Proceedings of Machine Learning Research 242:1–13, 2024

Towards Safe Multi-Task Bayesian Optimization

Jannis O. Lübsen[1] JANNIS.LUEBSEN@TUHH.DE
Christian Hespe[1] CHRISTIAN.HESPE@TUHH.DE
Annika Eichler[1,2] ANNIKA.EICHLER@TUHH.DE
[1]*Hamburg University of Technology, Germany*
[2]*Deutsches Elektronen-Synchrotron DESY, Germany*

Abstract

Bayesian optimization has emerged as a highly effective tool for the safe online optimization of systems, due to its high sample efficiency and noise robustness. To further enhance its efficiency, reduced physical models of the system can be incorporated into the optimization process, accelerating it. These models are able to offer an approximation of the actual system, and evaluating them is significantly cheaper. The similarity between the model and reality is represented by additional hyperparameters, which are learned within the optimization process. Safety is a crucial criterion for online optimization methods such as Bayesian optimization, which has been addressed by recent works that provide safety guarantees under the assumption of known hyperparameters. In practice, however, this does not apply. Therefore, we extend the robust Gaussian process uniform error bounds to meet the multi-task setting, which involves the calculation of a confidence region from the hyperparameter posterior distribution utilizing Markov chain Monte Carlo methods. Subsequently, the robust safety bounds are employed to facilitate the safe optimization of the system, while incorporating measurements of the models. Simulation results indicate that the optimization can be significantly accelerated for expensive to evaluate functions in comparison to other state-of-the-art safe Bayesian optimization methods, contingent on the fidelity of the models. The code is accessible on GitHub[1].

Keywords: Bayesian Optimization, Gaussian Processes, Controller Tuning, Safe Optimization

1. Introduction

Bayesian optimization (BO) is an iterative, learning-based, gradient-free, and global optimization method which gained attention in recent years. The method involves learning a probabilistic surrogate model of an arbitrary objective function in order to optimize it, while requiring minor assumptions. The utilization of Gaussian processes (GP) allows including prior knowledge of the objective which makes the procedure very suitable for expensive to evaluate functions and robust to noisy observations. Enhanced sample efficiency can be achieved by incorporating reduced models of the objective function into the optimization process, as demonstrated in a proof-of-principle study by Ferran Pousa et al. (2023); Letham and Bakshy (2019). The key tool, multi-task Gaussian process prediction, was initially presented in Bonilla et al. (2007). This approach utilizes correlation matrices to depict the influence between various tasks, learned from available data. Subsequently, Swersky et al. (2013) introduced the first multi-task Bayesian optimization algorithm, where its superior efficiency was highlighted. The key idea revolves around incorporating extra models of the

1. https://github.com/TUHH-ICS/2024-code-L4DC-Safe-Multi-Task-Bayesian-Optimization

actual function, allowing the prediction of the real function (main task) to be estimated by evaluating the model functions (supplementary tasks) when there exists some correlation. Since evaluating the model functions is significantly more cost-effective in practice, the optimization process can be accelerated considerably, depending upon the fidelity of the supplementary tasks.

