. This is a finding that aligns with classical theoretical results from Bayesian inference.

Nevertheless, the robust approaches to BO mentioned in Section 2.4.1 do not tackle this issue of selecting and specifying the surrogate model. They certainly render BO more robust towards false confidence in its prediction due to unreliable data (underestimation of data uncertainty) or other factors. Yet, robustness towards misspecification of the surrogate model is not taken into consideration. The python library with the promising name `RoBO` (robust Bayesian optimization) has implementations that are robust against model misspecification only to the extent that the package provides implementations of different surrogate models and acquisition functions [58, page 2].

As far as we know, there are only a few clear exceptions. Firstly, [4] come up with a simple, yet particularly thrilling idea: “Automating Bayesian optimization with Bayesian optimization”. They suggest to optimize over a space of models in an inner loop nested inside BO. Just like in the outer loop, BO is used as an optimizer as proposed in [59]. The model space is defined by multiplication and addition of base kernels, see [60,61]. In other words, from the four components of the GP prior (see Definition 1) only the functional form of the kernel is varied, which will be found to be the second-most influential component in the Bayesian sensitivity analysis conducted in Section 3. Secondly, [11] address the kernel parameters, the least influential prior GP component in our sensitivity analysis. Their idea is appealing nevertheless: Kernel parameters are corrected in an empirical-Bayes manner by performing distance-based active learning simultaneously to Bayesian optimization. Further examples of adaptive kernel selection in BO comprise kernel selection motivated by few-shot learning [62] and a comparative study [63]. Notably, adaptive surrogate model selection has also been discussed beyond Gaussian processes, see e.g. [64] for an application to materials science. Very recently, [12] proposed to use an ensemble of Gaussian processes, varying both kernel parameters and functional form, from which surrogates are sampled via Thompson sampling. By exploiting parallel computing schemes, [12] manage to speed up convergence as opposed to using a single GP. As opposed to [4,11,12] and other mentioned work, we do not touch the kernel at all and only vary the mean function’s parameter(s) since they were found most influential in the Bayesian sensitivity analysis conducted in Section 3 and proven to dramatically increase the growth rate of regret bounds in Section 4. To the best of our belief, such an approach has not appeared in the literature so far.

In addition to these works that explicitly address model imprecision in BO, there is growing interest in utilizing conformal prediction to hedge against potentially misspecified models. The charming idea here is to obtain guarantees on the inference without loosing any sleep over the correct model. This is due to conformal prediction’s coverage guarantees that hold for misspecified models. For recent examples of conformalizing BO’s surrogate model(s), we refer to [65–67].[12] On the theoretical side, regret bounds for GP prior misspecification w.r.t. a norm in the RKHS are provided by [69], while [70] address misspecified likelihoods. Note that our results presented in Section 4 are concerned with prior-mean parameter misspecification specifically, neither touching the likelihood nor addressing misspecification in the function space.

The rapidly growing field of meta (or transfer) learning based BO deploys similar techniques to address a related, yet different problem. Here, it is typically assumed that—while no explicit prior knowledge exists on the problem at hand (i.e., on the target function $f(\boldsymbol{x})$, see Eq. (2)), there is knowledge on other problems from the same problem class. By empirical Bayesian estimation of the GP prior through data from these related functions, meta learning BO then aims to outperform classical BO that uses non-informative GP priors or estimates them similarly by empirical Bayes, but from the initial sample from $f(\boldsymbol{x})$ only. Under the relatively strong assumption of data being sampled from the exact same prior as $f(\boldsymbol{x})$, [71] show that for meta learning BO the sublinear regret bounds for LCB [72] shrink near zero and collapse to a constant proportional to the noise. While [71] and also [71] estimate both mean function and kernel from the offline data from other problems, [73] only consider the kernel. For further applications of meta/transfer learning based BO, we refer to [74–77] as well as to [78] for a recent survey on this emerging field.

2.4.3. Bayesian optimization in materials science

In Section 6, we will demonstrate the efficiency of our method for the real world use case of graphene production, a longstanding challenge in materials science. Bayesian optimization has extensively been used in engineering since the seminal work of [79] and to optimize material production, in particular. We refer to [80–82] for popular and recent examples as well as to [83] for an extensive survey on Bayesian optimization across multiple experimental materials science domains including a benchmarking analysis.

3. Bayesian sensitivity analysis

One might naturally wonder about the sensitivity of Bayesian optimization to the choice of priors in the Gaussian process [4,12,84]. It is widely recognized that traditional inference using Gaussian processes (GPs) can be particularly sensitive to the specification of priors when the sample size (n) is small [57]. With fewer data points, inference increasingly depends on prior information. This concern is especially pertinent in the context of Bayesian optimization, which is often applied to functions that are costly to evaluate, implying situations where data is strongly limited.

3.1. Experimental setup

In this section, we closely follow [84]. We systematically investigate to what extent this sensitivity translates to BO’s returned optima and convergence rates. To the best of our knowledge, it is the first systematic empirical assessment of GP prior’s influence on BO. Analyzing the effect on optima and convergence rates is closely related, yet different. Both viewpoints have weaknesses. Focusing on the returned optima means conditioning the analysis on the termination criterion; considering convergence rates requires the optimizer to converge in computationally feasible time. To avoid these downsides, we analyze the mean optimizations paths.

Definition 6 (*Mean Optimization Path*). Given R repetitions of Bayesian optimization applied on a test function $\Psi(\boldsymbol{x})$ with T iterations each, let $\Psi(\boldsymbol{x}^*)_{r,t}$ be the best incumbent target value at iteration $t \in \{1, \dots, T\}$ from experimental repetition $r \in \{1, \dots, R\}$. That is, for fixed experiment r, we define $\Psi(\boldsymbol{x}^*)_{r,t} := \min_{\boldsymbol{x} \in D_t} \Psi(\boldsymbol{x})$ with $D_t := \{(\boldsymbol{x}^{(i)}, \Psi^{(i)})\}_{i=1,\dots,n_{init+t}}$ analogous to D in algorithm 1. The elements

$$MOP_t = \frac{1}{R} \sum_{r=1}^{R} \Psi(\boldsymbol{x}^*)_{r,t}$$

[12] See also [68] for conformalized robust optimization.

shall then constitute the T-dimensional vector MOP, which we call *mean optimization path (MOP)* henceforth.

As follows from Definition 1, specifying a GP prior comes down to choosing a mean function and a kernel. Both are in turn determined by a functional form (e.g. linear trend and Gaussian kernel) and its parameters (e.g. intercept and slope for the linear trend and a smoothness parameter for the Gaussian kernel). Hence, we vary the GP prior with regard to the mean functional form $m(\cdot)$, the mean function parameters, the kernel functional form $k(\cdot,\cdot)$ and the kernel parameters (see Definition 1). We run the analysis on 50 well-established synthetic test functions from the R package `smoof` [85]. These are analytically defined functions with known optima that are used to benchmark optimizers, see [86] for instance. The functions are selected at random, stratified across the covariate space dimensions $1, 2, 3, 4$ and 7. For each of them, a sensitivity analysis is conducted with regard to each of the four prior components. The initial design (line 1 in algorithm 1) of size $n_{init} = 10$ is randomly sampled anew for each of the $R = 40$ BO repetitions with $T = 20$ iterations each. This way, we make sure the results do not depend on a specific initial sample. For each test function we obtain an accumulated difference (AD) of mean optimization paths.

Definition 7 (*Accumulated Difference of Mean Optimization Paths*)**.** Consider an experiment comparing S different prior specifications on a test function with R repetitions per specification and T iterations per repetition. Let the results be stored in a $T \times S$-matrix of mean optimization paths for iterations $t \in \{1, \dots, T\}$ and prior specification $s \in \{1, \dots, S\}$ (e.g. constant, linear, quadratic etc. trend as mean functional form) with entries $MOP_{t,s} = \frac{1}{R}\sum_{r=1}^{R} \Psi(\boldsymbol{x}^*)_{r,t,s}$. The *accumulated difference (AD)* for this experiment shall then be:

$$AD = \sum_{t=1}^{T} \left(\max_s MOP_{t,s} - \min_s MOP_{t,s} \right).$$

3.2. Results of sensitivity analysis

The AD values vary strongly across functions. This can be explained by varying levels of difficulty of the optimization problem, mainly influenced by modality and smoothness. Table 1 shows accumulated differences of mean optimization paths for selected test functions. Tables E.4 and E.5 in the appendix have the complete results. Figs. 1 and 2 visualize the mean optimization paths for BO on Ackley function and the function itself, respectively. Since we are interested in an overall, systematic assessment of the prior's influence on Bayesian optimization, we sum the AD values over the stratified sample of 50 functions. This absolute sum, however, is likely driven by hard-to-optimize functions with generally higher AD values or by the scale of the functions' target values.[13] Thus, we divide each AD value by the mean AD of the respective function. Table 2 shows the sums of these relative AD values. It becomes evident that the optimization is affected the most by the functional form of the kernel and the mean parameters, while kernel parameters and the mean functional form play a minor role.

3.3. Discussion of sensitivity analysis

Bayesian optimization typically deals with expensive-to-evaluate functions. As such functions imply the availability of few data, it comes at no surprise that the GP's predictions in BO heavily depend on the prior. Our results suggest this translates to BO's convergence. It is more sensitive towards the functional form of the kernel than towards those of the mean function and more sensitive towards the mean function's parameters than towards those of the kernel, which appear to play a negligible role in BO's convergence, see Table 2. Overall, the mean parameters appear to have the strongest impact on BO's convergence.

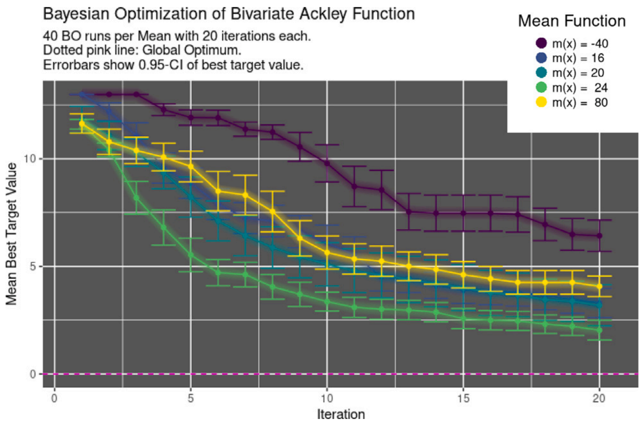


Fig. 1. Effect of Mean Function Parameters on Bayesian Optimization of Bivariate Ackley Function, see Fig. 2.

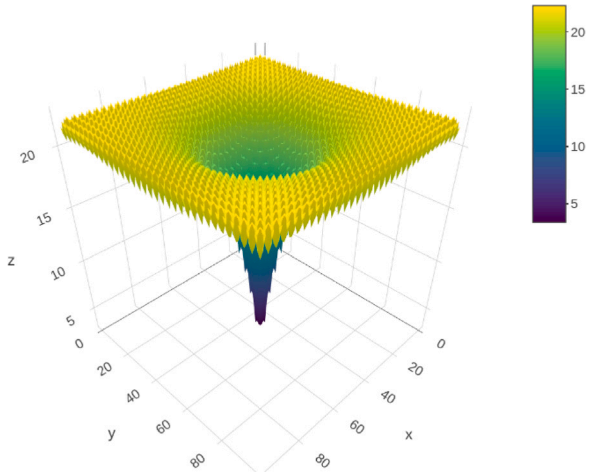


Fig. 2. Bivariate Ackley Function, see [85].

The kernel functional form determines the flexibility of the GP and thus has a strong effect on its capacity to model the functional relationship. What is more interesting, the mean parameters' effect may not only stem from the modeling capacity but also from the optimizational nature of the algorithm. While unintended in statistical modeling, a systematic under- or overestimation may be beneficial when facing an optimization problem. Further research on interpreting the effect of the GP prior's components on BO's performance is recommended.

Albeit the random sample of 50 test functions was drawn from a wide range of established benchmark functions, the analysis does by far not comprise all types of possible target functions, not to mention real-world optimization problems. Additionally, the presented findings regarding kernel and mean function parameters are influenced by the degree of variation, the latter being a subjective choice. Statements comparing the influence of the functional form with the parameters are thus to be treated with caution. Yet, the comparison between kernel and mean function parameters is found valid, as both have been altered by the same factors.

What weighs more, interaction effects between the four prior components were partly left to further research. The reported AD values for mean parameters and mean functional forms were computed using a Gaussian kernel. Since other kernels may interact differently with the mean function, the analysis was revisited using a power exponential kernel as well as a Matérn kernel. As we observe only small changes in

[13] Note that neither accumulated differences (Definition 7) nor mean optimization paths (Definition 6) are scale-invariant.

Table 1
Accumulated differences of mean optimization paths for Bayesian optimization of selected test functions from `smoof`. Please find complete results in Tables E.4 and E.5 in the Appendix.

Test function (Dimension of $\mathcal{X}$)	mean functional form	mean parameters	kernel functional form	kernel parameters
Ackley (1)	23	38	67	23
Cosine Mixture (1)	0.073	0.07	0.11	0.14
Six-Hump Camel Back (2)	1.3	3.3	2.9	0.71
Matyas (2)	0.28	0.59	2.7	0
Hartmann (3)	3	4.8	5.3	0.82
Alpine N. 2 (3)	14	25	30	4.6
Sum of Different Squares (4)	0.37	1.4	0.32	0
Bent-Cigar (4)	$3.6 \cdot 10^{9}$	$2 \cdot 10^{10}$	$7 \cdot 10^{9}$	$5.7 \cdot 10^{8}$
Deflected Corrugated Spring (7)	16	38	11	0
Sphere (7)	$1.5 \cdot 10^{2}$	$4.4 \cdot 10^{2}$	97	8.5

Table 2
Sum of relative ADs of all 50 MOPs per prior specification.

Mean functional form	Kernel functional form	Mean parameters	Kernel parameters
42.49	68.20	77.91	11.40

AD values, the sensitivity analysis can be seen as relatively robust in this regard, at least with respect to these three widely-used kernels.

4. Theoretical analysis

In light of the previous empirical results, we aim at a better understanding of the effect of prior mean parameter misspecification on BO's performance. We thus conduct a theoretical analysis of how prior mean parameter misspecification affects regret bounds of Bayesian optimization. Regret bounds are a well-established theoretical tool, with the help of which we can derive probabilistic guarantees for BO's performance. We will build on established regret bounds for Gaussian processes with (lower) confidence bound as acquisition function [72] which are still the tightest bounds in this general setup [87,88]. Note that they have been originally formulated for the bandit setup, but apply to Bayesian optimization analogously. Here, the action space and the reward function in the bandit setup corresponds to the parameter space $\mathcal{X}$ and the unknown target function in Bayesian optimization, respectively, see [89, section 10] for details. We will build on techniques in [72,90]. Our theoretical analysis will focus on the *cumulative* regret, i.e., a non-observable quantity that describes the accumulated difference between our incumbent best BO configuration and the *prima facie* unknown optimum.

Definition 8 (*Regret, Cumulative Regret*). Let $r_t = \Psi(\boldsymbol{x}_t) - \min_{\boldsymbol{x} \in \mathcal{X}} \Psi(\boldsymbol{x})$ be the instantaneous regret in iteration $t \in \{1, \dots, T\}$ with $\Psi(\boldsymbol{x}_t)$ the target value of proposal $\boldsymbol{x}_t$ in iteration t and $\min_{\boldsymbol{x} \in \mathcal{X}} \Psi(\boldsymbol{x})$ the universal optimum. Then $R_T = \sum_{t=1}^{T} r_t$ shall be called the cumulative regret.

In the following, we assume the GP's zero-mean noise ϵ, see Eq. (2), to be sub-Gaussian. This is a customary technical assumption in the context of regret analysis, see [70,73,91,92] for recent examples.

Definition 9 (*K-sub-Gaussian*). A zero-mean real-valued random variable X shall be called K-sub-Gaussian, if there exists a constant K^2 such that $\forall \lambda \in \mathbb{R}$ it holds $\mathbb{E}\left[e^{\lambda X}\right] \leq e^{\frac{\lambda^2 K^2}{2}}$.

In order to analyze the above described regret bound, we further need the concept of (maximum) information gains from information theory, see [93] for a textbook reference. It has previously been used to study regret bounds, see [24,72] for instance.

Definition 10 (*Maximum Information Gain [93]*). First denote by $\mathrm{I}\left(\boldsymbol{y}_A; \boldsymbol{\Psi}_A\right)$ the mutual information between $\boldsymbol{\Psi}_A = [\Psi(\boldsymbol{x})]_{\boldsymbol{x} \in A}$ and $\boldsymbol{y}_A = \boldsymbol{\Psi}_A + \boldsymbol{\epsilon}_A$, where $\boldsymbol{\epsilon}_A \sim \mathcal{N}\left(0, \sigma^2 \boldsymbol{I}\right)$, as

$$\mathrm{I}\left(\boldsymbol{y}_A; \boldsymbol{\Psi}\right) = \mathrm{H}\left(\boldsymbol{y}_A\right) - \mathrm{H}\left(\boldsymbol{y}_A \mid \boldsymbol{\Psi}\right),$$

where $H(\cdot)$ is the entropy and $H(\cdot \mid \cdot)$ the conditional entropy, as convention dictates. It quantifies the reduction in uncertainty about $\boldsymbol{\Psi}$ after observing $\boldsymbol{y}_A$ at points $A \subset \mathcal{X}$. The maximum information gain at iteration t shall be defined as

$$\gamma_t := \max_{A \subset \mathcal{X}: |A| = t} I\left(\boldsymbol{y}_A; \boldsymbol{\Psi}_A\right).$$

Note that γ_t is a problem-dependent quantity and can be found given the knowledge of the covariate space $\mathcal{X}$ and the kernel. This is the very reason why it can be expressed in terms of the predictive variances. The following lemma by [72] formalizes this fact. We will need it later in our regret analysis.

Lemma 1 (*Information Gain in Terms of Variances [72]*). *For real-valued $\Psi(\boldsymbol{x})$ it holds*

$$\mathrm{I}\left(\boldsymbol{y}_T; \boldsymbol{\Psi}_T\right) = \frac{1}{2} \sum_{t=1}^{T} \log\left(1 + \sigma^{-2} \sigma_{t-1}^2\left(\boldsymbol{x}_t\right)\right).$$

Proofs of all lemmas and theorems can be found in Appendix A. In order to facilitate a theoretical analysis of BO's sensitivity to GP prior mean, let us assume the ground truth f to be sampled from a Gaussian process.[14] We will point to ways of how to relax this assumption later. We mainly base our theoretical analysis on [72], where cumulative regret bounds are derived for bandit optimization of Gaussian processes, which corresponds to BO in case of the ground truth being sampled from a GP and the action (covariate) space to be infinite.

In order to analyze the effect of GP prior mean misspecification on BO performance, recall Definition 1 of a Gaussian process. Further bear in mind the experimental results from the sensitivity analysis, presented above in Section 3. The main takeaway was that prior mean parameters were the most influential GP prior components. We further learned that constant prior mean parameters can both slow down and speed up BO's convergence. Thus, we will restrict the analysis to constants as functional form of the prior mean to foster a *ceteris paribus* analysis of the prior mean parameter's worst case influence on BO. Let us consider a GP with zero prior mean as surrogate model first: $\mathcal{GP}(0, k(\boldsymbol{x}, \boldsymbol{x}'))$. Its predictive posterior mean in BO iteration T is $\mu_T(\boldsymbol{x}) = \boldsymbol{k}_T(\boldsymbol{x})' \left(\boldsymbol{K}_T + \sigma^2 \boldsymbol{I}\right)^{-1} \boldsymbol{y}_T$. In case of a non-zero prior mean $m(\boldsymbol{x})$ the predictive posterior mean corresponds to the one obtained when applying the usual zero mean GP to the difference between the

[14] Note that this translates to Ψ being sampled from a Gaussian process as well, since $\Psi(\boldsymbol{x}) = f(\boldsymbol{x}) + \epsilon, \epsilon \overset{i.i.d.}{\sim} N(0, \sigma^2 \boldsymbol{I})$, and the fact that a sum of *i.i.d.* normally distributed random variables is again normally distributed.

observations and the fixed mean function [22, page 27]. In our setup, this gives:

$$\begin{aligned}\mu_T(\boldsymbol{x}) &= m(\boldsymbol{x}) + \boldsymbol{k}_T(\boldsymbol{x})' \left(\boldsymbol{K}_T + \sigma^2 \boldsymbol{I}\right)^{-1} (\boldsymbol{y}_T - m_T(\boldsymbol{x}))\\ &= \underbrace{\boldsymbol{k}_T(\boldsymbol{x})' \left(\boldsymbol{K}_T + \sigma^2 \boldsymbol{I}\right)^{-1} \boldsymbol{y}_T}_{=\mu_T(\boldsymbol{x}) \text{ for prior mean zero}} + \underbrace{m(\boldsymbol{x}) - \boldsymbol{k}_T(\boldsymbol{x})' \left(\boldsymbol{K}_T + \sigma^2 \boldsymbol{I}\right)^{-1} m_T(\boldsymbol{x})}_{\epsilon_T(\boldsymbol{x})},\end{aligned} \tag{8}$$

where $m_T(\boldsymbol{x}) = [m(\boldsymbol{x}_1), \dots, m(\boldsymbol{x}_T)]'$. The term $\epsilon_T(\boldsymbol{x})$ is positive in case the prior function predicts higher values than the posterior based on $m(\boldsymbol{x})$, and vice versa. It quantifies the deviation from the predictive posterior mean with zero mean prior and will be the pivotal quantity in what follows. This is due to the fact that we will leverage results from [72] that apply to GPs with zero mean function, thereby in turn relying on techniques from [90]. In particular, the strategy will be to first prove a regret bound for finite $\mathcal{X}$ and then extend it to any compact and convex $\mathcal{X}$ along the lines of [72, Theorem 2].

We will base our analysis on using the lower confidence bound (Definition 5) as an acquisition function since it is the starting point for our robust extension in Section 5. Deploying the lower confidence bound as acquisition function with a GP surrogate model translates to proposing

$$\boldsymbol{x}_t = \arg\max_{\boldsymbol{x} \in \mathcal{X}} AF_{LCB} = \arg\max_{\boldsymbol{x} \in \mathcal{X}} \{-\mu_{t-1}(\boldsymbol{x}) + \tau_t \cdot \sigma_{t-1}(\boldsymbol{x})\} \tag{9}$$

in iteration $t \in \{1, \dots, T\}$. The idea now is to bound $|\Psi(\boldsymbol{x}) - \mu_{t-1}(\boldsymbol{x})|$ for all $t \in \mathbb{N}$ and all $\boldsymbol{x} \in \mathcal{X}$. Closely leaning on [72, Lemma 5.1], the following Lemma formalizes this rationale. Note that the proof directly follows from the proof of [72, Lemma 5.1] and Eq. (8).

Lemma 2 (*Confidence Bound*). *Assume finite $\mathcal{X}$, a GP with prior mean function $m(\boldsymbol{x})$ inducing $\epsilon_T(\boldsymbol{x})$, and BO proposing $\boldsymbol{x}_t$ according to Eq. (9). Pick $\delta \in (0,1)$ and set $\tau_t = 4\log\left(|\mathcal{X}|\pi_t/\delta\right)^2$, where $\sum_{t\geq 1}\pi_t^{-1} = 1, \pi_t > 0$. The following then holds $\forall \boldsymbol{x}\ \forall t \geq 1$ with probability $\geq 1-\delta$*

$$|\Psi(\boldsymbol{x}) - \tilde{\mu}_{t-1}(\boldsymbol{x})| \leq \begin{cases} \tau_t\sigma_{t-1}(\boldsymbol{x}) - \epsilon_{t-1}(\boldsymbol{x}), & \text{if } \epsilon_{t-1}(\boldsymbol{x}) \geq 0\\ \tau_t\sigma_{t-1}(\boldsymbol{x}) + \epsilon_{t-1}(\boldsymbol{x}), & \text{if } \epsilon_{t-1}(\boldsymbol{x}) < 0,\end{cases}$$

where $\tilde{\mu}_{t-1}(\boldsymbol{x})$ is the posterior mean of the GP with prior mean zero.

Based on Lemma 2, we make the simplifying assumption of the prior mean being uniformly too optimistic or too pessimistic, respectively. We simplify things this way, since we are interested in an extreme case analysis of how the GP prior mean affects BO regrets. That is, we focus on the cases

$$\forall t\ \forall \boldsymbol{x} : m(\boldsymbol{x}) > \boldsymbol{k}_T(\boldsymbol{x})' \left(\boldsymbol{K}_T + \sigma^2 \boldsymbol{I}\right)^{-1} m(\boldsymbol{x}) \iff \epsilon_T(\boldsymbol{x}) > 0 \tag{10}$$

and $\forall t\ \forall \boldsymbol{x} : m(\boldsymbol{x}) < \boldsymbol{k}_T(\boldsymbol{x})' \left(\boldsymbol{K}_T + \sigma^2 \boldsymbol{I}\right)^{-1} m(\boldsymbol{x}) \iff \epsilon_T(\boldsymbol{x}) < 0$, respectively. Theorem 1 will require the former, while Theorem 5 in Appendix B will address the latter in an analogous way. Notably, the sign of the misspecification (optimistic or pessimistic) will not affect the growth rate of any of the subsequently presented regret bounds, which are the primary target of our analysis, see Theorem 6, 7, and 8 in Appendix B for a detailed reasoning.