Many real-world optimization problems require the consideration of constraints to avoid the evaluation of inputs which invoke undesirable effects, e.g., damaging the system. Often, these constraints are also unknown and need to be learned online. The theoretical fundament for safe Bayesian optimization was inherited from bounding regrets via uniform error bounds in a multi armed bandit problem Srinivas et al. (2010) and improved with respect to performance by Chowdhury and Gopalan (2017). Subsequently, Sui et al. (2015) used the results from Srinivas et al. (2010) to describe `SafeOpt`, the first safe Bayesian optimization method. Uniform error bounds are defined by scaling the posterior standard deviation by a constant. Typically, these bounds are of probabilistic nature, meaning they hold with high probability. In the previous mentioned works, the assumption is made that the unknown function is deterministic and belongs to the reproducing kernel Hilbert space (RKHS) defined by chosen kernel. In addition, several additional methodologies have been developed such as the work by Lederer et al. (2019); Sun et al. (2021) where the unknown function is assumed to be a sample of the prior distribution defined by the Gaussian process. In statistics, the former approaches are known to operate in the frequentist setting whereas the latter operate in the Bayesian setting, see Murphy (2012) for more information about frequentist and Bayesian statistics. However, all these approaches assume that the complexity of the Gaussian process aligns perfectly with the objective, implying that the kernel hyperparameters are known. In case of the RKHS methods, the situation is even worse since they require also knowledge about an upper bound of the RKHS norm of the unknown function. Clearly, from a practical standpoint, this is usually not the case. Furthermore, this aspect is particularly crucial, especially in the context of a multi-task setting, wherein the main task is biased by the supplementary tasks. Capone et al. (2022) introduced a method to address this issue by establishing robust bounds on hyperparameters through a Bayesian framework (c.f. Section 2.1), and Fiedler et al. (2021) proposed a robustness approach for the frequentist setting where the upper bound on the RKHS norm is still valid. However, these results are not applicable to a multi-task setting.

Contribution The primary contribution involves introduction of a Bayesian optimization algorithm designed to ensure safe optimization of a system while incorporating measurements from various tasks. This is achieved by extending the outcomes of Capone et al. (2022) to a multi-task setting with the use of Lemma 2 and 3. Furthermore, we assume to operate in the bayesian setting. To the best of the authors' knowledge, this is the first robustly safe Bayesian optimization algorithm in a multi-task setting. Finally, we underscore the significance of the proposal by conducting a benchmark comparison with other state-of-the-art Bayesian optimization methods.

2. Fundamentals

In Bayesian optimization, Gaussian processes are used to model an unknown objective function $f : \mathcal{X} \longrightarrow \mathbb{R}$, where the domain $\mathcal{X} \subseteq \mathbb{R}^d$ is a compact set of input parameters. In general, f is non-convex and learned online by evaluating the function at some inputs $\boldsymbol{x} \in \mathcal{X}$. The function values themselves are not accessible, rather noisy observations are made. This behavior can be modeled by additive Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$, i.e., $y = f(\boldsymbol{x}) + \epsilon$, where y is the measured value and σ_n^2 denotes the noise variance. Furthermore, we assume to have a set of observations

by $\mathcal{D} \coloneqq \{(\boldsymbol{x}_k, y_k), k = 1, \dots, n\}$ which is composed of the evaluated inputs combined with the corresponding observations. This set can be considered as the training set. A more compact notation of all inputs of $\mathcal{D}$ is given by the matrix $X = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_n]^T$ and of all observations by the vector $\boldsymbol{y} = [y_1, \dots, y_n]^T$. With this data set, the Gaussian process creates a probabilistic model to predict $f(\boldsymbol{x})$. These predictions serve as inputs for an acquisition function $\alpha(\cdot)$, which identifies new promising inputs likely to minimize the objective. A common choice is expected improvement (EI) as initially described by Jones et al. (1998).

2.1. Gaussian Processes

A Gaussian process is fully defined by a mean function $m(\boldsymbol{x})$ and a kernel $k(\boldsymbol{x}, \boldsymbol{x}')$. The function values $f(\boldsymbol{x})$ are modeled by normal distributions and the kernel determines the dependency between function values at different inputs

$$\operatorname{cov}(f(\boldsymbol{x}), f(\boldsymbol{x}')) = k(\boldsymbol{x}, \boldsymbol{x}'). \tag{1}$$

Commonly used kernels are the spectral mixture, Matérn or squared exponential kernel, where the latter is defined as $k_{\mathrm{SE}}(\boldsymbol{x}, \boldsymbol{x}') = \sigma_f^2 \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}')^T \Delta^{-2} (\boldsymbol{x} - \boldsymbol{x}')\right)$ with $\Delta = \operatorname{diag}(\boldsymbol{\vartheta}) = \operatorname{diag}([\vartheta_1, \dots, \vartheta_d]^T)$. The signal variance σ_f^2, the lengthscales $\boldsymbol{\vartheta}$ and the noise variance σ_n^2 constitute the hyperparameters, allowing for the adjustment of the kernel function. For a more compact notation, we define the tuple of hyperparameters by $\boldsymbol{\theta} = (\sigma_f^2, \boldsymbol{\vartheta}, \sigma_n^2)$.