Theorem 1 (*Regret Bound For Optimistic GP Misspecification*). *Let $\delta \in (0,1)$ and $\tau_t = \sqrt{2\log\left(|\mathcal{X}|t^2\pi^2/6\delta\right)}$ with finite $\mathcal{X}$. Bayesian optimization with a GP surrogate with prior mean function $m(\boldsymbol{x})$ inducing $\epsilon_T(\boldsymbol{x})$ (Eq. (8)) with $\forall t\ \forall \boldsymbol{x} : m(\boldsymbol{x}) > \boldsymbol{k}_T(\boldsymbol{x})' \left(\boldsymbol{K}_T + \sigma^2 \boldsymbol{I}\right)^{-1} m(\boldsymbol{x})$ has a cumulative regret R_T such that*

$$\mathbb{P}\left\{ R_T \leq \sqrt{T}\sqrt{\tau_T^2 C_1 \gamma_T + \mathscr{E}\left(2\tau_t \mathcal{S} + \mathscr{E}\right)} \quad \forall T \geq 1 \right\} \geq 1 - \delta,$$

where $C_1 = 8/\log\left(1+\sigma^{-2}\right)$, $\mathcal{S} = \sum_{t=1}^{T} \sigma_{t-1}\left(\boldsymbol{x}_t\right)$ the accumulated GP variances of BO proposals, and $\mathscr{E} = \sum_{t=1}^{T} \epsilon_{t-1}(\boldsymbol{x}_t)$ the accumulated prior-mean induced error terms.

Proofs of all Lemmas and Theorems can be found in Appendix A. Note that this cumulative regret bound grows linear in $\mathscr{E}$. It is of order

$$\mathcal{O}\left(\sqrt{T}\sqrt{\gamma_T \log|\mathcal{X}| + \mathscr{E}\left(4\tau_t\mathcal{S} + \mathscr{E}\right)}\right) \tag{11}$$

with high probability. The same holds for pessimistic misspecification, see Theorem 5 in Appendix B. Under the same assumptions, [72] retrieve sublinear regret bounds. Note that the results do not contradict the sublinearity of the bound obtained in [72, Theorem 3], since the latter requires τ_t to depend on an upper bound for f (more restrictive than Theorem 1) and applies to any f from a RKHS corresponding to the kernel (less restrictive than Theorem 1).

Summing things up, we thus observe that misspecified GP prior means lift BO's regret bounds from sublinearity to linear growth. However, we have not made any further assumptions on the misspecification of the prior mean parameters. By having to account for *any* misspecification, the obtained general regret bounds might be wider than in specific scenarios. The question arises as to whether we can make the bounds tighter by restricting the misspecification. The following Theorem 2 gives an affirmative answer. The key is to bound the misspecification-induced error, not the misspecification itself.

Theorem 2 (*Regret Bound For Sub-Variance GP Misspecification*). *Let $\delta \in (0,1)$ and $\tau_t = \sqrt{2\log\left(|\mathcal{X}|t^2\pi^2/6\delta\right)}$ with finite $\mathcal{X}$. Bayesian optimization with a GP surrogate with prior mean function $m(\boldsymbol{x})$ inducing sub-variance error $\epsilon_T(\boldsymbol{x})$ (Eq. (8)) s.t. $\forall T : \epsilon_T(\boldsymbol{x}) \leq \sigma_T(\boldsymbol{x})$ with $\forall t\ \forall \boldsymbol{x} : m(\boldsymbol{x}) > \boldsymbol{k}_T(\boldsymbol{x})' \left(\boldsymbol{K}_T + \sigma^2 \boldsymbol{I}\right)^{-1} m(\boldsymbol{x})$ has a cumulative regret R_T such that*

$$\mathbb{P}\left\{ R_T \leq \sqrt{T(8\tau_T^2 + 8\tau_T + 2)\gamma_T / \log(1+\sigma^{-2})} \quad \forall T \geq 1 \right\} \geq 1 - \delta.$$

Crucially, this regret bound grows

$$\mathcal{O}\left(\sqrt{T\gamma_T \log|\mathcal{X}|}\right) \tag{12}$$

with high probability. That is, it is of the same order as the regret bound for BO with correctly specified zero mean GP, see [72].[15] In other words, the sublinearity of the regret bounds is restored in case of the misspecification error being upper-bounded by the variance. We have identified the pivotal property of the GP-prior-induced error, namely being bounded by the GP's variance or not.

Closely following [72], we will now lift the regret bound from finite $\mathcal{X}$ to (presumably practically more relevant) compact and convex $\mathcal{X}$. This endeavor requires mild conditions on the kernel of the GP, from which f is assumed to be sampled. The condition is fulfilled by any stationary kernel that is four times differentiable [72, section 4]. It ensures the samples from the corresponding GP are almost surely continuously differentiable [94, Theorem 5]. Examples of kernels that fulfill the condition comprise e.g., the previously introduced power-exponential kernels, see Eq. (C.5), with $p = 2$. It is also fulfilled by all Matérn-kernels, see Eq. (C.6), with $\nu > 2$.

Condition 1 (*Kernel Smoothness [72]*). *Consider any compact and convex $\mathcal{X}$ with $\dim(\mathcal{X}) = d$ and denote by Ψ a sample from a GP with kernel $k\left(\boldsymbol{x}, \boldsymbol{x}'\right)$ and by $\partial\Psi/\partial\boldsymbol{x}$ its partial derivative with regard to $\boldsymbol{x}$. A kernel $k\left(\boldsymbol{x}, \boldsymbol{x}'\right)$ satisfies the hereby defined smoothness condition if*

$$\mathbb{P}\left\{\sup_{\boldsymbol{x}\in\mathcal{X}}\left(\partial\Psi/\partial\boldsymbol{x}_j\right) > L\right\} \leq ae^{-(L/b)^2}, \quad j = 1, \dots, d,$$

for constants $a, b > 0$.

[15] The same holds for pessimistic misspecification, see Theorem 6 in Appendix B.

Theorem 3 (*Regret Bound For Optimistic GP Misspecification on Infinite $\mathcal{X}$*). *Let $\mathcal{X} \subset [0, r]^d$ be compact and convex, $d \in \mathbb{N}, r \in \mathbb{R}_{\geq 0}$. Fix $\delta \in (0, 1)$, and set*

$$\tau_t^2 = 2\log\left(t^2 2\pi^2/(3\delta)\right) + 2d\log\left(t^2 d b r\sqrt{\log(4da/\delta)}\right)$$

with a, b as in condition 1. If Bayesian optimization with misspecified prior mean inducing $\forall t\ :\ \epsilon_t(\boldsymbol{x}) > 0$ is run on Ψ that satisfies condition 1, we obtain the following cumulative regret bound

$$\mathbb{P}\left\{ R_T \leq \sqrt{\tau_T^2 C_1 \gamma_T + (\mathscr{E}+1)\left(2\mathscr{S}\tau_t + \mathscr{E}\right) + \frac{\pi^2}{6}} \quad \forall T \geq 1 \right\} \geq 1-\delta$$

with $C_1 = 8/\log\left(1+\sigma^{-2}\right)$ as in Theorem 1, $\mathscr{S} = \sum_{t=1}^T \sigma_{t-1}\left(\boldsymbol{x}_t\right)$, and $\mathscr{E} = \sum_{t=1}^T \epsilon_{t-1}(\boldsymbol{x}_t)$ the accumulated prior-mean induced error terms.

The idea of the proof is to show Lemma 2 in this setup $\forall t \geq 1$ and fixed x instead of showing it $\forall \boldsymbol{x} \in \mathcal{X}\ \forall t \geq 1$. Then consider a discretization $\mathcal{X}_t \subset \mathcal{X}$ for each t in order to prove Lemma 2 $\forall \boldsymbol{x} \in \mathcal{X}_t\ \forall t \leq 1$ and then let $\mathcal{X}_t$ get dense as t gets large. Note that the cumulative regret bound remains linear in $\mathscr{E}$ like in the finite $\mathcal{X}$ case, see Eq. (11). The same holds for pessimistic misspecification, see Theorem 7 in Appendix B.

The only thing that is left now is to lift Theorem 2 for the sub-variance GP misspecification to the case of infinite $\mathcal{X}$, too. Theorem 4 does the job.

Theorem 4 (*Regret Bound For Sub-Variance GP Misspecification on Infinite $\mathcal{X}$*). *Let $\mathcal{X} \subset [0, r]^d$ be compact and convex, $d \in \mathbb{N}, r \in \mathbb{R}_{\geq 0}$. Fix $\delta \in (0, 1)$, and set*

$$\tau_t^2 = 2\log\left(t^2 2\pi^2/(3\delta)\right) + 2d\log\left(t^2 d b r\sqrt{\log(4da/\delta)}\right)$$

with a, b as in condition 1. If Bayesian optimization with misspecified prior mean inducing sub-variance error $\epsilon_T(\boldsymbol{x})$, i.e., $\forall T\ :\ \ \epsilon_T(\boldsymbol{x}) \leq \sigma_T(\boldsymbol{x})$ with $\forall t\ :\ \epsilon_t(\boldsymbol{x}) > 0$ is run on target function Ψ that satisfies condition 1, we obtain the following cumulative regret bound.

$$\mathbb{P}\left\{ R_T \leq \sqrt{(8\tau_T^2 + 4\tau_T + 2)\gamma_T/\log\left(1+\sigma^{-2}\right)} + \frac{\pi^2}{6} \quad \forall T \geq 1 \right\} \geq 1-\delta.$$

5. PROBO: Prior-mean-robust Bayesian optimization

The sensitivity analysis in Section 3 has shown that the algorithm's convergence is especially sensitive towards the prior mean function's parameters. It was followed by a theoretical analysis in Section 4 suggesting that misspecification of the latter has the potential to lift the regret bounds of Bayesian optimization from sub-linearity to linearity. In summary, we conclude that Bayesian optimization heavily depends on its hyperparameters, in particular on the Gaussian process prior mean specification.

> *Ignoramus et ignorabimus.* (We do not know and we will never know.)
> – attributed to Emil Heinrich Du Bois-Reymond, cited after [95]

In light of these findings, it appears desirable to mitigate BO's dependence on the prior mean parameters by expressing a state of ignorance about the latter. Recall that Bayesian optimization is typically used for "black-box-functions", where very little, if any, prior knowledge about the functional relationship under study exists. The classical approach would be to specify a so-called non-informative hyperprior over the prior mean parameters. However, such a prior is not unique [96,97] and choosing different priors among the set of all non-informative priors would lead to different posterior inferences [14,98]. Such classical non-informative priors can thus not be regarded as fully uninformative and represent indifference rather than ignorance. Generally, unique priors describing a state of total ignorance are "missing ingredients required by the [Bayesian] prescription" [99, p. 162]. In response to this disillusioning fact, practitioners often turn to empirical Bayes. That is, they estimate prior parameters from the data, deliberately violating the Bayesian paradigm by peeking at the observations before stating prior knowledge. [99, p. 162] aptly characterizes empirical Bayes as a "pragmatic remedy for managing the headache created by the missing ingredients required by the prescription."

In this work, we argue that in the case of Bayesian optimization, this widely adopted pragmatic remedy of empirical Bayes might cause serious side effects. This is simply due to the fact that such Bayes-optimal estimation of a location parameter (empirical Bayes) does not necessarily equal the Bayes-optimal action in cumulative regret minimization as in BO. The deeper reason for this is that the prior components' effect on BO goes beyond mere inferential (or predictive) purposes. Instead, they interfere with the exploration–exploitation trade-off, which is essential to BO's convergence. Optimization and estimation of $f(\boldsymbol{x})$ can be competing aims. In other words, a prior that is optimal for inferential purposes not necessarily equates the one most favorable for fast convergence towards the optimum. Consider the bias of the estimator as an exemplary statistical property for illustrative intuition. The estimation of a population's location parameter through empirical Bayes via maximum likelihood from an *i.i.d.* sample is typically unbiased. The Bayes-optimal settings of the GP mean parameters in BO, however, might correspond to systematically over- or under-estimating the true mean, since this can speed up convergence, see experimental results in Section 3, depending on the target function and the explore–exploit setting. Since these Bayes-optimal settings, however, are generally unknown, arbitrary choices might as well hamper convergence. As our theoretical analysis in Section 4 has revealed, they can even lift the regret bounds from sublinear to linear growth. For even more intuition on how the GP prior mean can fiddle with the explore–exploit trade-off, think of the toy example of a simple-to-optimize (low-dimensional, unimodal and highly smooth) unknown target function that is being optimized by BO with LCB. Assume now a risk-averse decision maker with high τ in the LCB, see Definition 5. In this case, BO would waste budget by unnecessarily exploring regions of the covariate space with sub-optimal function values. This latter behavior can – loosely speaking – be both mitigated and enforced by the prior mean, since it dominates the posterior mean in regions with few observations. For instance, a severe underestimation of the true function would *ceteris paribus* lead to the LCB being dominated by the mean estimation, thus attenuating the influence of the variance term (Definition 5 of LCB) and reducing exploration, and vice versa.

Principled approaches argue that this dilemma of Bayesian inference in the absence of prior information cannot be solved within the framework of classical precise probabilities. Methods working with sets of priors have thus attracted increasing attention, see e.g. [100,101] for an introduction and [102–106] for applications. Truly uninformative priors, however, would entail sets of all possible probability distributions and thus lead to vacuous posterior inference. That is, prior beliefs would not change with data, which would make learning impossible. [98] thus propose prior *near*-ignorance models as a compromise that conciliates learning and *almost* non-informative priors. Prior near-ignorance models are characterized by placing a "probability interval" $[0, 1]$ on certain "standard" events, expressing ignorance about their tendency to occur. A popular example for a prior near-ignorance model is the imprecise Dirichlet model [107], which has wide reaching applications [108].

In the case of Gaussian processes, so-called imprecise Gaussian processes (IGP) were introduced by [13,14] as prior near-ignorance models for GP regression. The general idea of an IGP is to incorporate the model's imprecision regarding the choice of the prior's mean function parameter, given a constant mean function and a fully specified kernel. In the case of univariate regression, given a base kernel $k(x, x')$ and a degree of imprecision $c > 0$, [13, definition 2] defines a constant mean imprecise Gaussian process as a set of GP priors:

$$\mathcal{G}_c = \left\{ GP\left(Mh, k(x,x') + \frac{1+M}{c} \right) \, : \, h = \pm 1, M \geq 0 \right\}. \tag{13}$$

It can be shown that $\mathcal{G}_c$ verifies prior near-ignorance [13, page 194] and that $c \to 0$ yields the precise model [13, page 189]. Note that the mean functional form (constant) as well as both kernel functional form and its parameters do not vary in set $\mathcal{G}_c$, but only the mean parameter $Mh \in \,]-\infty, \infty[$. For each prior GP, a posterior GP can be inferred. This results in a set of posteriors and a corresponding set of mean estimates $\hat{\mu}(x)$, of which the upper and lower mean estimates $\underline{\hat{\mu}}(x)$, $\overline{\hat{\mu}}(x)$ can be derived analytically. To this very end, let $k(x,x')$ be a kernel function as defined in [22]. The finitely positive semi-definite matrix $\boldsymbol{K}$ is then formed by applying $k(x,x')$ on the training data vector $\boldsymbol{x} \in \mathcal{X}$:

$$\boldsymbol{K} = [k(x_i, x_j')]_{ij}. \tag{14}$$

Following [13], we call $\boldsymbol{K}$ base kernel matrix. Note that $\boldsymbol{K}$ is restricted only to be finitely positive semi-definite and not to have diagonal elements of 1. In statistical terms, $\boldsymbol{K}$ is a covariance matrix and not necessarily a correlation matrix. Hence, the variance $I\sigma^2$ is included. Now let x be a scalar input of test data, whose $f(x)$ is to be predicted. Then recall $\boldsymbol{k}(\boldsymbol{x}) = [k(x,x_1), \dots, k(x,x_n)]'$ is the vector of covariances between x and the training data. Furthermore, define $\boldsymbol{s}_k = \boldsymbol{K}^{-1}\mathbb{1}_n$ and $\boldsymbol{S}_k = \mathbb{1}_n'\boldsymbol{K}^{-1}\mathbb{1}_n$. Then [13] shows that upper and lower bounds of the posterior predictive mean function $\hat{\mu}(x)$ for $f(x)$ can be derived. If $|\frac{\boldsymbol{s}_k \boldsymbol{y}}{\boldsymbol{S}_k}| \leq 1 + \frac{c}{\boldsymbol{S}_k}$, they are:

$$\overline{\hat{\mu}}(x) = \boldsymbol{k}(\boldsymbol{x})'\boldsymbol{K}^{-1}\boldsymbol{y} + (1 - \boldsymbol{k}(\boldsymbol{x})'\boldsymbol{s}_k)\frac{\boldsymbol{s}_k'}{\boldsymbol{S}_k}\boldsymbol{y} + c\frac{|1 - \boldsymbol{k}(\boldsymbol{x})'\boldsymbol{s}_k|}{\boldsymbol{S}_k} \tag{15}$$

$$\underline{\hat{\mu}}(x) = \boldsymbol{k}(\boldsymbol{x})'\boldsymbol{K}^{-1}\boldsymbol{y} + (1 - \boldsymbol{k}(\boldsymbol{x})'\boldsymbol{s}_k)\frac{\boldsymbol{s}_k'}{\boldsymbol{S}_k}\boldsymbol{y} - c\frac{|1 - \boldsymbol{k}(\boldsymbol{x})'\boldsymbol{s}_k|}{\boldsymbol{S}_k} \tag{16}$$

If $|\frac{\boldsymbol{s}_k \boldsymbol{y}}{\boldsymbol{S}_k}| > 1 + \frac{c}{\boldsymbol{S}_k}$:

$$\overline{\hat{\mu}}(x) = \boldsymbol{k}(\boldsymbol{x})'\boldsymbol{K}^{-1}\boldsymbol{y} + (1 - \boldsymbol{k}(\boldsymbol{x})'\boldsymbol{s}_k)\frac{\boldsymbol{s}_k'}{\boldsymbol{S}_k}\boldsymbol{y} + c\frac{1 - \boldsymbol{k}(\boldsymbol{x})'\boldsymbol{s}_k}{\boldsymbol{S}_k} \tag{17}$$

$$\underline{\hat{\mu}}(x) = \boldsymbol{k}(\boldsymbol{x})'\boldsymbol{K}^{-1}\boldsymbol{y} + (1 - \boldsymbol{k}(\boldsymbol{x})'\boldsymbol{s}_k)\frac{\boldsymbol{s}_k'\boldsymbol{y}}{c + \boldsymbol{S}_k} \tag{18}$$

The corresponding variance estimate of both $\overline{\mu}(x)$ and $\underline{\mu}(x)$ is

$$\hat{\sigma}^2_{f(x)} = k(x,x) - \boldsymbol{k}(\boldsymbol{x})'\boldsymbol{K}^{-1}\boldsymbol{k}(\boldsymbol{x}) + \frac{(1 - \boldsymbol{k}(\boldsymbol{x})'\boldsymbol{s}_k)^2}{\boldsymbol{S}_k} \tag{19}$$

We can retrieve credible intervals for the predictions of an imprecise GP as follows. For $\alpha \in [0,1]$ and z_q the q-quantile of the standard normal distribution, the $1-\alpha$ credible intervals of the mean estimate for $f(x)$ is

$$CrI_\alpha = [\underline{f}_x = \underline{\mu}(x) - z_{1-\frac{\alpha}{2}} \cdot \hat{\sigma}^2_{f(x)}, \overline{f}_x = \overline{\mu}(x) + z_{1-\frac{\alpha}{2}} \cdot \hat{\sigma}^2_{f(x)}]. \tag{20}$$

[13, Theorem 4] shows that $CrI_\alpha = [\underline{f}_x, \overline{f}_x]$ satisfies $\overline{P}(f(x) < \underline{f}_x) \leq \frac{\alpha}{2}$ and $\overline{P}(f(x) > \overline{f}_x) \leq \frac{\alpha}{2}$. Fig. 3 visualizes upper and lower mean function estimates as well as corresponding credible intervals of an imprecise Gaussian process trained on data generated by $f(x) = x \cdot \sin(x) + 0.1x$. The prediction function including credible intervals of a precise (classical) Gaussian process is also depicted. As can be seen by comparing predictions in $x \in [-10,-5]$ to $x \in [4,7]$, the model imprecision $\overline{\mu}(x) - \underline{\mu}(x)$ is greater than the classical prediction uncertainty (credible interval of precise GP) in the absence of data. Here, the prior dominates the data in the posterior. The opposite holds in the abundance of data. Noteworthy, imprecise Gaussian processes require the estimated target function to be univariate, i.e., $\dim(\mathcal{X}) = 1$. Our proposed extension of BO will inherit this restriction, limiting its applicability. However, we emphasize that many multivariate optimization problems can be embedded in univariate subspaces, see [109,110] for instance, without substantial loss of the solution's efficiency. In the next Section 6 on PROBO's application, we practically demonstrate the feasibility of such embedding techniques both in general and in combination with our method in particular. We particularly point to embedding-based benchmarking of PROBO against the classical LCB for the problem of graphene, as summarized in Appendix G for random embedding and in Appendix H for embedding based on principal component analysis (PCA).

Inspired by multi-objective BO [111], one might think (despite knowing better) of an IGP and a GP as surrogate models for different target functions. A popular approach in multi-objective BO to proposing points based on various surrogate models is to scalarize their predictions by an acquisition function defined *a priori*. The proposed generalized lower confidence bound (GLCB) is such an acquisition function, since it combines mean and variance predictions of a precise GP with upper and lower mean estimates of an IGP, see also [1]. In this way, it generalizes the popular upper/lower confidence bound $LCB(\boldsymbol{x}) = \hat{\mu}(\boldsymbol{x}) - \tau_t \cdot \sqrt{var(\hat{\mu}(\boldsymbol{x}))}$, recall Definition 5.

Definition 11 (*Generalized Lower Confidence Bound (GLCB)*)**.** Let $\boldsymbol{x} \in \mathcal{X}$. As above, let $\overline{\hat{\mu}}(\boldsymbol{x}), \underline{\hat{\mu}}(\boldsymbol{x})$ be the upper/lower mean estimates of an IGP with imprecision c. Let $\hat{\mu}(\boldsymbol{x})$ and $var(\hat{\mu}(\boldsymbol{x}))$ be the mean and variance predictions of a precise GP. The prior-mean-robust acquisition function *generalized lower confidence bound (GLCB)* shall then be

$$GLCB(\boldsymbol{x}) = \hat{\mu}(\boldsymbol{x}) - \tau_t \cdot \sqrt{var(\hat{\mu}(\boldsymbol{x}))} - \rho \cdot (\overline{\hat{\mu}}(\boldsymbol{x}) - \underline{\hat{\mu}}(\boldsymbol{x})).$$

By explicitly accounting for the prior-induced imprecision, GLCB generalizes the trade-off between exploration and exploitation: $\tau_t > 0$ controls the classical "mean vs. data uncertainty" trade-off (degree of risk aversion) and $\rho > 0$ controls the "mean vs. model imprecision" trade-off (degree of ambiguity aversion). Notably, $\overline{\hat{\mu}}(\boldsymbol{x}) - \underline{\hat{\mu}}(\boldsymbol{x})$ simplifies to an expression only dependent on the kernel vector between x and the training data $\boldsymbol{k}(\boldsymbol{x}) = [k(x,x_1), \dots, k(x,x_n)]'$, the base kernel matrix $\boldsymbol{K}$ (Eq. (14)) and the degree of imprecision c, which follows from Eqs. (17) and (18) in case $|\frac{\boldsymbol{s}_k \boldsymbol{y}}{\boldsymbol{S}_k}| > 1 + \frac{c}{\boldsymbol{S}_k}$:

$$\overline{\hat{\mu}}(x) - \underline{\hat{\mu}}(x) = (1 - \boldsymbol{k}(\boldsymbol{x})'\boldsymbol{s}_k)\Big(\frac{\boldsymbol{s}_k'}{\boldsymbol{S}_k}\boldsymbol{y} + \frac{c}{\boldsymbol{S}_k} - \frac{\boldsymbol{s}_k'\boldsymbol{y}}{c + \boldsymbol{S}_k}\Big) \tag{21}$$

As can be seen by comparing Eqs. (15) and (16), in case of $|\frac{\boldsymbol{s}_k \boldsymbol{y}}{\boldsymbol{S}_k}| \leq 1 + \frac{c}{\boldsymbol{S}_k}$, the model imprecision $\overline{\hat{\mu}}(\boldsymbol{x}) - \underline{\hat{\mu}}(\boldsymbol{x})$ even simplifies further, as follows.

$$\overline{\hat{\mu}}(x) - \underline{\hat{\mu}}(x) = 2c\frac{|1 - \boldsymbol{k}(\boldsymbol{x})'\boldsymbol{s}_k|}{\boldsymbol{S}_k} \tag{22}$$

In this case, the GLCB comes down to $GLCB(\boldsymbol{x}) = \hat{\mu}(\boldsymbol{x}) - \tau_t \cdot \sqrt{var(\hat{\mu}(\boldsymbol{x}))} - 2 \cdot \rho c \frac{|1 - \boldsymbol{k}(\boldsymbol{x})'\boldsymbol{s}_k|}{\boldsymbol{S}_k}$ and the two hyperparameters ρ and c collapse to one. In both cases, the surrogate models $\underline{\hat{\mu}}(x)$ and $\overline{\hat{\mu}}(x)$ do not have to be fully implemented. Only $\boldsymbol{K}$ and $\boldsymbol{k}(\boldsymbol{x}) = [k(x,x_1), \dots, k(x,x_n)]'$ need to be computed. GLCB can thus be plugged into standard BO without much additional computational cost.[16] Algorithm 2 describes the procedure.

[16] Further note that with expensive target functions to optimize, the computational costs of surrogate models and acquisition functions in BO can be regarded as negligible. The computational complexity of PROBO is the same as for BO with GP.

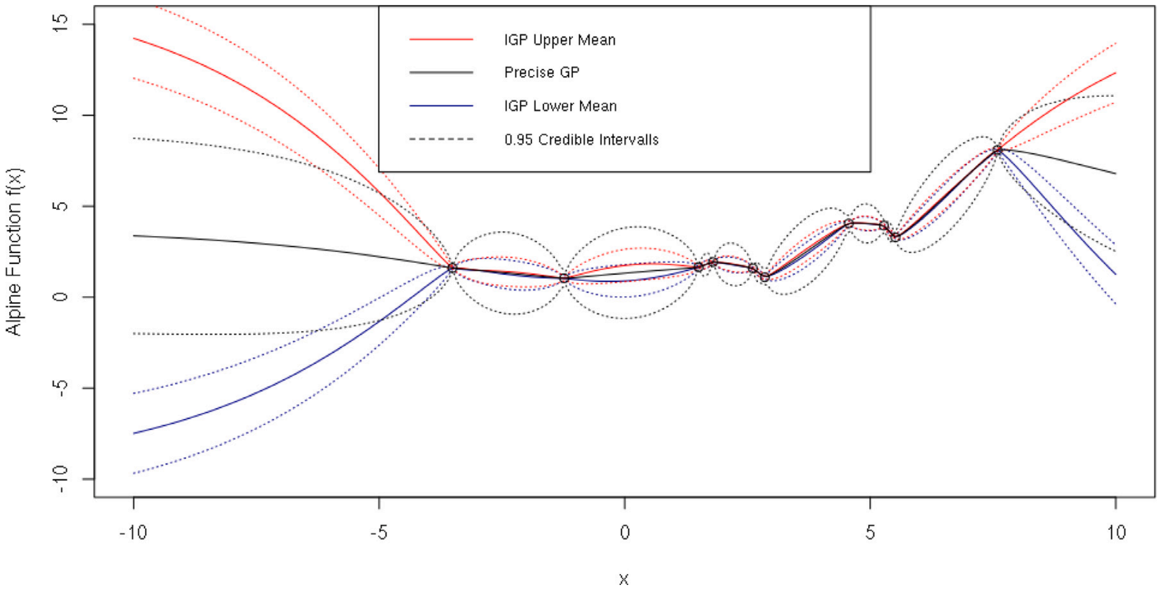


Fig. 3. Upper and lower mean estimates of imprecise GP and precise GP estimates from data generated by $f(x) = x \cdot \sin(x) + 0.1x$ ("alpine function"), see [84].

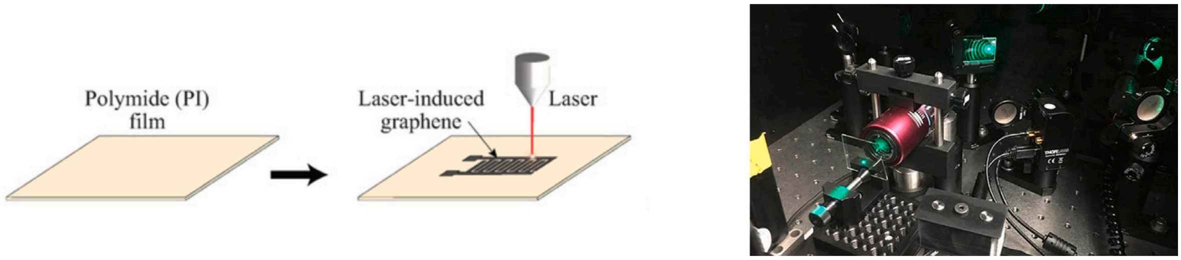


Fig. 4. Producing graphene via laser irradiation. Image credits: [112, page 3].

Algorithm 2 Prior-mean-RObust Bayesian Optimization (PROBO)

```
create an initial design $D = \{(\boldsymbol{x}^{(i)}, \Psi^{(i)})\}_{i=1,\dots,n_{init}}$ of size $n_{init}$
specify $c$ and $\rho$
while termination criterion is not fulfilled do
    train a precise GP on data $D$ and obtain $\hat{\mu}(\boldsymbol{x})$, $var(\hat{\mu}(\boldsymbol{x}))$
    compute $\boldsymbol{k}(\boldsymbol{x})$, $\boldsymbol{s}_k$ and $\boldsymbol{S}_k$
    if $|\frac{\boldsymbol{s}_k \boldsymbol{y}}{\boldsymbol{S}_k}| > 1 + \frac{c}{\boldsymbol{S}_k}$ then
        $\overline{\hat{\mu}}(x) - \underline{\hat{\mu}}(x) = (1 - \boldsymbol{k}(\boldsymbol{x})' \boldsymbol{s}_k)\big(\frac{\boldsymbol{s}_k'}{\boldsymbol{S}_k} \boldsymbol{y} + \frac{c}{\boldsymbol{S}_k} - \frac{\boldsymbol{s}_k' \boldsymbol{y}}{c + \boldsymbol{S}_k}\big)$
    else $\overline{\hat{\mu}}(x) - \underline{\hat{\mu}}(x) = 2c \frac{|1 - \boldsymbol{k}(\boldsymbol{x})' \boldsymbol{s}_k|}{\boldsymbol{S}_k}$
    compute $GLCB(\boldsymbol{x}) = -\hat{\mu}(\boldsymbol{x}) + \tau_t \cdot \sqrt{var(\hat{\mu}(\boldsymbol{x}))} + \rho \cdot (\overline{\hat{\mu}}(\boldsymbol{x}) - \underline{\hat{\mu}}(\boldsymbol{x}))$
    propose $\boldsymbol{x}^{new}$ that optimizes $GLCB(\boldsymbol{x})$
    evaluate $\Psi$ on $\boldsymbol{x}^{new}$
    update $D \leftarrow D \cup (\boldsymbol{x}^{new}, \Psi(\boldsymbol{x}^{new}))$
end while
return $\arg\min_{x \in D} \Psi(\boldsymbol{x})$ and respective $\Psi(\arg\min_{x \in D} \Psi(\boldsymbol{x}))$
```

Just like LCB, the generalized LCB balances optimization of $\hat{\mu}(x)$ and reduction of uncertainty with regard to the model's prediction variation $\sqrt{var(\hat{\mu}(\boldsymbol{x}))}$ through τ_t. What is more, GLCB aims at reducing model imprecision caused by the prior specification, controllable by ρ. This allows returning optima that are robust not only towards classical prediction uncertainty but also towards imprecision of the specified model, see Section 1.

6. Application on graphene production

We test our method on a univariate target function generated from a data set that describes the quality of experimentally produced graphene, an allotrope of carbon with potential use in semiconductors, smartphones and electric batteries [81]. The data set comprises $n = 210$ observations of an experimental manufacturing process of graphene. High-performance plastics like polyimide films, typically Kapton, are irradiated with a laser in a reaction chamber in order to trigger a chemical reaction that results in graphene, see Fig. 4. Four covariates influence the manufacturing process, namely power and time of the laser irradiation as well as gas in and pressure of the reaction chamber [81]. The target variable (to be maximized) is a measure for the quality of the induced graphene, ranging from 0.1 to 5.5 (see Table 3).

Table 3
Graphene data set [81].

covariate	min	max	type	description
power	10	5555	real-valued	power of the laser
time	500	20210	real-valued	irradiation time
gas			categorical	gas used in the reaction chamber (Nitrogen, Argon, Air)
pressure	0	1000	real-valued	pressure in the reaction chamber
target quality	0.1	5.5	real-valued	quality of induced graphene

In order to construct a univariate target function from the data set, a random forest was trained on subsets of it (target quality and power as well as target quality and time, see Fig. 5). The predictions of these random forests were then used as target functions to be optimized in order to compare the proposed BO method to existing ones on a real-world problem.

We compare GLCB to its classical counterpart LCB, see Definition 5, as well as to six other popular acquisition functions like expected improvement (EI), see Definition 4. Fig. 6 highlights the results for GLCB

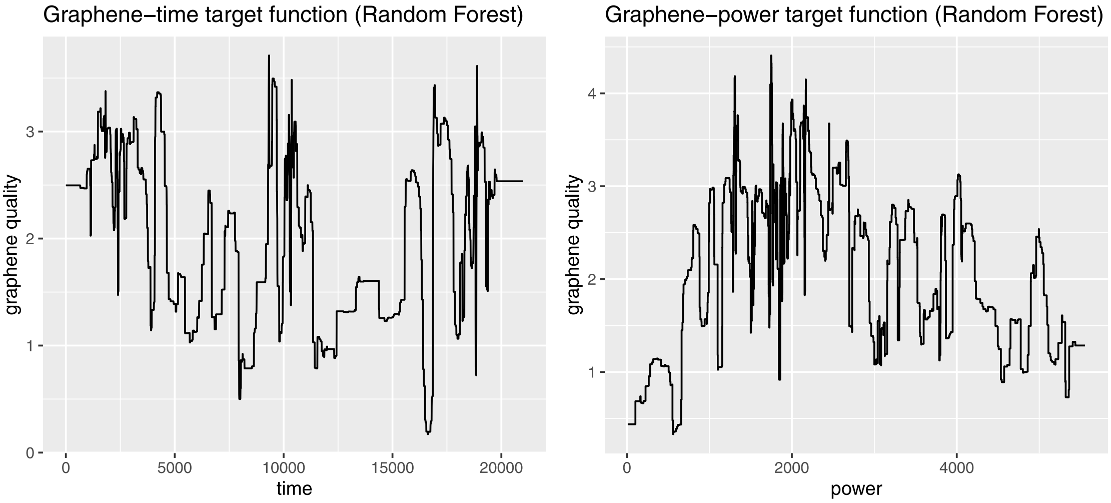


Fig. 5. Univariate target functions estimated from graphene data.

vs. LCB on the graphene-time function, while all remaining results can be found in Appendix F. These include, besides other competing acquisition functions, the benchmarks on the graphene-power target function. For each comparison, we observe $n = 60$ BO runs with a budget of 90 evaluations and an initial design of 10 data points generated by Latin hypercube sampling [113] each. Focus search [21, page 7] was used as AF optimizer with 1000 evaluations per round and 5 maximal restarts. All experiments were conducted in R version 4.0.3 [114] on a high-performance computing cluster using 20 cores (linux gnu). Fig. 6 depicts mean optimization paths of BO with GLCB compared to LCB on the graphene-time target function. The paths are shown for three different GLCB settings: $\rho = 1, c = 50$ and $\rho = 1, c = 100$ as well as $\rho = 10, c = 100$. We observe that GLCB surpasses LCB (all settings). Results in Appendix F show that we retrieve similar results for EI (with $\rho = 10, c = 100$) in late iterations and other acquisition functions, except for one purely exploratory and thus degenerated acquisition function. The results on the graphene-power target function reveal a similar pattern, except that GLCB is outperformed by EI, see also Appendix F.

Generally, it becomes evident that none of the methods reaches the global optimum of 5.5 or come close to it within the allocated budget. This can be attributed to the general hardness of the problem of graphene production and the high costs of conducting one experiment, which mainly stem from the required time to set up an experiment: "The preparation of a sample to be irradiated requires about one week to produce the graphene oxide powder and 1-2 days to create the ink and deposit it onto the subs" [112]. For more background on the nature of experimental graphene production, see [81,112].

Moreover, we benchmark our proposed methods on synthetic functions to study how these results generalize to applications beyond graphene production. We select a series of synthetic benchmarking problems for optimizer benchmarking from [85] which includes test functions from the well-established BBOB benchmarking suite [115–117]. The results can be summarized as follows: As long as the objective function is sufficiently wiggly and multimodal, PROBO achieves state-of-the-art results, see in particular the results on the noisy and multimodal drop-wave function in Appendix I. For smooth, uni- or bimodal target functions, however, PROBO is outperformed by competing methods, see particularly the results for the alpine function in Appendix J. Apparently, the superiority on noisy, multimodal targets (that are typically hard to optimize) like the graphene production problem does not come for free. It entails slower convergence in case of (very) simple optimization problems. In light of PROBO's motivation, this appears quite plausible: Our method hedges against the risk of model misspecification. The latter has only limited effect and does not outweight the additional exploration in case the model specification does not matter that much due to the true target function's simplicity. We recommend further research to better understand the determinants of optimization problems that can be solved more efficiently by accounting for model imprecision.

Another pattern from the results catches one's attention immediately, namely the late iterations, in which GLCB outperforms its competitors. Loosely speaking, accounting for model imprecision apparently needs time to play out its strengths. Only logically, the reduction of model imprecision needs a few iterations to impact the model's predictions that in turn impact the algorithm's proposals. This motivates an extension of our acquisition function to more complex multivariate target functions, as they usually require a higher budget of BO evaluations to be optimized. Recall that we restricted ourselves to the univariate case due to the one-dimensional nature of the imprecise Gaussian processes proposed by [13]. The fruitful application of imprecise Gaussian processes in BO might initiate a more general formulation of imprecise Gaussian processes.

We further point to benchmarking results based on other target functions resulting from univariate embeddings of all covariates power, time, gas (one-hot encoding), and pressure. As briefly mentioned in Section 5, these results are summarized in Appendix G for random embedding and in Appendix H for more statistically informed embedding based on principal component analysis (PCA).[17] These results confirm the competitive performance of PROBO. However, GLCB does not statistically outperform LCB on these embedding-based problems.

7. Discussion

The promising results presented above should not hide the fact that the proposed modification makes the optimizer robust only with regard to possible misspecification of the mean function parameter given a constant trend. Albeit the sensitivity analysis conducted in Section 3 demonstrated its importance, the mean parameter is clearly not the only influential component of the GP prior in BO. For instance, the functional form of the kernel also plays a major role, see Table 2. The question of how to specify this prior component is discussed in [4,61]. Apart from this, it is important to note that PROBO depends on a subjectively specified degree of imprecision c. It does not account for any imaginable prior mean (the model would become vacuous, see Section 5). What is more, it may be difficult to interpret c and

[17] The first principal component's scores were taken as univariate representation in this type of embedding strategy.

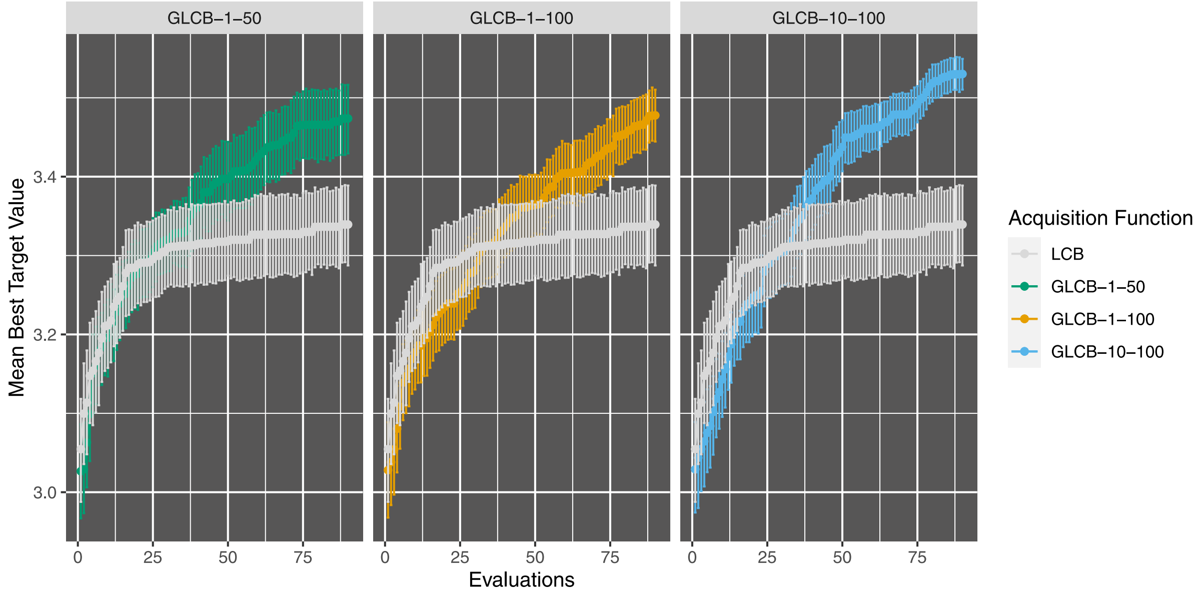


Fig. 6. Benchmarking results from graphene data: Generalized lower confidence bound (GLCB) vs. lower confidence bound (LCB). Shown are 60 runs per Acquisition Function with 90 evaluations and initial sample size 10 each. Error bars represent 95% confidence intervals. GLCB-1-100 means $\rho = 1$ and c = 100; $\tau_t = 1$ for all GLCBs and LCB.

thus specify it in practical applications. However, our method still offers more generality than a precise choice of the mean parameter. Specifying c corresponds to a weaker assumption than setting precise mean parameters.

Notwithstanding such deliberations concerning PROBO's robustness and generality, the method simply converges faster than BO when faced with graphene production. Such multimodal and wiggly (see Fig. 5) functions latter make up an arguably considerable part of problems not only materials science, but also in other relevant applications of BO such as hyperparameter-tuning [19], engineering [118] or drug discovery [16]. We thus conclude that PROBO has high potential in several real-world applications of Bayesian optimization, where a univariate embedding along the lines of [109] is feasible.

The herein proposed BO robustification framework PROBO as well as the empirical and theoretical analysis of BO under GP misspecification open up several venues for further work. First and foremost, multi-objective optimization problems appear to constitute a fruitful field of further study. We have already hinted at potential extensions in Section 2.4. Moreover, an extension of the rationale behind PROBO to other Bayesian surrogate models seems feasible, since there is a variety of prior near-ignorance models besides imprecise Gaussian processes. What is more, also non-Bayesian surrogate models like random forests, boosting methods or support vector machines can be altered such that they account for imprecision in their assumptions, see [119–125] for instance. Generally speaking, IP models appear very fruitful in the context of optimization based on surrogate models. They not only offer a vivid framework to represent prior ignorance, as demonstrated in this very paper, but may also be beneficial in applications where prior knowledge is abundant. In such situations, in the case of data contradicting the prior, precise probabilities often fail to adequately represent uncertainty, whereas IP models can handle these prior-data conflicts, see e.g., [105,106,126].

Furthermore, the theoretical analysis in Section 4 paves the way for several extensions, two of which shall be briefly outlined in what follows. First, the assumption of the ground truth f being sampled from a Gaussian process could be dropped. Analogous to [72, Theorem 3] and based on Theorem 3 and Theorem 4 regret bounds for f from a reproducing kernel Hilbert space (RKHS) might be within reach. In this agnostic setting, the challenge will be to define misspecification since a ground truth mean function is unavailable, as GPs from an RKHS are identified with their kernel only. Second, recall that the kernel's functional form was found to be the second most influential GP prior component in the empirical analysis in Section 3. It might be of interest to conduct a similar theoretical analysis of its influence on cumulative regret bounds. We point to the kernel-based Definition 10 of information gain and corresponding regret analyses in [23, section 4] and [24, section 3.1]. Since they are based on (conditional) entropy, they can be extended to account for sets of kernel functional forms using upper and lower entropy measures proposed by [127]. An additional theoretical extension of this work would be to focus on convergence rather than probabilistic regret bounds under GP prior misspecification, see [128,129] for pointers to convergence analyses.

As mentioned in Section 1, there are many more derivative-free optimizers with probabilistic elements. Consider, for instance, simulated annealing [130]. Inspired by cooling-down processes of metals and liquids, simulated annealing is a local search (i.e., it uniformly samples from a hypercube or an ϵ-ball around the current optimal value $\boldsymbol{x}_t$) that casually accepts parameters from the rejection area $\{\boldsymbol{x}_{t+1} : f(\boldsymbol{x}_{t+1}) - f(\boldsymbol{x}_t) > 0\}$. It uses the co-called Metropolis criterion: Accept parameter $\boldsymbol{x}_{t+1}$ from rejection area with $\mathbb{P} = \exp\left(\frac{f(x_{t+1})-f(x_t)}{T}\right)$, where T is the temperature of the system, which monotonically decreases with the iterations. $\mathbb{P}$ constitutes an exponential distribution with $\lambda = 1$. Different distributions from the same distributional family and their effect on the optimization path could be assessed in future work. Furthermore, the interaction of the two probability measures (uniform distribution and exponential distribution) might be investigated. Another example of optimizers relying on probabilistic assumptions are evolutionary algorithms mimicking the evolution of animal populations through natural selection. A crucial part of EA is the mutation operator, as it ensures diversity in following generations [131]. Besides uniformly sampling $\boldsymbol{x}_{t+1}$ from the covariate space, the Gauss-mutation has gained popularity in recent years: $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t \pm \sigma\mathcal{N}(\mathbf{0}, \mathbf{I})$. Leaning on so-called neighborhood models [132], one could sample from a set of (normal) distributions. The resulting set of populations could then be ordered by a fitness function. The induced ordering can itself be imprecise, as proposed by [133, Chapter 5.1] as "imprecise fitness comparison". Another point of attack from a robustness point of view can be the specification of covariance matrices in the highly popular covariance matrix adaptation evolutionary search (CMAES) algorithm [6].

CRediT authorship contribution statement

Julian Rodemann: Writing – review & editing, Writing – original draft, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Thomas Augustin:** Validation, Supervision, Writing – review & editing, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors are indebted to the three anonymous reviewers for valuable suggestions that helped to improve the paper, to Lars Kotthoff for providing data on graphene production, to Francesca Mangili for answering questions on imprecise Gaussian processes, as well as to Christoph Jansen and Georg Schollmeyer for valuable remarks on acquisition functions in light of the dominance principle.

Both authors acknowledge support by the Federal Statistical Office of Germany within the co-operation project "Machine Learning in Official Statistics". JR further acknowledges support by the Bavarian Academy of Sciences (BAS) through the Bavarian Institute for Digital Transformation (bidt) and the mentoring program by the Faculty of Mathematics, Informatics, and Statistics at Ludwig-Maximilians-Universität (LMU) Munich.

Appendix A. Proofs

Lemma 1 (*Information Gain in Terms of Variances [72]*)**.** For real-valued $\Psi(\boldsymbol{x})$ it holds

$$\mathrm{I}\left(\boldsymbol{y}_T;\boldsymbol{\Psi}_T\right)=\frac{1}{2}\sum_{t=1}^{\prime}\log\left(1+\sigma^{-2}\sigma_{t-1}^{2}\left(\boldsymbol{x}_t\right)\right).$$

Proof. See proof of [72, Lemma 5.3]. □

Lemma 2 (*Confidence Bound*)**.** Assume finite $\mathcal{X}$, a GP with prior mean function $m(\boldsymbol{x})$ inducing $\epsilon_T(\boldsymbol{x})$, and BO proposing $\boldsymbol{x}_t$ according to Eq. (9). Pick $\delta\in(0,1)$ and set $\tau_t=4\log\left(|\mathcal{X}|\pi_t/\delta\right)^2$, where $\sum_{t\geq 1}\pi_t^{-1}=1,\pi_t>0$. The following then holds $\forall\boldsymbol{x}\in\mathcal{X}\ \forall t\geq 1$ with probability $\geq 1-\delta$

$$\left|\Psi(\boldsymbol{x})-\tilde{\mu}_{t-1}(\boldsymbol{x})\right|\leq\begin{cases}\tau_t\sigma_{t-1}(\boldsymbol{x})-\epsilon_{t-1}(\boldsymbol{x}), & \text{if }\ \epsilon_{t-1}(\boldsymbol{x})\geq 0\\ \tau_t\sigma_{t-1}(\boldsymbol{x})+\epsilon_{t-1}(\boldsymbol{x}), & \text{if }\ \epsilon_{t-1}(\boldsymbol{x})<0,\end{cases}$$

where $\tilde{\mu}_{t-1}(\boldsymbol{x})$ is the posterior mean of the GP with prior mean zero.