Given the set of observations, the Gaussian process is used to predict the function values $f_* \mathrel{\widehat{=}} f(\boldsymbol{x}_*)$ at unobserved test points $\boldsymbol{x}_* \in \mathcal{X}$ by determining the posterior distribution $p(f_* | X, \boldsymbol{y}) = \mathcal{N}(\mu(\boldsymbol{x}_*), \sigma^2(\boldsymbol{x}_*))$. As shown by Williams and Rasmussen (2006) the posterior distribution is Gaussian given by

$$\begin{aligned}
\mu_{\boldsymbol{\theta}}(\boldsymbol{x}_*) &= K_{\boldsymbol{\theta}}(\boldsymbol{x}_*, X) \left(K_{\boldsymbol{\theta}} + \sigma_n^2 I\right)^{-1} \boldsymbol{y} \\
\sigma_{\boldsymbol{\theta}}^2(\boldsymbol{x}_*) &= K_{\boldsymbol{\theta},*} - K_{\boldsymbol{\theta}}(\boldsymbol{x}_*, X) \left(K_{\boldsymbol{\theta}} + \sigma_n^2 I\right)^{-1} K_{\boldsymbol{\theta}}(X, \boldsymbol{x}_*),
\end{aligned}$$

where $K_{\boldsymbol{\theta}} = K_{\boldsymbol{\theta}}(X, X)$ and $K_{\boldsymbol{\theta},*} = K_{\boldsymbol{\theta}}(\boldsymbol{x}_*, \boldsymbol{x}_*)$ are the Gram matrices of the training and test data, respectively. We use the subscript notation to emphasize the dependcy of the respective on the hyperparameters $\boldsymbol{\theta}$.

LEARNING HYPERPARAMETERS

The selection of hyperparameters is a challenging task due to their substantial impact on the predictive distribution. If selected inaccurately, the predictions do not agree with the true function and safeness conditions may not hold. The most common method is to select the hyperparameters such that the log marginal likelihood (2) is maximized, which provides a good trade-off between function complexity and accuracy with respect to the data, given by

$$\log p(\boldsymbol{y} | X, \boldsymbol{\theta}) = -\frac{1}{2} \log |\tilde{K}_{\boldsymbol{\theta}}| - \frac{N}{2} \log(2\pi) - \frac{1}{2} \boldsymbol{y}^T \tilde{K}_{\boldsymbol{\theta}}^{-1} \boldsymbol{y} \tag{2}$$

with $\tilde{K}_{\boldsymbol{\theta}} = K_{\boldsymbol{\theta}} + \sigma_n^2 I$. The maximization can be carried out efficiently using gradient based methods. However, the log marginal likelihood is non-convex in most cases which makes global optimization challenging. It is commonly assumed that the initial hyperparameters are proximate to

the global optimum. In scenarios with limited information about the underlying objective, selecting suitable hyperparameters becomes exceptionally difficult. Consequently, the optimization process may become trapped in a local optimum Williams and Rasmussen (2006).