Proof. Fix $t\geq 1$ and $\boldsymbol{x}\in\mathcal{X}$. Completely analogous to [72], consider $\Psi(\boldsymbol{x})\sim N\left(\mu_{t-1}(\boldsymbol{x}),\sigma_{t-1}^2(\boldsymbol{x})\right)$ and deterministic $\left\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{t-1}\right\}$ conditioned on observed vector $\boldsymbol{y}_{t-1}$ in iteration $t-1$. Defining

$$\rho:=\left(\Psi(\boldsymbol{x})-\mu_{t-1}(\boldsymbol{x})\right)/\sigma_{t-1}(\boldsymbol{x}),$$

it holds $\rho\sim N(0,1)$ and thus

$$\begin{aligned}\mathbb{P}\{\rho>\tau_t\}&=e^{-\tau_t^2/2}(2\pi)^{-1/2}\int e^{-(\rho-\tau_t)^2/2-\tau_t(\rho-\tau_t)}d\rho\\&\leq e^{-\tau_t^2/2}\ \mathbb{P}\{\rho>0\}=(1/2)e^{-\tau_t^2/2}\end{aligned}$$

for $\tau_t>0$, since $e^{-\tau_t(\rho-\tau_t)}\leq 1$ for $\rho\geq\tau_t$. Therefore

$$\mathbb{P}\left\{\left|\Psi(\boldsymbol{x})-\mu_{t-1}(\boldsymbol{x})\right|>\tau_t\sigma_{t-1}(\boldsymbol{x})\right\}\leq e^{-\beta_t/2}.$$

Boole's inequality (by σ-subadditivity of $\mathbb{P}$) delivers that

$$\left|\Psi(\boldsymbol{x})-\mu_{t-1}(\boldsymbol{x})\right|\leq\tau_t\sigma_{t-1}(\boldsymbol{x})\quad\forall\boldsymbol{x}\in\mathcal{X}$$

holds with probability $\geq 1-|\mathcal{X}|e^{-\beta_t/2}$. Choosing $|\mathcal{X}|e^{-\beta_t/2}=\delta/\pi_t$ and using Boole's inequality for $t\in\mathbb{N}$, see proof of [72, Lemma 5.1], it holds that

$$\left|\Psi(\boldsymbol{x})-\mu_{t-1}(\boldsymbol{x})\right|\leq\tau_t\sigma_{t-1}(\boldsymbol{x})\quad\forall\boldsymbol{x}\in\mathcal{X}\ \forall t\geq 1$$

with probability $\geq 1-\delta$. Now use Eq. (8) to get

$$\left|\Psi(\boldsymbol{x})-\tilde{\mu}_{t-1}(\boldsymbol{x})+\epsilon_{t-1}(\boldsymbol{x})\right|\leq\tau_t\sigma_{t-1}(\boldsymbol{x})\quad\forall\boldsymbol{x}\in\mathcal{X}\ \forall t\geq 1$$

with probability $\geq 1-\delta$, from which the statement follows directly. □

Theorem 1 (*Regret Bound For Optimistic GP Misspecification*)**.** Let $\delta\in(0,1)$ and $\tau_t=\sqrt{2\log\left(|\mathcal{X}|t^2\pi^2/6\delta\right)}$ with finite $\mathcal{X}$. Bayesian optimization with a GP surrogate with prior mean function $m(\boldsymbol{x})$ inducing $\epsilon_T(\boldsymbol{x})$ (Eq. (8)) with $\forall t\ \forall\boldsymbol{x}:\ m(\boldsymbol{x})>\boldsymbol{k}_T(\boldsymbol{x})'\left(\boldsymbol{K}_T+\sigma^2\boldsymbol{I}\right)^{-1}m(\boldsymbol{x})$ has a cumulative regret R_T such that

$$\mathbb{P}\left\{R_T\leq\sqrt{T}\sqrt{\tau_T^2C_1\gamma_T+\mathscr{E}\left(2\tau_t\mathscr{S}+\mathscr{E}\right)}\quad\forall T\geq 1\right\}\geq 1-\delta,$$

where $C_1=8/\log\left(1+\sigma^{-2}\right)$, $\mathscr{S}=\sum_{t=1}^{T}\sigma_{t-1}\left(\boldsymbol{x}_t\right)$ the accumulated GP variances of BO proposals, and $\mathscr{E}=\sum_{t=1}^{T}\epsilon_{t-1}(\boldsymbol{x}_t)$ the accumulated prior-mean induced error terms.

Proof. It directly follows from the premise $\forall t\ \forall\boldsymbol{x}:\ m(\boldsymbol{x})>\boldsymbol{k}_T(\boldsymbol{x})'\left(\boldsymbol{K}_T+\sigma^2\boldsymbol{I}\right)^{-1}m(\boldsymbol{x})$ and from Lemma 2 that

$$\left|\Psi(\boldsymbol{x})-\mu_{t-1}(\boldsymbol{x})\right|\leq\tau_t\sigma_{t-1}(\boldsymbol{x})+\epsilon_{t-1}(\boldsymbol{x}).$$

Consider any $t\geq 1$ and set $\boldsymbol{x}_{opt}=\arg\min_{\boldsymbol{x}\in\mathcal{X}}\Psi(\boldsymbol{x})$. Because of $\boldsymbol{x}_t=\arg\max_{\boldsymbol{x}\in\mathcal{X}}\{-\mu_{t-1}(\boldsymbol{x})+\tau_t\cdot\sigma_{t-1}(\boldsymbol{x})\}$, we have

$$-\mu_{t-1}\left(\boldsymbol{x}_t\right)+\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right)\geq-\mu_{t-1}\left(\boldsymbol{x}_{opt}\right)+\tau_t\sigma_{t-1}\left(\boldsymbol{x}_{opt}\right).$$

Besides, we can set τ_t (and we will, later on) such that $-\mu_{t-1}\left(\boldsymbol{x}_{opt}\right)+\tau_t\sigma_{t-1}\left(\boldsymbol{x}_{opt}\right)\geq-\Psi\left(\boldsymbol{x}_{opt}\right)$. Combining all three inequalities gives for the instantaneous regret (Definition 8):

$$\begin{aligned}r_t&=\Psi(\boldsymbol{x}_t)-\min_{\boldsymbol{x}\in\mathcal{X}}\Psi(\boldsymbol{x})=\Psi(\boldsymbol{x}_t)-\Psi(\boldsymbol{x}_{opt})\\&\leq\Psi(\boldsymbol{x}_t)-\mu_{t-1}\left(\boldsymbol{x}_t\right)+\tau\sigma_{t-1}\left(\boldsymbol{x}_t\right)\\&\leq 2\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right)+\epsilon_{t-1}(\boldsymbol{x}_t).\end{aligned}$$

We will now show that the following holds with probability $\geq 1-\delta$:

$$\begin{aligned}\sum_{t=1}^{T}r_t^2\leq&\sum_{t=1}^{T}4\tau_T^2\sigma^2C_2\log\left(1+\sigma^{-2}\sigma_{t-1}^2\left(\boldsymbol{x}_t\right)\right)\\&+4\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right)\cdot\epsilon_{t-1}(\boldsymbol{x}_t)+\epsilon_{t-1}(\boldsymbol{x}_t)^2\quad\forall T\geq 1\end{aligned}$$

with $C_2=\sigma^{-2}/\log\left(1+\sigma^{-2}\right)$ in order to use the representation from Lemma 1 of the maximum information gain, see Definition 10. We have that

$$r_t^2\leq(2\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right)+\epsilon_{t-1}(\boldsymbol{x}_t))^2\ \forall t\geq 1$$

with probability $\geq 1-\delta$. Recall that we have set $\tau_t=\sqrt{2\log\left(|\mathcal{X}|t^2\pi^2/6\delta\right)}$, i.e., τ_t is non-decreasing in t. Hence,

$$\begin{aligned}\sum_{t=1}^{T}r_t^2&\leq\sum_{t=1}^{T}(2\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right)+\epsilon_{t-1}(\boldsymbol{x}_t))^2\\&\leq\sum_{t=1}^{T}(2\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right))^2+4\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right)\cdot\epsilon_{t-1}(\boldsymbol{x}_t)+\epsilon_{t-1}(\boldsymbol{x}_t)^2\\&\leq\sum_{t=1}^{T}(4\tau_T^2\sigma^2\left(\sigma^{-2}\sigma_{t-1}^2\left(\boldsymbol{x}_t\right)\right)+4\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right)\cdot\epsilon_{t-1}(\boldsymbol{x}_t)+\epsilon_{t-1}(\boldsymbol{x}_t)^2\\&\leq\sum_{t=1}^{T}4\tau_T^2\sigma^2C_2\log\left(1+\sigma^{-2}\sigma_{t-1}^2\left(\boldsymbol{x}_t\right)\right)\\&\qquad+4\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right)\cdot\epsilon_{t-1}(\boldsymbol{x}_t)+\epsilon_{t-1}(\boldsymbol{x}_t)^2\end{aligned}$$

with probability $\geq 1-\delta$. We further used $C_2 = \sigma^{-2}/\log\left(1+\sigma^{-2}\right) \geq 1$, since $s^2 \leq C_2 \log\left(1+s^2\right)$ for $s \in \left[0,\sigma^{-2}\right]$, and $\sigma^{-2}\sigma^2_{t-1}\left(\boldsymbol{x}_t\right) \leq \sigma^{-2}k\left(\boldsymbol{x}_t,\boldsymbol{x}_t\right) \leq \sigma^{-2}$, see also [72, Lemma 5.4]. We are now using the following representation of the information gain from Lemma 1

$$\mathrm{I}\left(\boldsymbol{y}_T,\boldsymbol{\Psi}_T\right) = \frac{1}{2}\sum_{t=1}^{T}\log\left(1+\sigma^{-2}\sigma^2_{t-1}\left(\boldsymbol{x}_t\right)\right),$$

see [72, Lemma 5.3]. In the following, denote by $\mathcal{S} = \sum_{t=1}^{T}\sigma_{t-1}\left(\boldsymbol{x}_t\right)$ and $\mathcal{E} = \sum_{t=1}^{T}\epsilon_{t-1}(\boldsymbol{x}_t)$ the accumulated variances and prior-mean induced error terms. Recall Definition 10 of the maximum information gain $\gamma_t := \max \mathrm{I}\left(\boldsymbol{y}_T;\boldsymbol{f}_T\right)$ to get

$$\begin{aligned}\sum_{t=1}^{T} r_t^2 \leq \sum_{t=1}^{T} & 4\tau_T^2\sigma^2 C_2 \log\left(1+\sigma^{-2}\sigma^2_{t-1}\left(\boldsymbol{x}_t\right)\right)\\ &+ 4\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right)\cdot\epsilon_{t-1}(\boldsymbol{x}_t)+\epsilon_{t-1}(\boldsymbol{x}_t)^2\\ \leq\, & 8\tau_T^2\sigma^2C_2\gamma_T + 4\mathcal{S}\tau_t\mathcal{E}+\mathcal{E}^2\\ \leq\, & \tau_T^2 C_1\gamma_T + \mathcal{E}\left(4\mathcal{S}\tau_t+\mathcal{E}\right)\end{aligned}$$

with probability $\geq 1-\delta$ where we used $C_1 = 8\sigma^2 C_2$ and the trivial fact that the information gain is upper-bounded by the maximum information gain. Eventually, note that $R_T^2 \leq T\sum_{t=1}^{T} r_t^2$ by Cauchy–Schwarz. We thus have

$$R_T \leq \sqrt{T}\sqrt{\tau_T^2 C_1\gamma_T + \mathcal{E}\left(4\tau_t\mathcal{S}+\mathcal{E}\right)},$$

with probability $\geq 1-\delta$, which was to be demonstrated. □

Theorem 2 (*Regret Bound For Sub-Variance GP Misspecification*)**.** Let $\delta \in (0,1)$ and $\tau_t = \sqrt{2\log\left(|\mathcal{X}|t^2\pi^2/6\delta\right)}$ with finite $\mathcal{X}$. Bayesian optimization with a GP surrogate with prior mean function $m(\boldsymbol{x})$ inducing sub-variance error $\epsilon_T(\boldsymbol{x})$ (Eq. (8)) s.t. $\forall T\ :\ \epsilon_T(\boldsymbol{x}) \leq \sigma_T(\boldsymbol{x})$ with $\forall t\ \forall \boldsymbol{x}\ :\ m(\boldsymbol{x}) > \boldsymbol{k}_T(\boldsymbol{x})'\left(\boldsymbol{K}_T+\sigma^2\boldsymbol{I}\right)^{-1}m(\boldsymbol{x})$ has a cumulative regret R_T such that

$$\mathbb{P}\left\{R_T \leq \sqrt{T(8\tau_T^2+8\tau_T+2)\gamma_T/\log(1+\sigma^{-2})}\quad \forall T\geq 1\right\} \geq 1-\delta.$$

Proof. By exploiting Lemma 2, the proof of Theorem 1 entails

$$\begin{aligned}r_t &= \Psi(\boldsymbol{x}_t) - \min_{\boldsymbol{x}\in\mathcal{X}}\Psi(\boldsymbol{x}) = \Psi(\boldsymbol{x}_t) - \Psi(\boldsymbol{x}_{opt})\\ &\leq \Psi(\boldsymbol{x}_t) - \mu_{t-1}\left(\boldsymbol{x}_t\right) + \tau\sigma_{t-1}\left(\boldsymbol{x}_t\right)\\ &\leq 2\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right) + \epsilon_{t-1}(\boldsymbol{x}_t).\end{aligned}$$

We will now show that the following holds with probability $\geq 1-\delta$:

$$\sum_{t=1}^{T} r_t^2 \leq (8\tau_T^2+4\tau_T+2)\sigma^2 C_2\gamma_T \quad \forall T\geq 1$$

with $C_2 = \sigma^{-2}/\log\left(1+\sigma^{-2}\right)$) and γ_t the maximum information gain, see Definition 10. We have that

$$r_t^2 \leq (2\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right)+\epsilon_{t-1}(\boldsymbol{x}_t))^2\ \forall t\geq 1$$

with probability $\geq 1-\delta$. Recall that we have set $\tau_t = 4\log\left(|\mathcal{X}|t^2\pi^2/6\delta\right)^2$, i.e., τ_t is non-decreasing in t. Hence,

$$\begin{aligned}\sum_{t=1}^{T} r_t^2 &\leq \sum_{t=1}^{T}(2\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right)+\epsilon_{t-1}(\boldsymbol{x}_t))^2\\ &\leq \sum_{t=1}^{T}(2\tau_T\sigma_{t-1}\left(\boldsymbol{x}_t\right))^2 + 4\tau_T\sigma_{t-1}\left(\boldsymbol{x}_t\right)\cdot\epsilon_{t-1}(\boldsymbol{x}_t)+\epsilon_{t-1}(\boldsymbol{x}_t)^2\\ &\leq \sum_{t=1}^{T} 4\tau_T^2\sigma^2_{t-1}(\boldsymbol{x}_t) + 4\tau_T\sigma^2_{t-1}\left(\boldsymbol{x}_t\right)+\sigma^2_{t-1}\left(\boldsymbol{x}_t\right),\end{aligned}$$

with probability $\geq 1-\delta$ using the critical assumption $\forall t\ :\ \sigma_{t-1}\left(\boldsymbol{x}_t\right) \leq \epsilon_{t-1}(\boldsymbol{x}_t)$. Analogous to the proof of Theorem 1, we can now exploit $C_2 = \sigma^{-2}/\log\left(1+\sigma^{-2}\right) \geq 1$, and $\sigma^{-2}\sigma^2_{t-1}\left(\boldsymbol{x}_t\right) \leq \sigma^{-2}k\left(\boldsymbol{x}_t,\boldsymbol{x}_t\right) \leq \sigma^{-2}$ again, see also [72, Lemma 5.4]. This yields

$$\begin{aligned}&\sum_{t=1}^{T} 4\tau_T^2\sigma^2_{t-1}(\boldsymbol{x}_t) + 4\tau_T\sigma^2_{t-1}\left(\boldsymbol{x}_t\right)+\sigma^2_{t-1}\left(\boldsymbol{x}_t\right),\\ =&\sum_{t=1}^{T}\sigma^2\left(\sigma^{-2}\sigma^2_{t-1}\left(\boldsymbol{x}_t\right)\right)\left(4\tau_T^2+4\tau_T+1\right)\\ \leq&\sum_{t=1}^{T}\sigma^2 C_2\log\left(1+\sigma^{-2}\sigma^2_{t-1}\left(\boldsymbol{x}_t\right)\right)\left(2\tau_T+1\right)^2\end{aligned}$$

with probability $\geq 1-\delta$. Just like in the proof of Theorem 1, we are now using the representation of the information gain $\mathrm{I}\left(\boldsymbol{y}_T,\boldsymbol{\Psi}_T\right) = \frac{1}{2}\sum_{t=1}^{T}\log\left(1+\sigma^{-2}\sigma^2_{t-1}\left(\boldsymbol{x}_t\right)\right)$. Furthermore, recall Definition 10 of the maximum information gain $\gamma_t := \max \mathrm{I}\left(\boldsymbol{y}_T;\boldsymbol{f}_T\right)$ once more to get

$$\begin{aligned}\sum_{t=1}^{T} r_t^2 \leq \sum_{t=1}^{T} & \sigma^2 C_2\log\left(1+\sigma^{-2}\sigma^2_{t-1}\left(\boldsymbol{x}_t\right)\right)\left(4\tau_T^2+4\tau_T+1\right)\\ &\leq (8\tau_T^2+8\tau_T+2)\sigma^2 C_2\gamma_T\end{aligned}$$

with probability $\geq 1-\delta$ where we again used the trivial fact that the information gain is upper-bounded by the maximum information gain. Finally, recall that $R_T^2 \leq T\sum_{t=1}^{T} r_t^2$ by Cauchy–Schwarz. We thus have

$$R_T \leq \sqrt{T(8\tau_T^2+8\tau_T+2)\sigma^2C_2\gamma_T},$$

with probability $\geq 1-\delta$, or

$$\mathbb{P}\left\{R_T \leq \sqrt{T(8\tau_T^2+8\tau_T+2)\gamma_T/\log(1+\sigma^{-2})}\right\} \geq 1-\delta,$$

which was to be demonstrated. □

Theorem 3 (*Regret Bound For Optimistic GP Misspecification on Infinite $\mathcal{X}$*)**.** Let $\mathcal{X} \subset [0,r]^d$ be compact and convex, $d\in\mathbb{N}, r\in\mathbb{R}_{\geq 0}$. Fix $\delta\in(0,1)$, and set

$$\tau_t^2 = 2\log\left(t^2 2\pi^2/(3\delta)\right) + 2d\log\left(t^2 dbr\sqrt{\log(4da/\delta)}\right)$$

with a, b as in condition 1. If Bayesian optimization with misspecified prior mean inducing $\forall t\ :\ \epsilon_t(\boldsymbol{x}) > 0$ is run on Ψ that satisfies condition 1, we obtain the following cumulative regret bound

$$\mathbb{P}\left\{R_T \leq \sqrt{\tau_T^2 C_1\gamma_T + (\mathcal{E}+1)\left(2\mathcal{S}\tau_t+\mathcal{E}\right)+\frac{\pi^2}{6}}\quad \forall T\geq 1\right\} \geq 1-\delta$$

with $C_1 = 8/\log\left(1+\sigma^{-2}\right)$ as in Theorem 1, and $\mathcal{S} = \sum_{t=1}^{T}\sigma_{t-1}\left(\boldsymbol{x}_t\right)$, and $\mathcal{E} = \sum_{t=1}^{T}\epsilon_{t-1}(\boldsymbol{x}_t)$ the accumulated prior-mean induced error terms.

Proof. Based on the discretization of $\mathcal{X}$ in [72, Lemmas 5.6 and 5.7] and [72, Lemma 5.8] it holds that $\forall t\geq 1$:

$$r_t \leq 2\tau_t\sigma_{t-1}(\boldsymbol{x}) + \frac{1}{t^2},$$

in the case of zero-mean GP. We can directly apply the discretization technique from [72] to our instantaneous regret bound $r_t \leq 2\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right) + \epsilon_{t-1}(\boldsymbol{x}_t)$ from the case of finite $\mathcal{X}$ to get the following instantaneous regret bound for convex and compact $\mathcal{X}$:

$$r_t \leq 2\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right) + \epsilon_{t-1}(\boldsymbol{x}_t) + \frac{1}{t^2}, \forall t\geq 1,$$

with probability greater than $1-\delta$. Thus,

$$r_t^2 \leq \left(2\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right)+\epsilon_{t-1}(\boldsymbol{x}_t)+\frac{1}{t^2}\right)^2\ \forall t\geq 1$$

with probability greater than $1-\delta$. Now we are ready to complete the proof in analogy to the proof of Theorem 1, just that we have to expand a trinomial instead of a binomial.

$$\begin{aligned}\sum_{t=1}^{T} r_t^2 \leq \sum_{t=1}^{T} & 4\tau_T^2\sigma^2 C_2 \log\left(1+\sigma^{-2}\sigma_{t-1}^2\left(\boldsymbol{x}_t\right)\right) \\ & + 4\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right)\cdot\epsilon_{t-1}(\boldsymbol{x}_t)+\epsilon_{t-1}(\boldsymbol{x}_t)^2 \\ & + \frac{4\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right)}{t^2}+\frac{2\epsilon_{t-1}(\boldsymbol{x}_t)}{t^2}+\frac{1}{t^4} \\ \leq & 8\tau_T^2\sigma^2 C_2\gamma_T + 4\mathcal{S}\tau_t\mathcal{E}+\mathcal{E}^2+4\mathcal{S}\tau_t+2\mathcal{E}+\sum\frac{1}{t^4} \\ \leq & \tau_T^2 C_1\gamma_T+\mathcal{E}\left(4\mathcal{S}\tau_t+\mathcal{E}+2\right)+4\mathcal{S}\tau_t+\frac{\pi^2}{6},\end{aligned}$$

utilizing $t \geq 1$, $C_1 = 8\sigma^2 C_2$, and $\sum\frac{1}{t^2} = \frac{\pi^2}{6}$ (Euler's solution to the Basel problem). The assertion now follows by the Cauchy–Schwarz inequality. □

Theorem 4 (*Regret Bound For Sub-Variance GP Misspecification on Infinite* $\mathcal{X}$)**.** Let $\mathcal{X} \subset [0,r]^d$ be compact and convex, $d \in \mathbb{N}, r \in \mathbb{R}_{\geq 0}$. Fix $\delta \in (0,1)$, and set

$$\tau_t^2 = 2\log\left(t^2 2\pi^2/(3\delta)\right)+2d\log\left(t^2 dbr\sqrt{\log(4da/\delta)}\right)$$

with a, b as in condition 1. If Bayesian optimization with misspecified prior mean inducing sub-variance error $\epsilon_T(\boldsymbol{x})$, i.e., $\forall T : \epsilon_T(\boldsymbol{x}) \leq \sigma_T(\boldsymbol{x})$ with $\forall t : \epsilon_t(\boldsymbol{x}) > 0$ is run on target function Ψ that satisfies condition 1, we obtain the following cumulative regret bound.

$$\mathbb{P}\left\{R_T \leq \sqrt{(8\tau_T^2+4\tau_T+2)\gamma_T/\log\left(1+\sigma^{-2}\right)}+\frac{\pi^2}{6} \quad \forall T \geq 1\right\} \geq 1-\delta.$$

Proof. From the proof of Theorem 2 we have the following $\forall T \geq 1$:

$$\begin{aligned}\sum_{t=1}^{T} r_t^2 \leq & \sum_{t=1}^{T}(2\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right)+\epsilon_{t-1}(\boldsymbol{x}_t))^2 \\ \leq & \sum_{t=1}^{T}\sigma^2 C_2\log\left(1+\sigma^{-2}\sigma_{t-1}^2\left(\boldsymbol{x}_t\right)\right)\left(2\tau_T+1\right)^2 \\ \leq & (8\tau_T^2+8\tau_T+2)\sigma^2 C_2\gamma_T\end{aligned}$$

with probability $\geq 1-\delta$ where the critical assumption $\forall T : \epsilon_T(\boldsymbol{x}) \leq \sigma_T(\boldsymbol{x})$ was used as well as the representation of the information gain $\mathrm{I}\left(\boldsymbol{y}_T, \boldsymbol{\Psi}_T\right) = \frac{1}{2}\sum_{t=1}^{T}\log\left(1+\sigma^{-2}\sigma_{t-1}^2\left(\boldsymbol{x}_t\right)\right)$, and $C_2 = \sigma^{-2}/\log\left(1+\sigma^{-2}\right)$, see the proof of Theorem 1. By Cauchy–Schwarz this implies

$$\sum_{t=1}^{T} 2\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right)+\epsilon_{t-1}(\boldsymbol{x}_t) \leq \sqrt{(8\tau_T^2+8\tau_T+2)\sigma^2 C_2\gamma_T} \quad \forall T \geq 1 \tag{A.1}$$

From the proof of Theorem 3 we further have for the regret in the case of infinite $\mathcal{X}$:

$$\mathbb{P}\left\{r_t \leq 2\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right)+\epsilon_{t-1}(\boldsymbol{x}_t)+\frac{1}{t^2} \quad \forall t \geq 1\right\} \geq 1-\delta.$$

Summing over and using Eq. (A.1) now directly delivers that

$$\mathbb{P}\left\{\sum_{t=1}^{T} r_t \leq \sqrt{(8\tau_T^2+8\tau_T+2)\sigma^2 C_2\gamma_T}+\frac{\pi^2}{6} \quad \forall T \geq 1\right\} \geq 1-\delta,$$

where we again use Euler's solution to the Basel problem: $\sum\frac{1}{t^2} = \frac{\pi^2}{6}$. With $C_2 = \sigma^{-2}/\log\left(1+\sigma^{-2}\right)$ and $R_T = \sum_{t=1}^{T} r_t$ the assertion follows. □

Appendix B. Regret bounds for pessimistic GP misspecification

Crucially, the cumulative regret bound for pessimistic Gaussian process misspecification s.t. $\forall t\ \forall\boldsymbol{x} : m(\boldsymbol{x}) < \boldsymbol{k}_T(\boldsymbol{x})'\left(\boldsymbol{K}_T+\sigma^2\boldsymbol{I}\right)^{-1} m(\boldsymbol{x})$ are of the same order as the ones for optimistic GP, namely linear in $\mathcal{E}$:

$$\mathcal{O}\left(\sqrt{T}\sqrt{\gamma_T\log|\mathcal{X}|-\mathcal{E}4\tau_t\mathcal{S}+\mathcal{E}^2}\right). \tag{B.1}$$

For the sake of completeness, we provide the respective theorem in what follows. The proof is equivalent to the proof of Theorem 1 up to the sign of $\epsilon_{t-1}(\boldsymbol{x}_t)$.

Theorem 5 (*Regret Bound For Pessimistic GP Misspecification*)**.** *Let* $\delta \in (0,1)$ *and* $\tau_t = \sqrt{2\log\left(|\mathcal{X}|t^2\pi^2/6\delta\right)}$*. In case of finite* $\mathcal{X}$*, Bayesian optimization with a GP surrogate with prior mean function* $m(\boldsymbol{x})$ *inducing* $\epsilon_T(\boldsymbol{x})$ *(Eq. (8)) with* $\forall t\ \forall\boldsymbol{x} : m(\boldsymbol{x}) < \boldsymbol{k}_T(\boldsymbol{x})'\left(\boldsymbol{K}_T+\sigma^2\boldsymbol{I}\right)^{-1} m(\boldsymbol{x})$ *has a cumulative regret* R_T *such that*

$$\mathbb{P}\left\{R_T \leq \sqrt{T}\sqrt{\tau_T^2 C_1\gamma_T-\mathcal{E}\left(4\tau_t\mathcal{S}-\mathcal{E}\right)} \quad \forall T \geq 1\right\} \geq 1-\delta,$$

where $C_1 = 8/\log\left(1+\sigma^{-2}\right)$*,* $\mathcal{S} = \sum_{t=1}^{T}\sigma_{t-1}\left(\boldsymbol{x}_t\right)$ *the accumulated GP variances of BO proposals, and* $\mathcal{E} = \sum_{t=1}^{T}\epsilon_{t-1}(\boldsymbol{x}_t)$ *the accumulated prior-mean induced error terms.*

Proof. It directly follows from the premise $\forall t\ \forall\boldsymbol{x} : m(\boldsymbol{x}) > \boldsymbol{k}_T(\boldsymbol{x})'\left(\boldsymbol{K}_T+\sigma^2\boldsymbol{I}\right)^{-1} m(\boldsymbol{x})$ and from Lemma 2 that

$$\left|\Psi(\boldsymbol{x})-\mu_{t-1}(\boldsymbol{x})\right| \leq \tau_t\sigma_{t-1}(\boldsymbol{x})-\epsilon_{t-1}(\boldsymbol{x}).$$

Consider any $t \geq 1$ and set $\boldsymbol{x}_{opt} = \arg\min_{\boldsymbol{x}\in\mathcal{X}}\Psi(\boldsymbol{x})$. Because of $\boldsymbol{x}_t = \arg\max_{\boldsymbol{x}\in\mathcal{X}}\{-\mu_{t-1}(\boldsymbol{x})+\tau_t\cdot\sigma_{t-1}(\boldsymbol{x})\}$, we have

$$-\mu_{t-1}\left(\boldsymbol{x}_t\right)+\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right) \geq -\mu_{t-1}\left(\boldsymbol{x}_{opt}\right)+\tau_t\sigma_{t-1}\left(\boldsymbol{x}_{opt}\right).$$

Besides, we can set τ_t (and we will, later on) such that $-\mu_{t-1}\left(\boldsymbol{x}_{opt}\right)+\tau_t\sigma_{t-1}\left(\boldsymbol{x}_{opt}\right) \geq -\Psi\left(\boldsymbol{x}_{opt}\right)$. Combining all three inequalities gives for the instantaneous regret (Definition 8):

$$\begin{aligned}r_t &= \Psi(\boldsymbol{x}_t)-\min_{\boldsymbol{x}\in\mathcal{X}}\Psi(\boldsymbol{x}) = \Psi(\boldsymbol{x}_t)-\Psi(\boldsymbol{x}_{opt}) \\ &\leq \Psi(\boldsymbol{x}_t)-\mu_{t-1}\left(\boldsymbol{x}_t\right)+\tau\sigma_{t-1}\left(\boldsymbol{x}_t\right) \\ &\leq 2\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right)-\epsilon_{t-1}(\boldsymbol{x}_t).\end{aligned}$$

The strategy will now be to show that the following holds with probability $\geq 1-\delta$:

$$\begin{aligned}\sum_{t=1}^{T} r_t^2 \leq & \sum_{t=1}^{T} 4\tau_T^2\sigma^2 C_2\log\left(1+\sigma^{-2}\sigma_{t-1}^2\left(\boldsymbol{x}_t\right)\right) \\ & - 4\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right)\cdot\epsilon_{t-1}(\boldsymbol{x}_t)+\epsilon_{t-1}(\boldsymbol{x}_t)^2 \quad \forall T \geq 1\end{aligned}$$

with $C_2 = \sigma^{-2}/\log\left(1+\sigma^{-2}\right)$ in order to use the representation from Lemma 1 of the maximum information gain, see Definition 10. We have that

$$r_t^2 \leq (2\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right)-\epsilon_{t-1}(\boldsymbol{x}_t))^2\ \forall t \geq 1$$

with probability $\geq 1-\delta$. Recall that we have set $\tau_t = \sqrt{2\log\left(|\mathcal{X}|t^2\pi^2/6\delta\right)}$, i.e., τ_t is non-decreasing in t. Hence,

$$\begin{aligned}\sum_{t=1}^{T} r_t^2 \leq & \sum_{t=1}^{T}(2\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right)-\epsilon_{t-1}(\boldsymbol{x}_t))^2 \\ \leq & \sum_{t=1}^{T}(2\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right))^2-4\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right)\cdot\epsilon_{t-1}(\boldsymbol{x}_t)+\epsilon_{t-1}(\boldsymbol{x}_t)^2 \\ \leq & \sum_{t=1}^{T}(4\tau_T^2\sigma^2\left(\sigma^{-2}\sigma_{t-1}^2\left(\boldsymbol{x}_t\right)\right)-4\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right)\cdot\epsilon_{t-1}(\boldsymbol{x}_t)+\epsilon_{t-1}(\boldsymbol{x}_t)^2 \\ \leq & \sum_{t=1}^{T} 4\tau_T^2\sigma^2 C_2\log\left(1+\sigma^{-2}\sigma_{t-1}^2\left(\boldsymbol{x}_t\right)\right) \\ & - 4\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right)\cdot\epsilon_{t-1}(\boldsymbol{x}_t)+\epsilon_{t-1}(\boldsymbol{x}_t)^2\end{aligned}$$

with probability $\geq 1-\delta$. We further used $C_2 = \sigma^{-2}/\log\left(1+\sigma^{-2}\right) \geq 1$, since $s^2 \leq C_2\log\left(1+s^2\right)$ for $s \in \left[0,\sigma^{-2}\right]$, and $\sigma^{-2}\sigma_{t-1}^2\left(x_t\right) \leq \sigma^{-2}k\left(x_t,x_t\right) \leq \sigma^{-2}$, see also [72, Lemma 5.4]. We are now using the following representation of the information gain from Lemma 1

$$\mathrm{I}\left(\boldsymbol{y}_T,\boldsymbol{\Psi}_T\right) = \frac{1}{2}\sum_{t=1}^{T}\log\left(1+\sigma^{-2}\sigma_{t-1}^2\left(\boldsymbol{x}_t\right)\right),$$

see [72, Lemma 5.3]. In the following, denote by $\mathcal{S} = \sum_{t=1}^{T} \sigma_{t-1}\left(\boldsymbol{x}_t\right)$ and $\mathcal{E} = \sum_{t=1}^{T} \epsilon_{t-1}(\boldsymbol{x}_t)$ the accumulated variances and prior-mean induced error terms. Recall Definition 10 of the maximum information gain $\gamma_t := \max \mathrm{I}\left(\boldsymbol{y}_T; \boldsymbol{f}_T\right)$ to get

$$\begin{aligned}\sum_{t=1}^{T} r_t^2 \leq \sum_{t=1}^{T} & 4\tau_T^2\sigma^2 C_2 \log\left(1+\sigma^{-2}\sigma_{t-1}^2\left(\boldsymbol{x}_t\right)\right) \\ & - 4\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right)\cdot\epsilon_{t-1}(\boldsymbol{x}_t) + \epsilon_{t-1}(\boldsymbol{x}_t)^2 \\ \leq & 8\tau_T^2\sigma^2 C_2\gamma_T - 4\mathcal{S}\tau_t\mathcal{E} + \mathcal{E}^2 \\ \leq & \tau_T^2 C_1\gamma_T - \mathcal{E}\left(4\mathcal{S}\tau_t - \mathcal{E}\right)\end{aligned}$$

with probability $\geq 1-\delta$ where we used $C_1 = 8\sigma^2 C_2$ and the trivial fact that the information gain is upper-bounded by the maximum information gain. Eventually, note that $R_T^2 \leq T\sum_{t=1}^{T} r_t^2$ by Cauchy–Schwarz. We thus have

$$R_T \leq \sqrt{T}\sqrt{\tau_T^2 C_1\gamma_T - \mathcal{E}\left(4\tau_t\mathcal{S} - \mathcal{E}\right)},$$

with probability $\geq 1-\delta$, which was to be demonstrated. □

Similar reasoning applies to the case of sub-variance GP misspecification.

Theorem 6 (*Regret Bound For Pessimistic Sub-Variance GP Misspecification*). *Let $\delta \in (0,1)$ and $\tau_t = \sqrt{2\log\left(|\mathcal{X}|t^2\pi^2/6\delta\right)}$. In case of finite $\mathcal{X}$, Bayesian optimization with a GP surrogate with prior mean function $m(\boldsymbol{x})$ inducing sub-variance error $\epsilon_T(\boldsymbol{x})$ (Eq. (8)) s.t. $\forall T:\ \epsilon_T(\boldsymbol{x}) \leq \sigma_T(\boldsymbol{x})$ with $\forall t\ \forall \boldsymbol{x}: m(\boldsymbol{x}) < \boldsymbol{k}_T(\boldsymbol{x})'\left(\boldsymbol{K}_T + \sigma^2\boldsymbol{I}\right)^{-1} m(\boldsymbol{x})$ has a cumulative regret R_T such that*

$$\mathbb{P}\left\{R_T \leq \sqrt{T(8\tau_T^2 - 8\tau_T + 2)\gamma_T/\log(1+\sigma^{-2})} \quad \forall T \geq 1\right\} \geq 1-\delta.$$

Again, this regret bound is of the same order as the one for optimistic sub-variance GP misspecification:

$$\mathcal{O}\left(\sqrt{T\gamma_T\log|\mathcal{X}|}\right). \tag{B.2}$$

Proof. By exploiting Lemma 2, the proof of Theorem 1 entails

$$\begin{aligned} r_t &= \Psi(\boldsymbol{x}_t) - \min_{\boldsymbol{x}\in\mathcal{X}}\Psi(\boldsymbol{x}) = \Psi(\boldsymbol{x}_t) - \Psi(\boldsymbol{x}_{opt}) \\ &\leq \Psi(\boldsymbol{x}_t) - \mu_{t-1}\left(\boldsymbol{x}_t\right) + \tau\sigma_{t-1}\left(\boldsymbol{x}_t\right) \\ &\leq 2\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right) - \epsilon_{t-1}(\boldsymbol{x}_t).\end{aligned}$$

We will now show that the following holds with probability $\geq 1-\delta$:

$$\sum_{t=1}^{T} r_t^2 \leq (8\tau_T^2 + 4\tau_T + 2)\sigma^2 C_2\gamma_T \quad \forall T \geq 1$$

with $C_2 = \sigma^{-2}/\log\left(1+\sigma^{-2}\right)$) and γ_t the maximum information gain, see Definition 10. We have that

$$r_t^2 \leq (2\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right) - \epsilon_{t-1}(\boldsymbol{x}_t))^2\ \forall t \geq 1$$

with probability $\geq 1-\delta$. Recall that we have set $\tau_t = 4\log\left(|\mathcal{X}|t^2\pi^2/6\delta\right)^2$, i.e., τ_t is non-decreasing in t. Hence,

$$\begin{aligned}\sum_{t=1}^{T} r_t^2 &\leq \sum_{t=1}^{T}(2\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right) - \epsilon_{t-1}(\boldsymbol{x}_t))^2 \\ &\leq \sum_{t=1}^{T}(2\tau_T\sigma_{t-1}\left(\boldsymbol{x}_t\right))^2 - 4\tau_T\sigma_{t-1}\left(\boldsymbol{x}_t\right)\cdot\epsilon_{t-1}(\boldsymbol{x}_t) + \epsilon_{t-1}(\boldsymbol{x}_t)^2 \\ &\leq \sum_{t=1}^{T} 4\tau_T^2\sigma_{t-1}^2(\boldsymbol{x}_t) - 4\tau_T\sigma_{t-1}^2\left(\boldsymbol{x}_t\right) + \sigma_{t-1}^2\left(\boldsymbol{x}_t\right),\end{aligned}$$

with probability $\geq 1-\delta$ using the critical assumption $\forall t:\ \sigma_{t-1}\left(\boldsymbol{x}_t\right) \leq \epsilon_{t-1}(\boldsymbol{x}_t)$. Analogous to the proof of Theorem 1, we can now exploit $C_2 = \sigma^{-2}/\log\left(1+\sigma^{-2}\right) \geq 1$, and $\sigma^{-2}\sigma_{t-1}^2\left(x_t\right) \leq \sigma^{-2}k\left(x_t, x_t\right) \leq \sigma^{-2}$ again, see also [72, Lemma 5.4]. This yields

$$\begin{aligned} &\sum_{t=1}^{T} 4\tau_T^2\sigma_{t-1}^2(\boldsymbol{x}_t) - 4\tau_T\sigma_{t-1}^2\left(\boldsymbol{x}_t\right) + \sigma_{t-1}^2\left(\boldsymbol{x}_t\right), \\ = &\sum_{t=1}^{T}\sigma^2\left(\sigma^{-2}\sigma_{t-1}^2\left(\boldsymbol{x}_t\right)\right)\left(4\tau_T^2 - 4\tau_T + 1\right) \\ \leq &\sum_{t=1}^{T}\sigma^2 C_2\log\left(1+\sigma^{-2}\sigma_{t-1}^2\left(\boldsymbol{x}_t\right)\right)\left(2\tau_T - 1\right)^2\end{aligned}$$

with probability $\geq 1-\delta$. Just like in the proof of Theorem 1, we are now using the representation of the information gain $\mathrm{I}\left(\boldsymbol{y}_T, \boldsymbol{\Psi}_T\right) = \frac{1}{2}\sum_{t=1}^{T}\log\left(1+\sigma^{-2}\sigma_{t-1}^2\left(\boldsymbol{x}_t\right)\right)$. Furthermore, recall Definition 10 of the maximum information gain $\gamma_t := \max \mathrm{I}\left(\boldsymbol{y}_T; \boldsymbol{f}_T\right)$ once more to get

$$\begin{aligned}\sum_{t=1}^{T} r_t^2 \leq &\sum_{t=1}^{T}\sigma^2 C_2\log\left(1+\sigma^{-2}\sigma_{t-1}^2\left(\boldsymbol{x}_t\right)\right)\left(4\tau_T^2 - 4\tau_T + 1\right) \\ &\leq (8\tau_T^2 - 8\tau_T + 2)\sigma^2 C_2\gamma_T\end{aligned}$$

with probability $\geq 1-\delta$ where we again used the trivial fact that the information gain is upper-bounded by the maximum information gain. Finally, recall that $R_T^2 \leq T\sum_{t=1}^{T} r_t^2$ by Cauchy–Schwarz. We thus have

$$R_T \leq \sqrt{T(8\tau_T^2 - 8\tau_T + 2)\sigma^2 C_2\gamma_T},$$

or

$$R_T \leq \sqrt{T(8\tau_T^2 - 8\tau_T + 2)\gamma_T/\log(1+\sigma^{-2})},$$

with probability $\geq 1-\delta$, which was to be demonstrated. □

Considering infinite $\mathcal{X}$ does not change the order of the regret bounds either. The proofs are equivalent to the proofs of Theorem 3 and Theorem 4 up to the sign of the misspecification-induced error.

Theorem 7 (*Regret Bound For Pessimistic GP Misspecification on Infinite $\mathcal{X}$*). *Let $\mathcal{X} \subset [0,r]^d$ be compact and convex, $d \in \mathbb{N}, r \in \mathbb{R}_{\geq 0}$. Fix $\delta \in (0,1)$, and set*

$$\tau_t^2 = 2\log\left(t^2 2\pi^2/(3\delta)\right) + 2d\log\left(t^2 dbr\sqrt{\log(4da/\delta)}\right)$$

with a, b as in condition 1. If Bayesian optimization with misspecified prior mean inducing $\forall t:\ \epsilon_t(\boldsymbol{x}) < 0$ is run on Ψ that satisfies condition 1, we obtain the following cumulative regret bound

$$\mathbb{P}\left\{R_T \leq \sqrt{\tau_T^2 C_1\gamma_T - \mathcal{E}\left(4\mathcal{S}\tau_t - \mathcal{E} + 2\right) + 4\mathcal{S}\tau_t + \frac{\pi^2}{6}} \quad \forall T \geq 1\right\} \geq 1-\delta$$

with $C_1 = 8/\log\left(1+\sigma^{-2}\right)$ as in Theorem 1, and $\mathcal{S} = \sum_{t=1}^{T}\sigma_{t-1}\left(\boldsymbol{x}_t\right)$, and $\mathcal{E} = \sum_{t=1}^{T}\epsilon_{t-1}(\boldsymbol{x}_t)$ the accumulated prior-mean induced error terms.

The idea of the proof is to show Lemma 2 $\forall t \geq 1$ and fixed x instead of $\forall \boldsymbol{x} \in \mathcal{X}\ \forall t \geq 1$. Then consider a discretization $\mathcal{X}_t \subset \mathcal{X}$ for each t in order to prove Lemma 2 $\forall \boldsymbol{x} \in \mathcal{X}_t\ \forall t \leq 1$ and then let $\mathcal{X}_t$ get dense as t gets large. Note that the cumulative regret bound remains linear in $\mathcal{E}$ like in the finite $\mathcal{X}$ case, see Eq. (11).

Proof. Based on the discretization of $\mathcal{X}$ in [72, Lemmas 5.6 and 5.7] and [72, Lemma 5.8] it holds that $\forall t \geq 1$:

$$r_t \leq 2\tau_t\sigma_{t-1}(\boldsymbol{x}) + \frac{1}{t^2},$$

in the case of zero-mean GP. We can directly apply the discretization technique from [72] to our instantaneous regret bound $r_t \leq 2\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right) - \epsilon_{t-1}(\boldsymbol{x}_t)$ from the case of finite $\mathcal{X}$ to get the following instantaneous regret bound for convex and compact $\mathcal{X}$:

$$r_t \leq 2\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right) - \epsilon_{t-1}(\boldsymbol{x}_t) + \frac{1}{t^2}, \forall t \geq 1,$$

with probability greater than $1-\delta$. Thus,

$$r_t^2 \leq \left(2\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right) - \epsilon_{t-1}(\boldsymbol{x}_t) + \frac{1}{t^2}\right)^2\ \forall t \geq 1$$

with probability greater than $1-\delta$. Now we are ready to complete the proof in analogy to the proof of Theorem 1, just that we have to expand a trinomial instead of a binomial.

$$\begin{aligned}\sum_{t=1}^{T} r_t^2 \leq &\sum_{t=1}^{T} 4\tau_T^2\sigma^2 C_2 \log\left(1+\sigma^{-2}\sigma_{t-1}^2\left(\boldsymbol{x}_t\right)\right)\\ &- 4\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right)\cdot\epsilon_{t-1}(\boldsymbol{x}_t)+\epsilon_{t-1}(\boldsymbol{x}_t)^2\\ &+\frac{4\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right)}{t^2}-\frac{2\epsilon_{t-1}(\boldsymbol{x}_t)}{t^2}+\frac{1}{t^4}\\ \leq &\, 8\tau_T^2\sigma^2C_2\gamma_T-4\mathcal{S}\tau_t\mathcal{E}+\mathcal{E}^2+4\mathcal{S}\tau_t-2\mathcal{E}+\sum\frac{1}{t^4}\\ \leq &\, \tau_T^2C_1\gamma_T-\mathcal{E}\left(4\mathcal{S}\tau_t-\mathcal{E}+2\right)+4\mathcal{S}\tau_t+\frac{\pi^2}{6},\end{aligned}$$

utilizing $t \geq 1$, $C_1 = 8\sigma^2C_2$, and $\sum\frac{1}{t^2} = \frac{\pi^2}{6}$ (Euler's solution to the Basel problem). The assertion now follows by the Cauchy–Schwarz inequality. □

The only thing that is left now is to lift Theorem 2 for the sub-variance GP misspecification to the case of infinite $\mathcal{X}$, too. Theorem 4 does the job.

Theorem 8 (*Regret Bound For Pessimistic Sub-Variance GP Misspec. on Infinite $\mathcal{X}$*). *Let $\mathcal{X} \subset [0,r]^d$ be compact and convex, $d \in \mathbb{N}, r \in \mathbb{R}_{\geq 0}$. Fix $\delta \in (0,1)$, and set*

$$\tau_t^2 = 2\log\left(t^2 2\pi^2/(3\delta)\right)+2d\log\left(t^2 dbr\sqrt{\log(4da/\delta)}\right)$$

with a, b as in condition 1. If Bayesian optimization with misspecified prior mean inducing sub-variance error $\epsilon_T(\boldsymbol{x})$, i.e., $\forall T : \;\epsilon_T(\boldsymbol{x}) \leq \sigma_T(\boldsymbol{x})$ with $\forall t : \epsilon_t(\boldsymbol{x}) < 0$ is run on target function Ψ that satisfies condition 1, we obtain the following cumulative regret bound.

$$\mathbb{P}\left\{R_T \leq \sqrt{(8\tau_T^2+4\tau_T+2)\gamma_T/\log\left(1+\sigma^{-2}\right)}+\frac{\pi^2}{6}\quad \forall T\geq 1\right\}\geq 1-\delta.$$

We reason the cumulative regret stays sublinear for sub-variance GP prior mean parameter misspecification for infinite $\mathcal{X}$. Our main observation from above thus also holds for the practicably more relevant case of infinite $\mathcal{X}$.

Proof. From the proof of Theorem 2 we have the following $\forall T \geq 1$:

$$\begin{aligned}\sum_{t=1}^{T} r_t^2 &\leq \sum_{t=1}^{T}(2\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right)-\epsilon_{t-1}(\boldsymbol{x}_t))^2\\ &\leq \sum_{t=1}^{T}\sigma^2C_2\log\left(1+\sigma^{-2}\sigma_{t-1}^2\left(\boldsymbol{x}_t\right)\right)\left(2\tau_T-1\right)^2\\ &\leq (8\tau_T^2-8\tau_T+2)\sigma^2C_2\gamma_T\end{aligned}$$

with probability $\geq 1-\delta$ where the critical assumption $\forall T : \;\epsilon_T(\boldsymbol{x}) \leq \sigma_T(\boldsymbol{x})$ was used as well as the representation of the information gain $\mathrm{I}\left(\boldsymbol{y}_T, \boldsymbol{\Psi}_T\right) = \frac{1}{2}\sum_{t=1}^{T}\log\left(1+\sigma^{-2}\sigma_{t-1}^2\left(\boldsymbol{x}_t\right)\right)$, and $C_2 = \sigma^{-2}/\log\left(1+\sigma^{-2}\right)$, see the proof of Theorem 1. By Cauchy–Schwarz this implies

$$\sum_{t=1}^{T} 2\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right)-\epsilon_{t-1}(\boldsymbol{x}_t) \leq \sqrt{(8\tau_T^2-8\tau_T+2)\sigma^2C_2\gamma_T}\quad \forall T\geq 1 \tag{B.3}$$

From the proof of Theorem 3 we further have for the regret in the case of infinite $\mathcal{X}$

$$r_t \leq 2\tau_t\sigma_{t-1}\left(\boldsymbol{x}_t\right)-\epsilon_{t-1}(\boldsymbol{x}_t)+\frac{1}{t^2}\quad \forall t\geq 1$$

with probability greater than $1-\delta$. Summing over and using Eq. (A.1) now directly delivers that

$$\sum_{t=1}^{T} r_t \leq \sqrt{(8\tau_T^2-8\tau_T+2)\sigma^2C_2\gamma_T}+\frac{\pi^2}{6}\quad \forall T\geq 1,$$

with probability greater than $1-\delta$, where we again used Euler's solution to the Basel problem: $\sum\frac{1}{t^2} = \frac{\pi^2}{6}$. With $C_2 = \sigma^{-2}/\log\left(1+\sigma^{-2}\right)$ and $R_T = \sum_{t=1}^{T} r_t$ the assertion follows. □

Appendix C. Kernels

Recall Definition 2 from Section 2: A function $f : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ is finitely positive semi-definite if it is symmetric ($\forall \boldsymbol{x}, \boldsymbol{z} \in \mathcal{X} : f(\boldsymbol{x},\boldsymbol{z}) = f(\boldsymbol{z},\boldsymbol{x})$) and the matrix $\boldsymbol{K}$ formed by applying f to any finite subset of $\mathcal{X}$ is positive semi-definite, i.e. for its quadratic form it holds $\boldsymbol{x}'\boldsymbol{K}\boldsymbol{x} \geq 0$ $\forall \boldsymbol{x} \in \mathcal{X}$. A kernel is said to be isotropic if it is a function of the distance $\|\boldsymbol{x}-\boldsymbol{x}'\|$, conditioned on a norm, mostly the L2-Norm. Popular isotropic kernel families are linear kernels

$$k(\boldsymbol{x},\boldsymbol{x}') = \sigma_b^2+\sigma^2(\boldsymbol{x}-c)(\boldsymbol{x}'-c), \tag{C.1}$$

polynomial kernels

$$k(\boldsymbol{x},\boldsymbol{x}') = \left(\sigma_b^2+\sigma^2(\boldsymbol{x}-c)(\boldsymbol{x}'-c)\right)^p, \tag{C.2}$$

Gaussian kernels

$$k(\boldsymbol{x},\boldsymbol{x}') = \sigma^2\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|^2}{2\ell^2}\right), \tag{C.3}$$

exponential kernels

$$k(\boldsymbol{x},\boldsymbol{x}') = \sigma^2\cdot\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|}{2\ell}\right), \tag{C.4}$$

power-exponential kernels[18]

$$k(\boldsymbol{x},\boldsymbol{x}') = \sigma^2\cdot\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|}{2\ell}\right)^p \tag{C.5}$$

and Matérn-kernels

$$\begin{aligned}k(\boldsymbol{x},\boldsymbol{x}') = \sigma^2&\left(1+\frac{\sqrt{\nu}*\|\boldsymbol{x}-\boldsymbol{x}'\|}{\ell}+\frac{\nu}{3}\left(\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|}{\ell}\right)^\rho\right)\\ &\times\exp\left(-\sqrt{\nu}\cdot\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|}{\ell}\right), \nu,\rho\in\mathbb{R},\end{aligned} \tag{C.6}$$

to name only a few. In all kernels, σ^2 is the variance that can be viewed as the average distance away from the mean. In kernels with offset c, the base variance σ_b^2 additionally determines the uncertainty around c. Parameter ℓ determines the smoothness of the GP. For isotropic kernels, there even exists an exact mapping from ℓ to the expected number of up-crossings at level 0 in the unit interval (with $m(\boldsymbol{x}) = 0$, of course). Sometimes the effect of the kernel on the GP is reduced to this smoothness parameter ℓ.[19] However, as any finitely positive semi-definite function is a kernel, it can include various other parameters and represent all possible covariance structures.

Appendix D. Prior conference publication

A prior version of parts of this work had been presented at the *Ninth International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making* (IUKM) in March 2022 and published under the title "Accounting for Gaussian Process Imprecision in Bayesian Optimization" in the conference proceedings as part of the Lecture Notes in Computer Science book series (LNAI, volume 13199), see [1]. In that paper, we proposed the conceptual idea of PROBO without any detailed analysis and theoretical results. The paper at hand goes substantially beyond this proceedings paper. In particular,

- Section 1 (Introduction) is completely new and unrelated to the proceedings paper.
- Section 2 (Background and Related Work) borrows Definition 1 (Gaussian Process) and Algorithm 1 (Bayesian Optimization) from the IUKM paper, but additionally introduces and motivates definitions 3 (Reproducing Kernel Hilbert Space), 4 (Expected Improvement), and 5 (Lower Confidence Bound). Moreover, it includes novel discussions of optimality and convergence of Bayesian optimization.

[18] For $p = 2$: Gaussian kernel.

[19] Also called kernel-bandwidth or length-scale parameter.

Table E.4
Accumulated difference for BO of 50 randomly selected test functions from `smoof`, see Section 5.

Test Function	Mean functional form	Mean parameters	Kernel functional form	Kernel parameters
1-d Ackley Function	23	38	67	23
1-d Alpine01 Function	2.8	1.8	2	1.2
1-d Alpine N. 2 Function	0.11	0.15	0.16	0.079
1-d Chung Reynolds Function	9.1e+03	5.4e+03	9.3e+03	5.4e+03
Cosine Mixture Function	0.073	0.07	0.11	0.14
1-d Deflected Corrugated Spring function	1.4	1.7	2.1	0.77
1-d Double-Sum Function	0.076	0.044	6.9	0.021
1-d Exponential Function	0.00036	0.0015	0.00064	0.00017
1-d Generalized Drop-Wave Function	0.7	1.4	1.7	0.59
1-d Griewank Function	1.1	0.71	1.9	0.69
2-d Hyper-Ellipsoid Function	0.24	2	3.7	0.0012
Six-Hump Camel Back Function	1.3	3.3	2.9	0.71
Price Function N. 4	1.9e+16	1.1e+16	1.1e+16	3.5e+15
Schaffer Function N. 2	0.79	1.3	0.9	0
Beale Function	27	17	20	0.76
Matyas Function	0.25	0.67	3.8	0.0014
Engvall Function	1.7e+11	3.7e+11	2.7e+11	1.2e+11
El-Attar-Vidyasagar-Dutta Function	3e+07	4.6e+07	7.7e+07	4e+07
Cube Function	2.6e+03	6.3e+03	4.8e+03	0
Holder Table Function N. 1	1.1e+02	62	74	1.8
Goldstein-Price Function	8e+02	5.2e+02	6.8e+02	0
3-d Dixon-Price function	1.5e+03	2.2e+03	9.6e+02	1.3e+03
Schaffer Function N. 2	0.32	0.9	0.94	0.078
Giunta Function	0.16	0.29	0.1	0.00018
Chichinadze Function	19	29	66	3.6
Kearfott Function	3.4	7.8	5.7	0
3-d Hartmann Function	3	4.8	5.3	0.82
3-d Alpine N. 2 Function	14	25	30	4.6
Complex Function	3.1	4	1.4	0
Carrom Table Function	71	71	81	0.2
4-d Alpine N. 2 Function	86	33	66	4.7
Adjiman Function	0.34	0.024	1.6	0.00099
Bird Function	1.4e+02	5.1e+02	1.5e+02	0
4-d Generalized Drop-Wave Function	1.5	2.1	1	0.015
Chichinadze Function	54	47	44	16
Brent Function	0.44	0.47	13	4.4e−05
Bukin Function N. 2	3.1	2.2	1.6e+02	0.089
4-d Sum of Different Squares Function	0.37	1.4	0.32	0
Bent-Cigar Function	3.6e+09	2e+10	7e+09	5.7e+08
Booth Function	13	14	47	15
Bartels Conn Function	2.2e+03	1.3e+04	4e+04	27
7-d Sphere Function	1.5e+02	4.4e+02	97	8.5
Goldstein-Price Function	5.8e+02	4e+02	4.4e+02	0
Price Function N. 2	0.36	1.7	0.84	0.21
Engvall Function	1.4e+11	5.1e+11	2.5e+11	4.7e+10
7-d Deflected Corrugated Spring function	16	38	11	0
7-d Hyper-Ellipsoid function	5e+02	1.7e+03	3.4e+02	0
Bent-Cigar Function	4.8e+10	1.5e+11	3.4e+10	0
Trecanni Function	1.5	3.4	7	0
Matyas Function	0.28	0.59	2.7	0

- Section 2.5 on related work is entirely new and unrelated to the proceedings paper.
- Section 3 (Bayesian Sensitivity Analysis) builds on Section 2 in the proceedings paper. It now provides more details on the experiments and more comprehensive results.
- Section 4 (Theoretical Analysis) is completely new and unrelated to the proceedings paper. The herein-derived regret bounds are novel.
- Section 5 (PROBO: Prior-Mean-Robust Bayesian Optimization) borrows from Section 3 in the proceedings paper. It additionally includes derivations of the GLCB as well as further illustrations of the methods (e.g., figure 3).
- Section 6 (Application on Graphene Production) reports on a superset of experiments described in the proceedings paper and entails further illustrations of graphene productions (e.g., figures 4 and 5). In particular, all results in the power setup are novel as well as all results in the time setup except for LCB and EI (see Appendix B).
- Section 7 (Discussion) contains two revised paragraphs (roughly 50 percent) from the proceedings paper and two paragraphs that are entirely novel, containing – amongst other things – a novel outlook to future work.

In summary, we have extended the conference proceedings paper by novel theoretical results, additional experiments, a general introduction, a summary of related work as well as a broader discussion of the results and an outlook to future work. Furthermore, we have thoroughly revised the parts of the earlier paper that are included in the manuscript at hand.

This manuscript further builds on and includes parts of the master thesis “Robust Generalizations of Stochastic Derivative-Free Optimization”, see [84]. As part of the standard examination procedure, this master thesis was made available via a preprint server hosted by

Table E.5
Standardized accumulated differences for BO of 50 randomly selected test functions from `smoof`, see Section 5.

Test function	Mean functional form	Mean parameters	Kernel functional form	Kernel parameters
1-d Ackley Function	0.62	1	1.8	0.61
1-d Alpine01 Function	1.4	0.94	1	0.62
1-d Alpine N. 2 Function	0.87	1.2	1.3	0.62
1-d Chung Reynolds Function	1.3	0.74	1.3	0.73
Cosine Mixture Function	0.75	0.72	1.1	1.4
1-d Deflected Corrugated Spring function	0.96	1.1	1.4	0.51
1-d Double-Sum Function	0.043	0.025	3.9	0.012
1-d Exponential Function	0.54	2.3	0.95	0.25
1-d Generalized Drop-Wave Function	0.65	1.3	1.5	0.54
1-d Griewank Function	1	0.64	1.7	0.62
2-d Hyper-Ellipsoid Function	0.16	1.3	2.5	0.00083
Six-Hump Camel Back Function	0.63	1.6	1.4	0.35
Price Function N. 4	1.7	0.99	0.99	0.32
Schaffer Function N. 2	1.1	1.7	1.2	0
Beale Function	1.7	1.1	1.2	0.047
Matyas Function	0.22	0.57	3.2	0.0012
Engvall Function	0.72	1.6	1.2	0.51
El-Attar-Vidyasagar-Dutta Function	0.62	0.96	1.6	0.82
Cube Function	0.75	1.8	1.4	0
Holder Table Function N. 1	1.8	0.99	1.2	0.028
Goldstein-Price Function	1.6	1	1.4	0
3-d Dixon-Price function	1	1.5	0.65	0.85
Schaffer Function N. 2	0.56	1.6	1.7	0.14
Giunta Function	1.2	2.1	0.73	0.0013
Chichinadze Function	0.64	0.99	2.3	0.12
Kearfott Function	0.8	1.8	1.4	0
3-d Hartmann Function	0.86	1.4	1.5	0.24
3-d Alpine N. 2 Function	0.75	1.4	1.6	0.25
Complex Function	1.5	1.9	0.64	0
Carrom Table Function	1.3	1.3	1.4	0.0036
4-d Alpine N. 2 Function	1.8	0.7	1.4	0.1
Adjiman Function	0.69	0.049	3.3	0.002
Bird Function	0.68	2.6	0.76	0
4-d Generalized Drop-Wave Function	1.3	1.8	0.88	0.013
Chichinadze Function	1.3	1.2	1.1	0.39
Brent Function	0.13	0.14	3.7	1.3e−05
Bukin Function N. 2	0.074	0.053	3.9	0.0021
4-d Sum of Different Squares Function	0.73	2.6	0.63	0
Bent-Cigar Function	0.46	2.6	0.9	0.073
Booth Function	0.59	0.61	2.1	0.67
Bartels Conn Function	0.16	0.93	2.9	0.002
7-d Sphere Function	0.84	2.6	0.56	0.049
Goldstein-Price Function	1.6	1.1	1.2	0
Price Function N. 2	0.47	2.2	1.1	0.27
Engvall Function	0.59	2.1	1.1	0.2
7-d Deflected Corrugated Spring function	1	2.3	0.65	0
7-d Hyper-Ellipsoid function	0.79	2.7	0.53	0
Bent-Cigar Function	0.83	2.6	0.59	0
Trecanni Function	0.51	1.1	2.4	0
Matyas Function	0.31	0.66	3	0

Ludwig-Maximilians-Universität (LMU) Munich, but has not been submitted for publication to a peer-reviewed venue. Prior ideas have also been presented on a poster at the International Symposium on Imprecise Probabilities (ISIPTA) 2021 [134]

Appendix E. Bayesian sensitivity analysis

E.1. Accumulated differences of mean optimization paths for Bayesian optimization of 50 randomly selected test functions

See Table E.4.

E.2. Standardized accumulated differences of mean optimization paths for BO of 50 randomly selected test functions

See Table E.5.

Appendix F. PROBO application on graphene data

See Figs. F.7–F.10.

Appendix G. PROBO application on graphene with random embedding

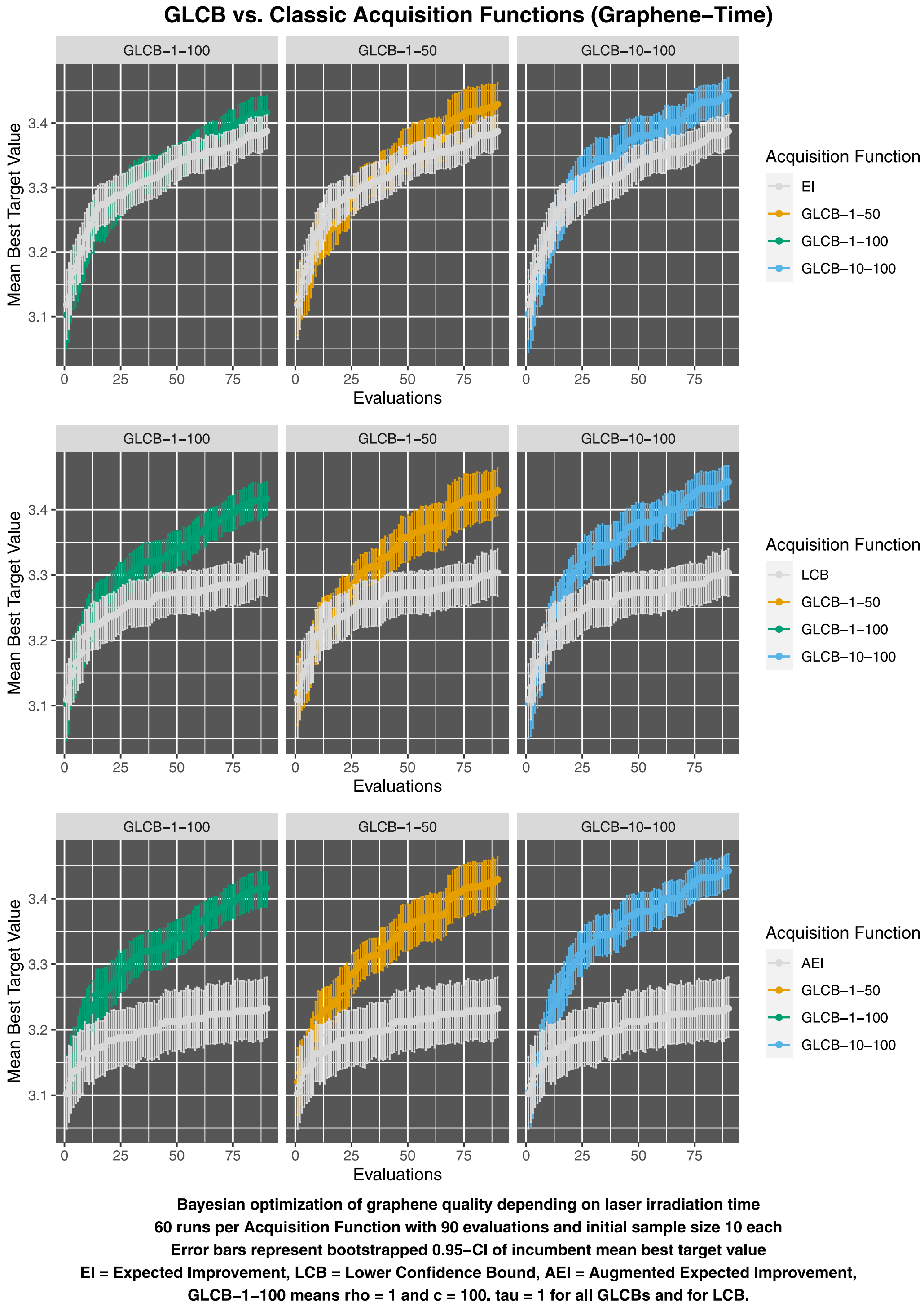
See Fig. G.11.

Appendix H. PROBO application on graphene with PCA-based embedding

See Fig. H.12.

Appendix I. PROBO application on drop-wave functions

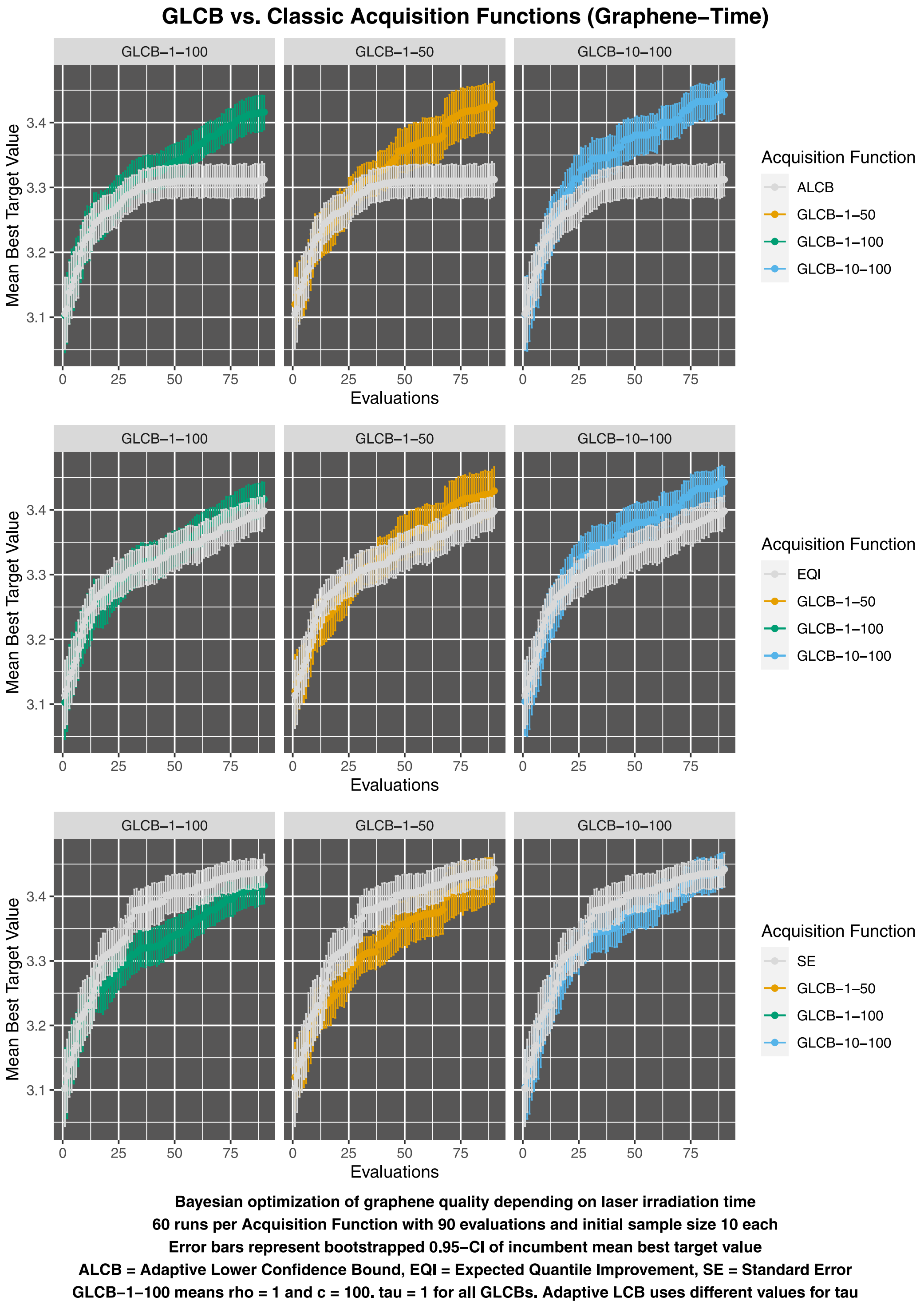
See Figs. I.13 and I.14.



GLCB vs. Classic Acquisition Functions (Graphene–Time)

Bayesian optimization of graphene quality depending on laser irradiation time
60 runs per Acquisition Function with 90 evaluations and initial sample size 10 each
Error bars represent bootstrapped 0.95–CI of incumbent mean best target value
EI = Expected Improvement, LCB = Lower Confidence Bound, AEI = Augmented Expected Improvement,
GLCB–1–100 means rho = 1 and c = 100. tau = 1 for all GLCBs and for LCB.

Fig. F.7. Graphene and time: Benchmarking results from graphene quality as function of laser irradiation time: GLCB vs. several established Acquisition Functions (1).


GLCB vs. Classic Acquisition Functions (Graphene–Time)

Bayesian optimization of graphene quality depending on laser irradiation time
60 runs per Acquisition Function with 90 evaluations and initial sample size 10 each
Error bars represent bootstrapped 0.95–CI of incumbent mean best target value
ALCB = Adaptive Lower Confidence Bound, EQI = Expected Quantile Improvement, SE = Standard Error
GLCB–1–100 means rho = 1 and c = 100. tau = 1 for all GLCBs. Adaptive LCB uses different values for tau

Fig. F.8. Graphene and time: Benchmarking results from graphene quality as function of laser irradiation time: GLCB vs. several established Acquisition Functions (2).

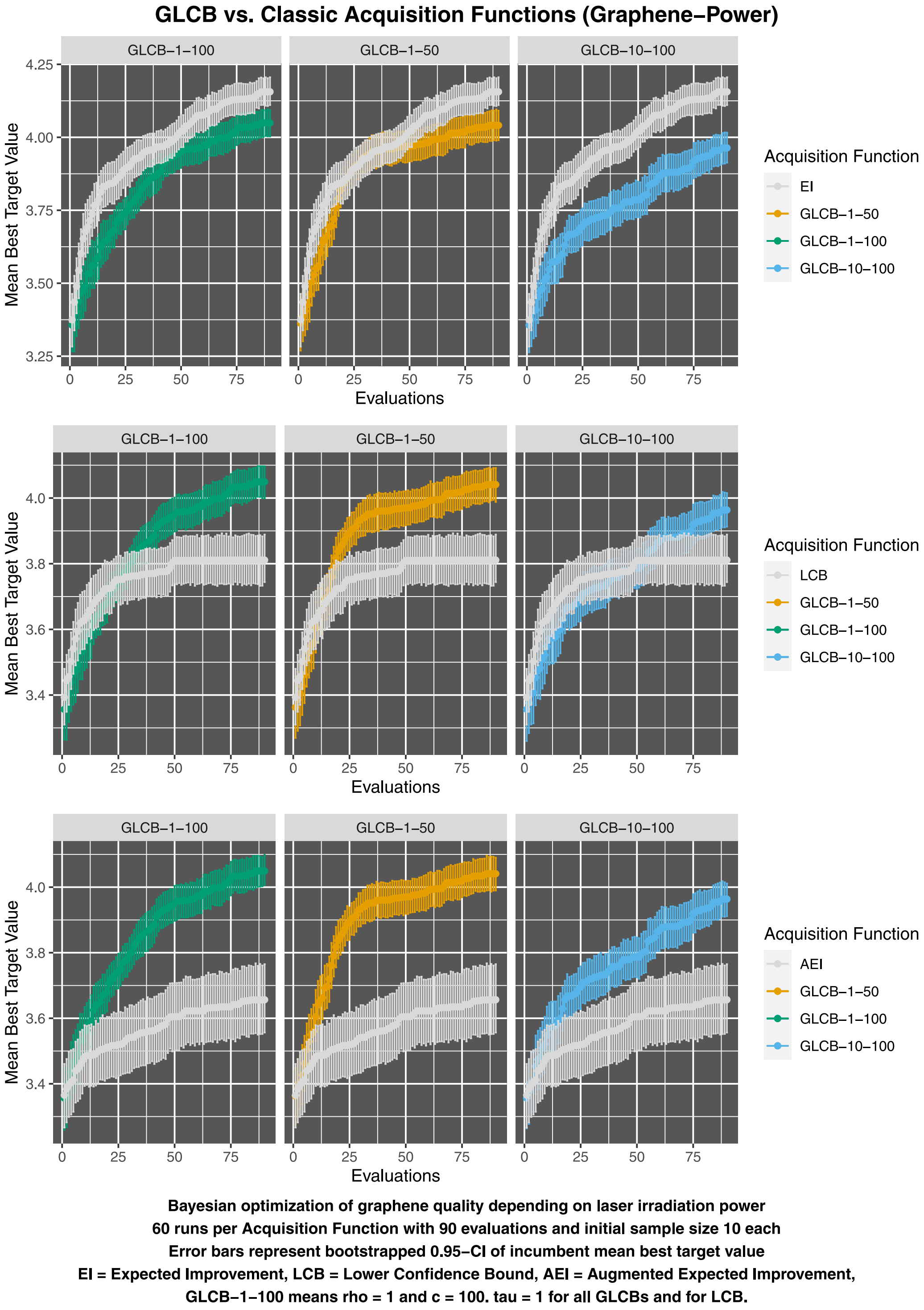
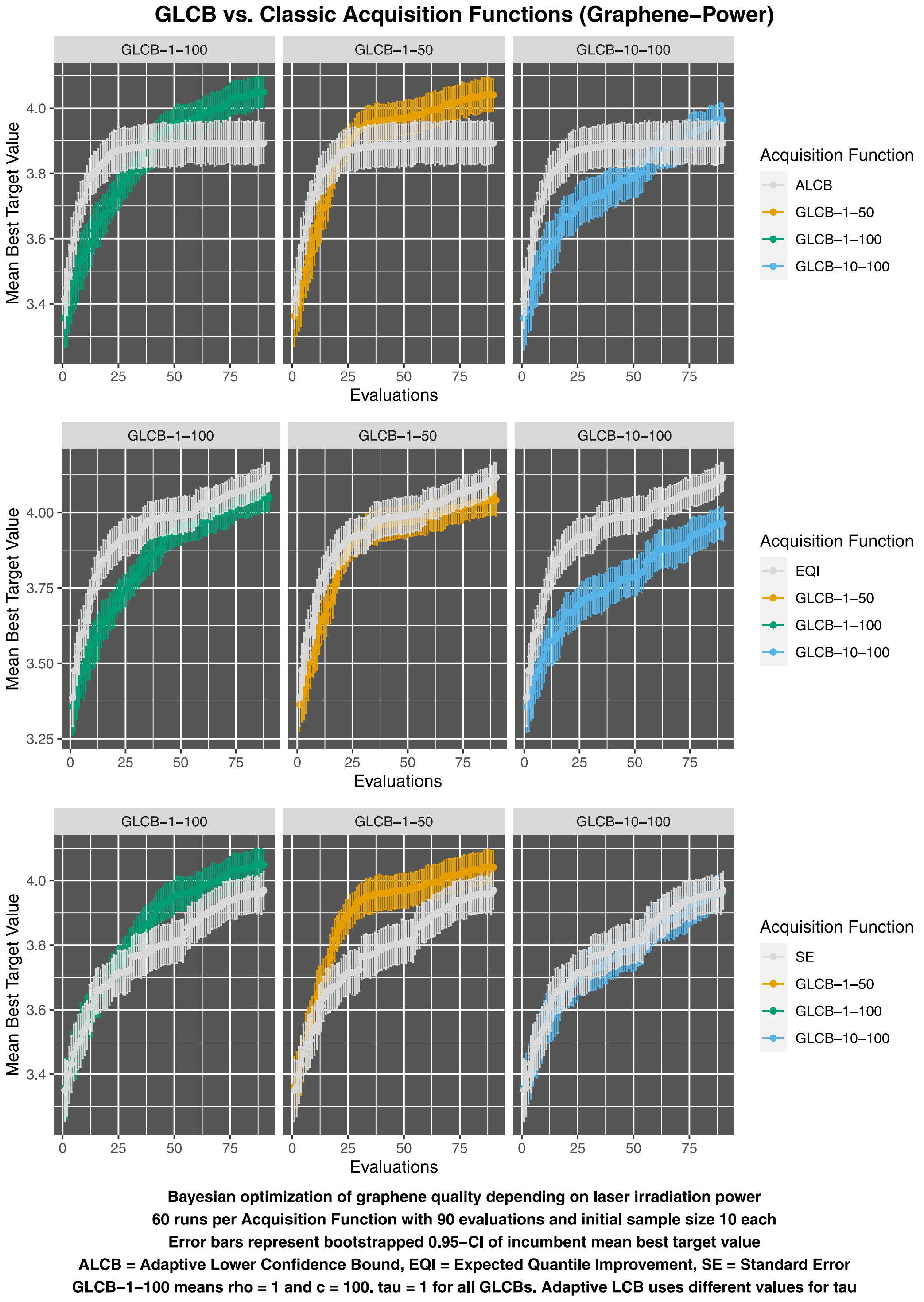


Fig. F.9. Graphene and power: Benchmarking results from graphene quality as function of laser irradiation power: GLCB vs. several established Acquisition Functions (1).



GLCB vs. Classic Acquisition Functions (Graphene–Power)

Bayesian optimization of graphene quality depending on laser irradiation power
60 runs per Acquisition Function with 90 evaluations and initial sample size 10 each
Error bars represent bootstrapped 0.95–CI of incumbent mean best target value
ALCB = Adaptive Lower Confidence Bound, EQI = Expected Quantile Improvement, SE = Standard Error
GLCB–1–100 means rho = 1 and c = 100. tau = 1 for all GLCBs. Adaptive LCB uses different values for tau

Fig. F.10. Graphene and power: Benchmarking results from graphene quality as function of laser irradiation power: GLCB vs. several established Acquisition Functions (2).

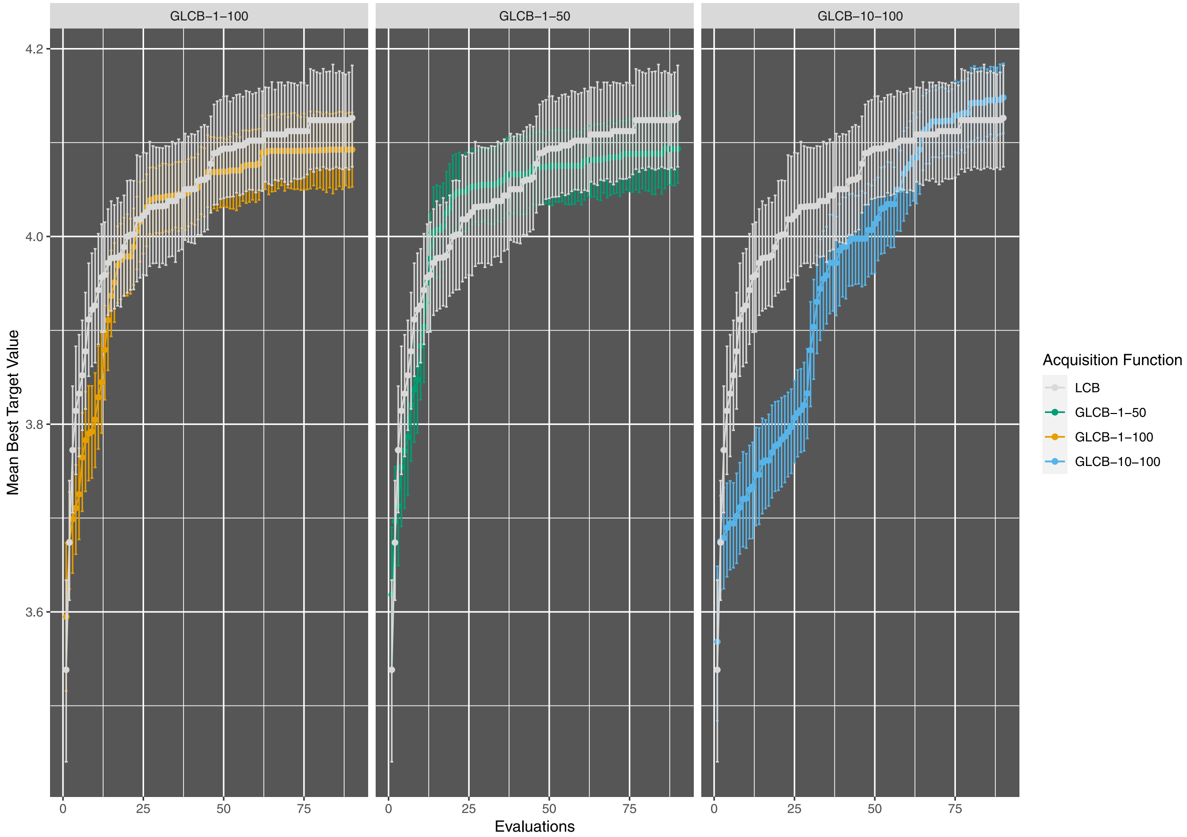


Fig. G.11. Benchmarking results from graphene data with random embedding of covariates power, time, gas (one-hot encoding), and pressure, see also Table 3 and explanations on embeddings in Sections 5 and 6. Generalized lower confidence bound (GLCB) vs. lower confidence bound (LCB). Figure shows 60 runs per Acquisition Function with 90 evaluations and initial sample size 10 each. Error bars represent 95% confidence intervals. GLCB-1-100 means $\rho = 1$ and c = 100; $\tau_t = 1$ for all GLCBs and LCB.

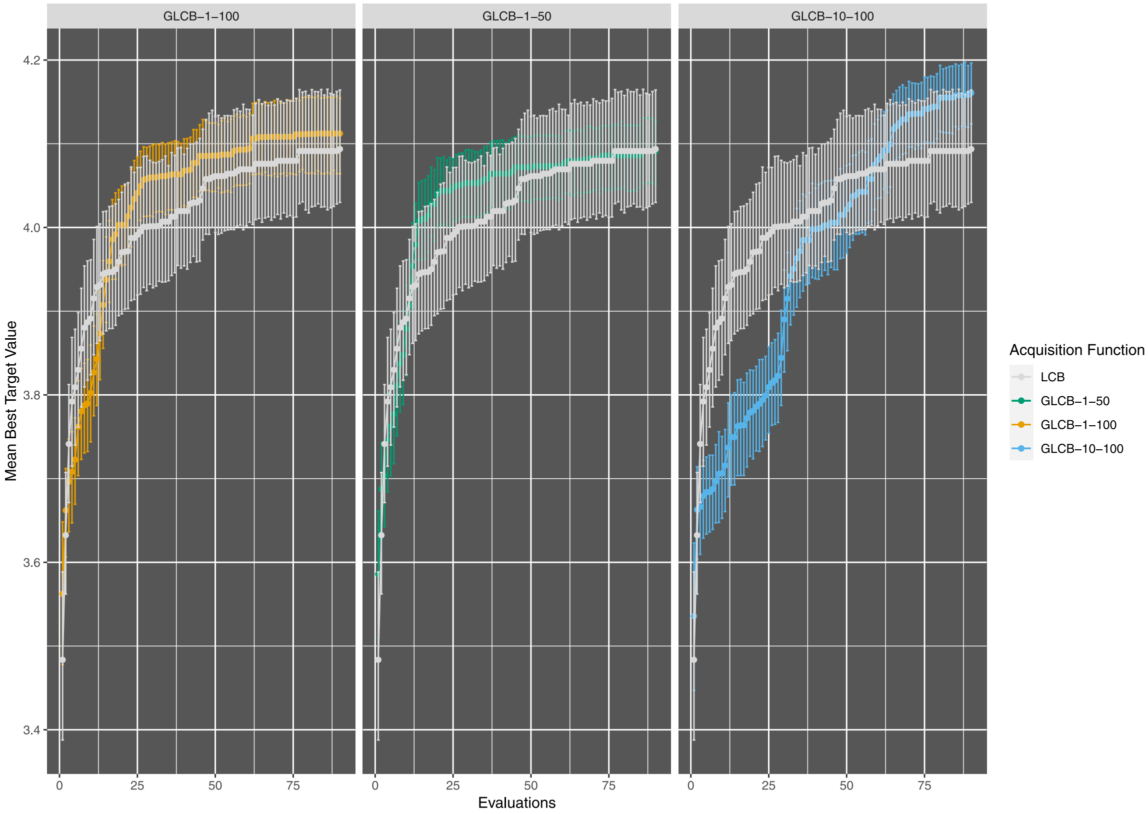
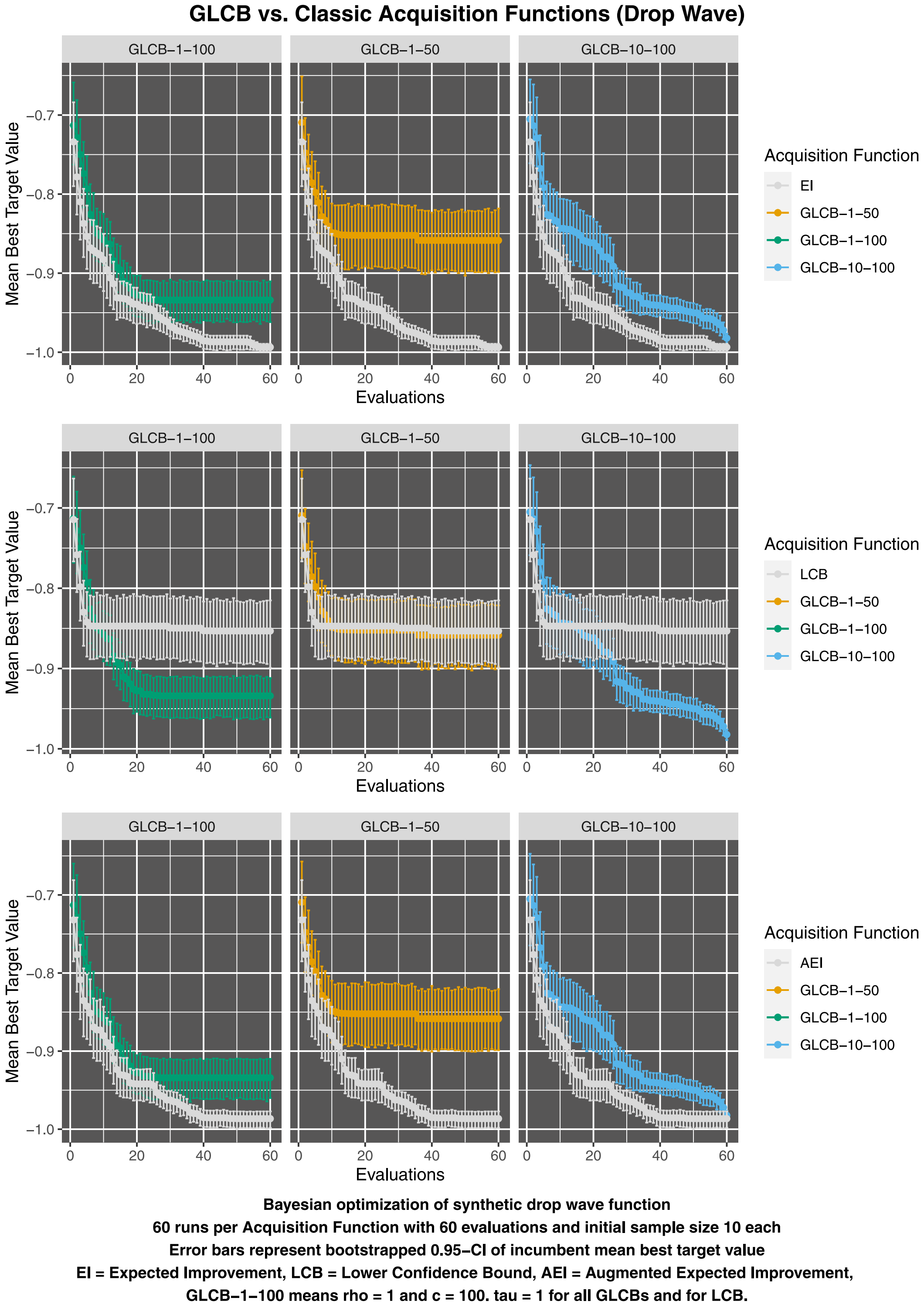


Fig. H.12. Benchmarking results from graphene data with principal component analysis (PCA) based embedding of covariates power, time, gas (one-hot encoding), and pressure, see also Table 3 and explanations on embeddings in Sections 5 and 6. Generalized lower confidence bound (GLCB) vs. lower confidence bound (LCB). Figure shows are 60 runs per Acquisition Function with 90 evaluations and initial sample size 10 each. Error bars represent 95% confidence intervals. GLCB-1-100 means $\rho = 1$ and c = 100; $\tau_t = 1$ for all GLCBs and LCB.


GLCB vs. Classic Acquisition Functions (Drop Wave)

Bayesian optimization of synthetic drop wave function
60 runs per Acquisition Function with 60 evaluations and initial sample size 10 each
Error bars represent bootstrapped 0.95–CI of incumbent mean best target value
EI = Expected Improvement, LCB = Lower Confidence Bound, AEI = Augmented Expected Improvement,
GLCB–1–100 means rho = 1 and c = 100. tau = 1 for all GLCBs and for LCB.

Fig. I.13. Benchmarking results from synthetic drop wave function: Generalized Lower Confidence Bound (GLCB) vs. established acquisition functions.

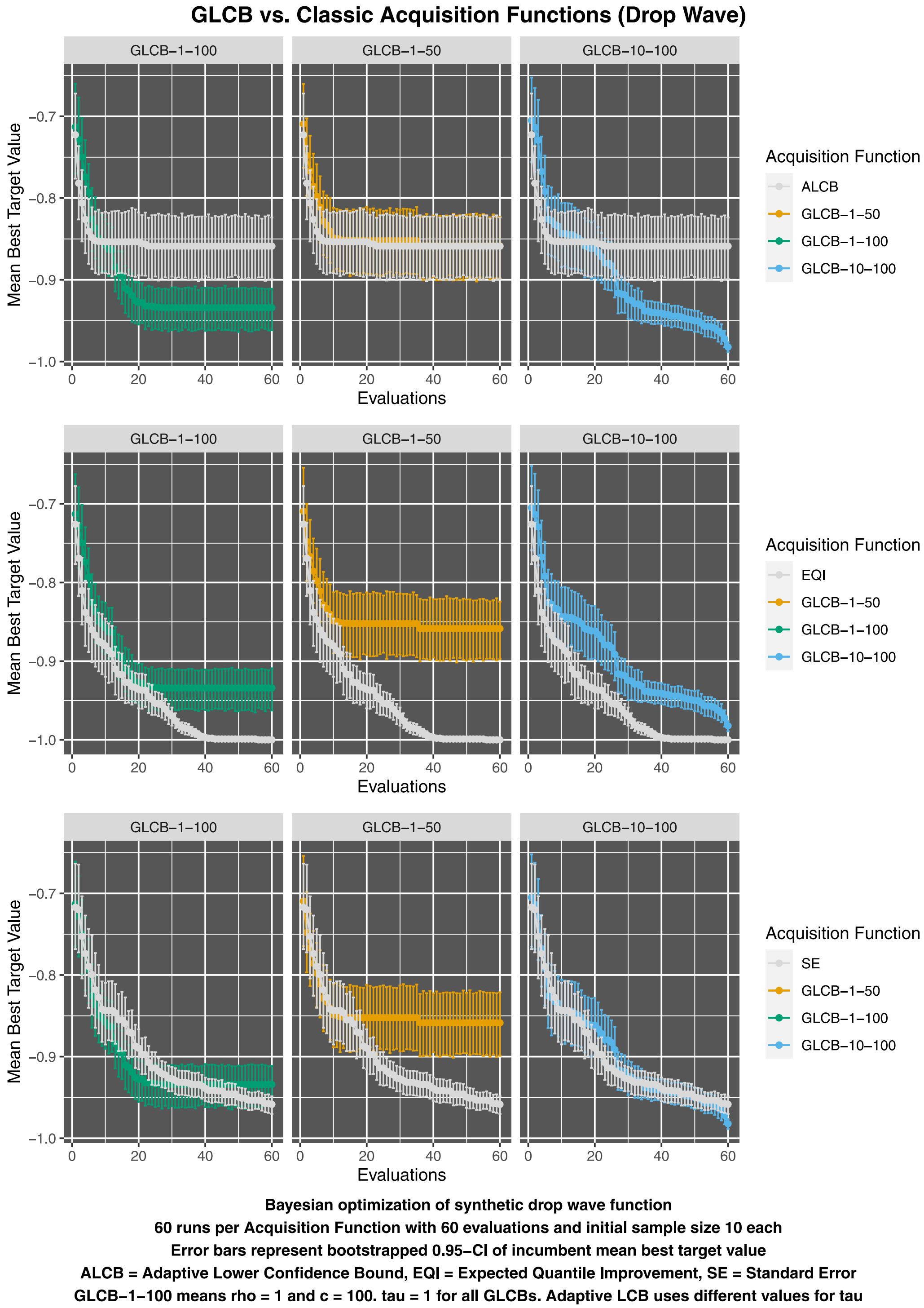


Fig. I.14. Benchmarking results from synthetic drop wave function: Generalized Lower Confidence Bound (GLCB) vs. established acquisition functions.

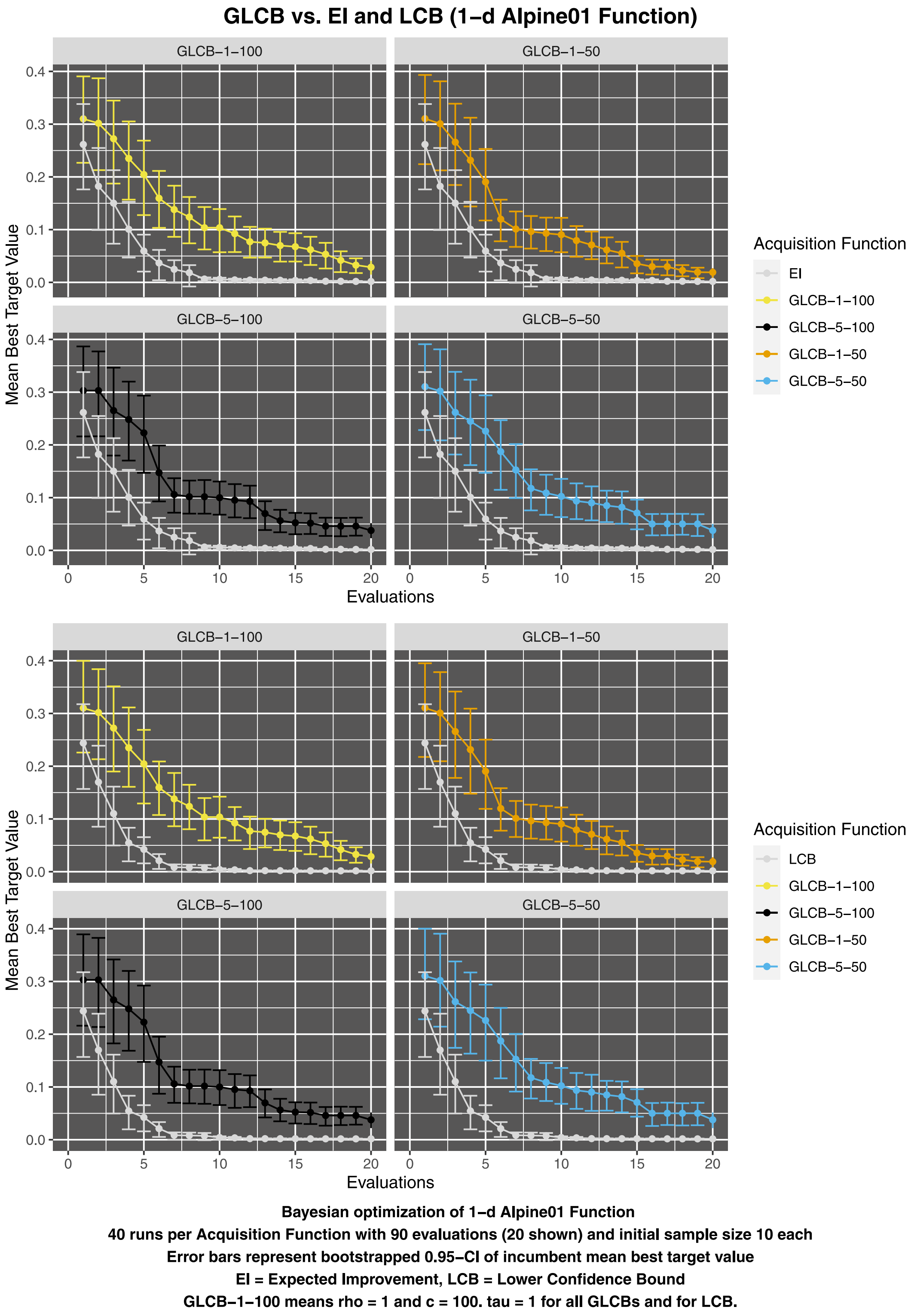


Fig. J.15. Benchmarking results from synthetic Alpine function: Generalized Lower Confidence Bound (GLCB) vs. EI and LCB.

Appendix J. PROBO application on alpine functions

See Fig. J.15.

References

[1] J. Rodemann, T. Augustin, Accounting for Gaussian process imprecision in Bayesian optimization, in: International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making, Springer, 2022, pp. 92–104.

[2] J. Zittrain, The hidden costs of automated thinking, The New Yorker (2019) URL https://www.newyorker.com/tech/annals-of-technology/the-hidden-costs-of-automated-thinking, July 23, 2019.

[3] J. Močkus, On Bayesian methods for seeking the extremum, in: Optimization Techniques IFIP Technical Conference, Springer, 1975, pp. 400–404.

[4] G. Malkomes, R. Garnett, Automating Bayesian optimization with Bayesian optimization, Adv. Neural Inf. Process. Syst. 31 (2018).

[5] D. Bertsimas, J. Tsitsiklis, Simulated annealing, Statist. Sci. 8 (1) (1993) 10–15.

[6] N. Hansen, S.D. Müller, P. Koumoutsakos, Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES), Evol. Comput. 11 (1) (2003) 1–18.

[7] C.F. Manski, Partial Identification of Probability Distributions, Springer Science & Business Media, 2003.

[8] G. De Ath, R.M. Everson, A.A. Rahat, J.E. Fieldsend, Greed is good: Exploration and exploitation trade-offs in Bayesian optimisation, ACM Trans. Evol. Learn. Optim. 1 (1) (2021) 1–22.

[9] E. Hüllermeier, W. Waegeman, Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods, Mach. Learn. 110 (3) (2021) 457–506.

[10] X.-M. Dong, Y.-H. Gu, J. Shi, K. Xiang, Random multi-scale kernel-based Bayesian distribution regression learning, Knowl.-Based Syst. 201–202 (2020) 106073.

[11] C. Hvarfner, E. Hellsten, F. Hutter, L. Nardi, Self-correcting Bayesian optimization through Bayesian active learning, Adv. Neural Inf. Process. Syst. 36 (2024).

[12] Q. Lu, K.D. Polyzos, B. Li, G.B. Giannakis, Surrogate modeling for Bayesian optimization beyond a single Gaussian process, IEEE Trans. Pattern Anal. Mach. Intell. 45 (9) (2023) 11283–11296.

[13] F. Mangili, A prior near-ignorance Gaussian process model for nonparametric regression, in: ISIPTA '15: 9th International Symposium on Imprecise Probability: Theories and Applications, 2015, pp. 187–196.

[14] F. Mangili, A prior near-ignorance Gaussian process model for nonparametric regression, Internat. J. Approx. Reason. 78 (2016) 153–171.

[15] P. Frazier, J. Wang, Bayesian optimization for materials design, in: Information Science for Materials Discovery and Design, Springer, 2016, pp. 45–75.

[16] E.O. Pyzer-Knapp, Bayesian optimization for accelerated drug discovery, IBM J. Res. Dev. 62 (6) (2018) 2:1–2:7.

[17] M.A. Awal, M. Masud, M.S. Hossain, A.A. Bulbul, S.M.H. Mahmud, A.K. Bairagi, A novel Bayesian optimization-based machine learning framework for COVID-19 detection from inpatient facility data, IEEE Access 9 (2021) 10263–10281.

[18] D.P. Kuttichira, S. Gupta, D. Nguyen, S. Rana, S. Venkatesh, Verification of integrity of deployed deep learning models using Bayesian optimization, Knowl.-Based Syst. 241 (2022) 108238.

[19] V. Nguyen, Bayesian optimization for accelerating hyper-parameter tuning, in: 2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering, AIKE, IEEE, 2019, pp. 302–305.

[20] J. Snoek, H. Larochelle, R.P. Adams, Practical Bayesian optimization of machine learning algorithms, Adv. Neural Inf. Process. Syst. 25 (2012).

[21] B. Bischl, J. Richter, J. Bossek, D. Horn, J. Thomas, M. Lang, mlrMBO: A modular framework for model-based optimization of expensive black-box functions, 2017, arXiv preprint arXiv:1703.03373.

[22] C.E. Rasmussen, Gaussian processes in machine learning, in: Summer School on Machine Learning, Springer, 2003, pp. 63–71.

[23] A.K.A. Venkatesh, S. Rana, A. Shilton, S. Venkatesh, Human-AI collaborative Bayesian optimisation, Adv. Neural Inf. Process. Syst. 35 (2022).

[24] S.R. Chowdhury, A. Gopalan, On kernelized multi-armed bandits, in: International Conference on Machine Learning, PMLR, 2017, pp. 844–853.

[25] B. Bischl, S. Wessing, N. Bauer, K. Friedrichs, C. Weihs, MOI-MBO: multiobjective infill for parallel model-based optimization, in: International Conference on Learning and Intelligent Optimization, Springer, 2014, pp. 173–186.

[26] F. Hutter, L. Kotthoff, J. Vanschoren, Automated Machine Learning: Methods, Systems, Challenges, Springer, 2018.

[27] D.D. Cox, S. John, A statistical method for global optimization, in: Proceedings of 1992 IEEE International Conference on Systems, Man, and Cybernetics, IEEE, 1992, pp. 1241–1246.

[28] A. Benavoli, D. Azzimonti, D. Piga, Preferential Bayesian optimisation with skew Gaussian processes, in: Genetic and Evolutionary Computation Conference Companion, 2021, pp. 1842–1850.

[29] B. Liu, Q. Zhang, F.V. Fernández, G. Gielen, Self-adaptive lower confidence bound: A new general and effective prescreening method for Gaussian process surrogate model assisted evolutionary algorithms, in: 2012 IEEE Congress on Evolutionary Computation, IEEE, 2012, pp. 1–6.

[30] A. Ben-Tal, L. El Ghaoui, A. Nemirovski, Robust Optimization, vol. 28, Princeton University Press, 2009.

[31] Y. Wang, B. Chaib-draa, KNN-based Kalman filter: An efficient and non-stationary method for Gaussian process regression, Knowl.-Based Syst. 114 (2016) 148–155.

[32] S. Sun, J. Wang, Multiview learning with variational mixtures of Gaussian processes, Knowl.-Based Syst. 200 (2020) 105990.

[33] M. Papež, A. Quinn, Transferring model structure in Bayesian transfer learning for Gaussian process regression, Knowl.-Based Syst. 251 (2022) 108875.

[34] A. Makarova, H. Shen, V. Perrone, A. Klein, J.B. Faddoul, A. Krause, M. Seeger, C. Archambeau, Overfitting in Bayesian optimization: an empirical study and early-stopping solution, in: 2nd Workshop on Neural Architecture Search (NAS 2021 Collocated with the 9th ICLR 2021), 2021.

[35] R. Moriconi, K.S. Kumar, M.P. Deisenroth, High-dimensional Bayesian optimization with projections using quantile Gaussian processes, Optim. Lett. 14 (1) (2020) 51–64.

[36] A. Shah, A. Wilson, Z. Ghahramani, Student-t Processes as Alternatives to Gaussian Processes, in: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, in: Proceedings of Machine Learning Research, Vol. 33, PMLR, Reykjavik, Iceland, 2014, pp. 877–885.

[37] J. Kirschner, I. Bogunovic, S. Jegelka, A. Krause, Distributionally robust Bayesian optimization, in: Twenty Third International Conference on Artificial Intelligence and Statistics, Vol. 108, PMLR, 2020, pp. 2174–2184.

[38] T. Nguyen, S. Gupta, H. Ha, S. Rana, S. Venkatesh, Distributionally robust Bayesian quadrature optimization, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2020, pp. 1921–1931.

[39] J. Garcia-Barcos, R. Martinez-Cantin, Parallel robust Bayesian optimization with off-policy evaluations, Technical report, 2019, URL https://jgbarcos.github.io/.

[40] S. Daulton, S. Cakmak, M. Balandat, M.A. Osborne, E. Zhou, E. Bakshy, Robust multi-objective Bayesian optimization under input noise, in: 39th International Conference on Machine Learning, PMLR, 2022, pp. 4831–4866.

[41] J.A. Paulson, G. Makrygiorgos, A. Mesbah, Adversarially robust Bayesian optimization for efficient auto-tuning of generic control structures under uncertainty, AIChE J. 68 (6) (2022) e17591.

[42] J.T. Springenberg, A. Klein, S. Falkner, F. Hutter, Bayesian optimization with robust Bayesian neural networks, Adv. Neural Inf. Process. Syst. 29 (2016).

[43] J.B. Mockus, L.J. Mockus, Bayesian approach to global optimization and application to multiobjective and constrained problems, J. Optim. Theory Appl. 70 (1991) 157–172.

[44] N. Khan, D.E. Goldberg, M. Pelikan, Multi-objective Bayesian optimization algorithm, in: Proceedings of the 4th Annual Conference on Genetic and Evolutionary Computation, 2002, pp. 684–684.

[45] M.T. Emmerich, K. Giannakoglou, B. Naujoks, Single- and multiobjective evolutionary optimization assisted by Gaussian random field metamodels, in: IEEE International Conference on Evolutionary Computation, IEEE, 2006, pp. 508–515.

[46] A.J. Keane, Statistical improvement criteria for use in multiobjective design optimization, AIAA J. 44 (4) (2006) 879–891.

[47] J. Knowles, ParEGO: A hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems, IEEE Trans. Evol. Comput. 10 (1) (2006) 50–66.

[48] P. Hennig, C.J. Schuler, Entropy search for information-efficient global optimization, J. Mach. Learn. Res. 13 (Jun) (2012) 1809–1837.

[49] A. Shah, Z. Ghahramani, Pareto frontier learning with expensive correlated objectives, in: International Conference on Machine Learning, PMLR, 2016, pp. 1919–1927.

[50] S. Daulton, M. Balandat, E. Bakshy, Differentiable expected hypervolume improvement for parallel multi-objective Bayesian optimization, Adv. Neural Inf. Process. Syst. 33 (2020).

[51] J.M. Hernández-Lobato, M.W. Hoffman, Z. Ghahramani, Predictive entropy search for multi-objective Bayesian optimization, in: International Conference on Machine Learning, PMLR, 2016, pp. 1492–1501.

[52] M. Abdolshah, A. Shilton, S. Rana, S. Gupta, S. Venkatesh, Multi-objective Bayesian optimisation with preferences over objectives, Adv. Neural Inf. Process. Syst. 32 (2019).

[53] C. Vişan, O. Pascu, M. Stănescu, E.-D. Şandru, C. Diaconu, A. Buzo, G. Pelz, H. Cucu, Automated circuit sizing with multi-objective optimization based on differential evolution and Bayesian inference, Knowl.-Based Syst. 258 (2022) 109987.

[54] M. Kaedi, N. Ghasem-Aghaee, Biasing Bayesian optimization algorithm using case based reasoning, Knowl.-Based Syst. 24 (8) (2011) 1245–1253.

[55] A. Ramachandran, S. Gupta, S. Rana, C. Li, S. Venkatesh, Incorporating expert prior in Bayesian optimisation via space warping, Knowl.-Based Syst. 195 (2020) 105663.

[56] J. Foldager, M. Jordahn, L.K. Hansen, M.R. Andersen, On the role of model uncertainties in Bayesian optimisation, in: Thirty-Ninth Conference on Uncertainty in Artificial Intelligence, Vol. 216, PMLR, 2023, pp. 592–601.

[57] A.M. Schmidt, M.d.F.d.G. Conceição, G.A. Moreira, Investigating the sensitivity of Gaussian processes to the choice of their correlation function and prior specifications, J. Stat. Comput. Simul. 78 (8) (2008) 681–699.

[58] A. Klein, S. Falkner, N. Mansur, F. Hutter, Robo: A flexible and robust Bayesian optimization framework in python, in: NIPS 2017 Bayesian Optimization Workshop, 2017.

[59] G. Malkomes, C. Schaff, R. Garnett, Bayesian optimization for automated model selection, in: Workshop on Automatic Machine Learning, PMLR, 2016, pp. 41–47.

[60] D. Duvenaud, J. Lloyd, R. Grosse, J. Tenenbaum, G. Zoubin, Structure discovery in nonparametric regression through compositional kernel search, in: S. Dasgupta, D. McAllester (Eds.), 30th International Conference on Machine Learning, Vol. 28, (3) PMLR, 2013, pp. 1166–1174.

[61] D. Duvenaud, Automatic Model Construction with Gaussian Processes (Ph.D. thesis), University of Cambridge, 2014.

[62] M. Wistuba, J. Grabocka, Few-shot Bayesian optimization with deep kernel surrogates, 2021, arXiv preprint arXiv:2101.07667.

[63] I. Roman, R. Santana, A. Mendiburu, J.A. Lozano, An experimental study in adaptive kernel selection for Bayesian optimization, IEEE Access 7 (2019) 184294–184302.

[64] B. Lei, T.Q. Kirk, A. Bhattacharya, D. Pati, X. Qian, R. Arroyave, B.K. Mallick, Bayesian optimization with adaptive surrogate models for automated experimental design, NPJ Comput. Mater. 7 (1) (2021) 194.

[65] D. Salinas, J. Golebiowski, A. Klein, M. Seeger, C. Archambeau, Optimizing hyperparameters with conformal quantile regression, in: International Conference on Machine Learning, PMLR, 2023, pp. 29876–29893.

[66] S. Stanton, W. Maddox, A.G. Wilson, Bayesian optimization with conformal prediction sets, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2023, pp. 959–986.

[67] Y. Zhang, S. Park, O. Simeone, Bayesian optimization with formal safety guarantees via online conformal prediction, 2023, arXiv preprint arXiv:2306.17815.

[68] C. Johnstone, B. Cox, Conformal uncertainty sets for robust optimization, in: Conformal and Probabilistic Prediction and Applications, 2021.

[69] I. Bogunovic, A. Krause, Misspecified Gaussian process bandit optimization, Adv. Neural Inf. Process. Syst. 34 (2021).

[70] A. Makarova, I. Usmanova, I. Bogunovic, A. Krause, Risk-averse heteroscedastic Bayesian optimization, Adv. Neural Inf. Process. Syst. 34 (2021).

[71] Z. Wang, B. Kim, L.P. Kaelbling, Regret bounds for meta Bayesian optimization with an unknown Gaussian process prior, Adv. Neural Inf. Process. Syst. 31 (2018).

[72] N. Srinivas, A. Krause, S. Kakade, M. Seeger, Gaussian process optimization in the bandit setting: No regret and experimental design, in: 27th International Conference on Machine Learning, 2010, pp. 1015–1022.

[73] F. Berkenkamp, A.P. Schoellig, A. Krause, No-regret Bayesian optimization with unknown hyperparameters, J. Mach. Learn. Res. 20 (50) (2019) 1–24.

[74] Z. Fan, X. Han, Z. Wang, Transfer learning for Bayesian optimization on heterogeneous search spaces, 2023, arXiv preprint arXiv:2309.16597.

[75] Z. Fan, X. Han, Z. Wang, HyperBO+: Pre-training a universal prior for Bayesian optimization with hierarchical Gaussian processes, 2022, arXiv preprint arXiv:2212.10538.

[76] T.T. Joy, S. Rana, S. Gupta, S. Venkatesh, A flexible transfer learning framework for Bayesian optimization with convergence guarantee, Expert Syst. Appl. 115 (2019) 656–672.

[77] P. Tighineanu, K. Skubch, P. Baireuther, A. Reiss, F. Berkenkamp, J. Vinogradska, Transfer learning with Gaussian processes for Bayesian optimization, in: G. Camps-Valls, F.J.R. Ruiz, I. Valera (Eds.), Proceedings of the 25th International Conference on Artificial Intelligence and Statistics, in: Proceedings of Machine Learning Research, Vol. 151, PMLR, 2022, pp. 6152–6181.

[78] T. Bai, Y. Li, Y. Shen, X. Zhang, W. Zhang, B. Cui, Transfer learning for Bayesian optimization: A survey, 2023, arXiv preprint arXiv:2302.05927.

[79] D.R. Jones, M. Schonlau, W.J. Welch, Efficient global optimization of expensive black-box functions, J. Global Optim. 13 (4) (1998) 455–492.

[80] C. Li, D. Rubín de Celis Leal, S. Rana, S. Gupta, A. Sutti, S. Greenhill, T. Slezak, M. Height, S. Venkatesh, Rapid Bayesian optimisation for synthesis of short polymer fiber materials, Sci. Rep. 7 (1) (2017) 5683.

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

[82] K. Deane, Y. Yang, J.J. Licavoli, V. Nguyen, S. Rana, S. Gupta, S. Venkatesh, P.G. Sanders, Utilization of Bayesian optimization and KWN modeling for increased efficiency of Al-Sc precipitation strengthening, Metals 12 (6) (2022) 975.

[83] Q. Liang, A.E. Gongora, Z. Ren, A. Tiihonen, Z. Liu, S. Sun, J.R. Deneault, D. Bash, F. Mekki-Berrada, S.A. Khan, et al., Benchmarking the performance of Bayesian optimization across multiple experimental materials science domains, NPJ Comput. Mater. 7 (1) (2021) 188.

[84] J. Rodemann, Robust generalizations of stochastic derivative-free optimization, (Master's thesis), LMU Munich, 2021.

[85] J. Bossek, Smoof: Single- and multi-objective optimization test functions, R J. 9 (1) (2017) 103.

[86] N. Hansen, A. Auger, R. Ros, O. Mersmann, T. Tušar, D. Brockhoff, COCO: A platform for comparing continuous optimizers in a black-box setting, Optim. Methods Softw. 36 (1) (2021) 114–144.

[87] S. Vakili, K. Khezeli, V. Picheny, On information gain and regret bounds in Gaussian process bandits, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 82–90.

[88] J. Whitehouse, A. Ramdas, S. Wu, On the sublinear regret of GP-UCB, in: Thirty-seventh Conference on Neural Information Processing Systems, Vol. 37, 2023.

[89] R. Garnett, Bayesian Optimization, Cambridge University Press, 2023.

[90] V. Dani, T.P. Hayes, S.M. Kakade, Stochastic linear optimization under bandit feedback, in: Conference on Learning Theory, COLT, Vol. 2, 2008, p. 3.

[91] S. Curi, F. Berkenkamp, A. Krause, Efficient model-based reinforcement learning through optimistic policy search and planning, Adv. Neural Inf. Process. Syst. 33 (2020).

[92] S. Sussex, A. Makarova, A. Krause, Model-based causal Bayesian optimization, in: International Conference on Learning Representations, 2023.

[93] J.A. Thomas, Elements of information theory, 1991.

[94] S. Ghosal, A. Roy, Posterior consistency of Gaussian process prior for nonparametric binary regression, Ann. Statist. 34 (5) (2006) 2413–2429.

[95] D. Hilbert, Mathematical problems, Bull. Amer. Math. Soc. 9 (3) (1902) 437–479.

[96] R.E. Kass, L. Wasserman, The selection of prior distributions by formal rules, J. Amer. Statist. Assoc. 91 (435) (1996) 1343–1370.

[97] J.O. Berger, L.R. Pericchi, The intrinsic Bayes factor for model selection and prediction, J. Amer. Statist. Assoc. 91 (433) (1996) 109–122.

[98] A. Benavoli, M. Zaffalon, Prior near ignorance for inferences in the k-parameter exponential family, Statistics 49 (5) (2015) 1104–1140.

[99] X.-L. Meng, A BFFer's exploration with nuisance constructs: Bayesian p-value, H-likelihood, and Cauchyanity, in: J. Berger, X.-L. Meng, N. Reid, M.-g. Xie (Eds.), Handbook of Bayesian, Fiducial, and Frequentist Inference, Chapman and Hall/CRC, 2024, pp. 161–187.

[100] D. Rios Insua, F. Ruggeri, Robust Bayesian Analysis, Springer, New York, 2000.

[101] T. Augustin, G. Walter, F.P. Coolen, Statistical inference, in: T. Augustin, F. Coolen, G. de Cooman, M. Troffaes (Eds.), Introduction to Imprecise Probabilities, Wiley Online Library, 2014, pp. 135–189.

[102] J. Rodemann, C. Jansen, G. Schollmeyer, T. Augustin, In all likelihoods: Robust selection of pseudo-labeled data, in: International Symposium on Imprecise Probability: Theories and Applications, PMLR, 2023, pp. 412–425.

[103] S. Dietrich, J. Rodemann, C. Jansen, Semi-supervised learning guided by the generalized Bayes rule under soft revision, 2024, arXiv preprint arXiv:2405.15294.

[104] J. Rodemann, Pseudo label selection is a decision problem, Proc. of the 46th German Conference on Artificial Intelligence, Springer, 2023.

[105] M. Caprio, S. Dutta, K.J. Jang, V. Lin, R. Ivanov, O. Sokolsky, I. Lee, Imprecise Bayesian neural networks, 2023, arXiv preprint arXiv:2302.09656.

[106] A. Marquardt, J. Rodemann, T. Augustin, An empirical study of prior-data conflicts in Bayesian neural networks, in: Poster presented at ISIPTA '23: International Symposium on Imprecise Probability: Theories and Applications, 2023.

[107] P. Walley, Inferences from multinomial data: learning about a bag of marbles, J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1) (1996) 3–34.

[108] S. Moral-Garcia, J.G. Castellano, C.J. Mantas, J. Abellan, Using extreme prior probabilities on the naive credal classifier, Knowl.-Based Syst. 237 (2022) 107707.

[109] A. Nayebi, A. Munteanu, M. Poloczek, A framework for Bayesian optimization in embedded subspaces, in: International Conference on Machine Learning, PMLR, 2019, pp. 4752–4761.

[110] F. Llorente, P.M. Djurić, Dynamic random feature Gaussian processes for Bayesian optimization of time-varying functions, in: ICASSP 2024 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2024, pp. 9756–9760, http://dx.doi.org/10.1109/ICASSP48485.2024.10447767.

[111] D. Horn, T. Wagner, D. Biermann, C. Weihs, B. Bischl, Model-based multi-objective optimization: taxonomy, multi-point proposal, toolbox and benchmark, in: International Conference on Evolutionary Multi-Criterion Optimization, Springer, 2015, pp. 64–78.

[112] L. Kotthoff, V. Jain, A. Tyrrell, H. Wahab, P. Johnson, AI for materials science: Tuning laser-induced graphene production, in: Presentation at NASA AME, 2019, 2019.

[113] M.D. McKay, R.J. Beckman, W.J. Conover, A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, Technometrics 42 (1) (2000) 55–61.

[114] R. Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2020.

[115] N. Hansen, A. Auger, R. Ros, S. Finck, P. Pošík, Comparing results of 31 algorithms from the black-box optimization benchmarking BBOB-2009, in: Proceedings of the 12th Annual Conference Companion on Genetic and Evolutionary Computation, 2010.

[116] O. Mersmann, M. Preuss, H. Trautmann, B. Bischl, C. Weihs, Analyzing the BBOB results by means of benchmarking concepts, Evol. Comput. 23 (1) (2015) 161–185.

[117] J. Rodemann, H. Blocher, Partial rankings of optimizers, in: International Conference on Learning Representations (ICLR), Tiny Papers Track, 2024.

[118] J. Rodemann, F. Croppi, P. Arens, Y. Sale, J. Herbinger, B. Bischl, E. Hüllermeier, T. Augustin, C.J. Walsh, G. Casalicchio, Explaining Bayesian optimization by Shapley values facilitates human-ai collaboration, 2024, arXiv preprint arXiv:2403.04629.

[119] L.V. Utkin, Y.A. Zhuk, Robust boosting classification models with local sets of probability distributions, Knowl.-Based Syst. 61 (2014) 59–75.

[120] L.V. Utkin, M.S. Kovalev, A.A. Meldo, A deep forest classifier with weights of class probability distribution subsets, Knowl.-Based Syst. 173 (2019) 15–27.

[121] L.V. Utkin, M. Kovalev, A. Meldo, F. Coolen, Imprecise extensions of random forests and random survival forests, in: International Symposium on Imprecise Probabilities: Theories and Applications, PMLR, 2019, pp. 404–413.

[122] L.V. Utkin, M.A. Ryabinin, A Siamese deep forest, Knowl.-Based Syst. 139 (2018) 13–22.

[123] J. Abellan, C.J. Mantas, J.G. Castellano, A random forest approach using imprecise probabilities, Knowl.-Based Syst. 134 (2017) 72–84.

[124] L.V. Utkin, Y.A. Zhuk, An one-class classification support vector machine model by interval-valued training data, Knowl.-Based Syst. 120 (2017) 43–56.

[125] M. Nalenz, J. Rodemann, T. Augustin, Learning de-biased regression trees and forests from complex samples, Mach. Learn. 113 (6) (2024) 3379–3398.

[126] G. Walter, T. Augustin, Imprecision and prior-data conflict in generalized Bayesian inference, J. Stat. Theory Pract. 3 (1) (2009) 255–271.

[127] J. Abellan, S. Moral, Maximum of entropy for credal sets, Internat. J. Uncertain. Fuzziness Knowledge-Based Systems 11 (05) (2003) 587–597.

[128] J. Rodemann, C. Jansen, G. Schollmeyer, Reciprocal learning, arxiv (2024) Preprint, submitted for publication.

[129] J. Rodemann, Towards Bayesian data selection, in: 5th Workshop on Data-Centric Machine Learning Research (DMLR) at ICML 2024, 2024.

[130] M. Pincus, A Monte Carlo method for the approximate solution of certain types of constrained optimization problems, Oper. Res. 18 (6) (1970) 1225–1228.

[131] K. Deb, D. Deb, Analysing mutation schemes for real-parameter genetic algorithms, Int. J. Artif. Intell. Soft Comput. 4 (1) (2014) 1–28.

[132] T. Augustin, R. Hable, On the impact of robust statistics on imprecise probability models: a review, Struct. Saf. 32 (6) (2010) 358–365.

[133] M. Abrams, Natural selection with objective imprecise probability, in: 11. International Symposium on Imprecise Probabilities: Theories and Applications, 2019, pp. 2–13.

[134] J. Rodemann, T. Augustin, Accounting for imprecision of model specification in Bayesian optimization, in: Poster Presented at International Symposium on Imprecise Probabilities, ISIPTA, 2021.