An advanced strategy was introduced by Capone et al. (2022). Here, the authors consider the lengthscales in a single-task setting and perform model selection in a Bayesian sense. The idea is to replace the deterministically chosen initial lengthscales $\boldsymbol{\vartheta} \in \mathbb{R}^d$ by a prior distribution $p(\boldsymbol{\vartheta})$. Then, for a given data set $\mathcal{D} = \{X, \boldsymbol{y}\}$ the posterior distribution of the lengthscales can be computed by applying Bayes rule

$$p(\boldsymbol{\vartheta}|X, \boldsymbol{y}) = \frac{p(\boldsymbol{y}|X, \boldsymbol{\vartheta})p(\boldsymbol{\vartheta})}{p(\boldsymbol{y}|X)}. \tag{3}$$

Note that computing (3) analytically is not tractable in general, however, one can use approximations such as Laplace approximation or Markov chain Monte Carlo (MCMC) methods to estimate the posterior. Considering the posterior distribution rather than deterministic values offers the opportunity to establish robust probability bounds wherein the majority of the probability mass lies, i.e.,

$$P_\delta = \left\{ (\boldsymbol{\vartheta}', \boldsymbol{\vartheta}'') \in \Theta^2 \middle| \int_{\boldsymbol{\vartheta}' \leq \boldsymbol{\vartheta} \leq \boldsymbol{\vartheta}''} p(\boldsymbol{\vartheta}|X, \boldsymbol{y}) \mathrm{d}\boldsymbol{\vartheta} \geq 1 - \delta \right\}, \tag{4}$$

where $1 - \delta$ is the confidence interval with $\delta \in (0, 1)$. After estimating the ranges, one can apply Lemma 3.3 and Theorem 3.5 from Capone et al. (2022) to obtain the new scaling factor

$$\bar{\beta} = \gamma^2 \left(\max_{\boldsymbol{\vartheta}' \leq \boldsymbol{\vartheta} \leq \boldsymbol{\vartheta}''} \beta^{\frac{1}{2}}(\boldsymbol{\vartheta}) + \frac{2\,\|y\|_2}{\sigma_n} \right)^2, \tag{5}$$

with $\gamma^2 = \prod_{i=1}^{d} \frac{\vartheta_i''}{\vartheta_i'}$, that guarantees with probability $(1 - \delta)(1 - \rho)$ that $|f(\boldsymbol{x}) - \mu_{\boldsymbol{\vartheta}_0}(\boldsymbol{x})| \leq \bar{\beta}^{\frac{1}{2}} \sigma_{\boldsymbol{\vartheta}'}(\boldsymbol{x}), \forall \boldsymbol{x} \in \mathcal{X}$, where $\rho > 0$ denotes the failure probability.

2.2. Multi-Task Gaussian Processes

So far, only scalar functions are considered, which is different from the considered setting. Since we want to include simulator observations, the GP needs to be extended to model vector-valued functions $\boldsymbol{f} = [f_1, \dots f_u] : \mathcal{X} \to \mathbb{R}^u$. To tackle inter-function correlations, the covariance function from (1) is extended by additional task inputs, i.e., $k((\boldsymbol{x}, t), (\boldsymbol{x}', t')) = \mathrm{cov}(f_t(\boldsymbol{x}), f_{t'}(\boldsymbol{x}'))$ where $t, t' \in \{1, \dots, u\}$ denote the task indices and u is the total number of information sources. If the kernel can be separated into a task-depending k_t and an input-depending kernel k_x, i.e., $k((\boldsymbol{x}, t), (\boldsymbol{x}', t')) = k_t(t, t')\, k_x(\boldsymbol{x}, \boldsymbol{x}')$, then the kernel is called separable which is used in most literature, e.g., Bonilla et al. (2007); Letham and Bakshy (2019); Swersky et al. (2013). Typically, this is represented by the intrinsic coregionalization model (ICM) Álvarez and Lawrence (2011), in which k_x measures the dependency in the input space, while k_t measures the dependency between tasks.

Since the task indices are integers and independent of the inputs, the task covariance function is usually substituted by constants $\sigma^2_{t,t'}$ which are treated as additional hyperparameters. Hence, defining the correlation matrix $\Sigma = [\sigma_{t,t'}]_{t,t'=1}^{u}$, we have $\Sigma\, k(\boldsymbol{x}, \boldsymbol{x}') = \mathrm{cov}(\boldsymbol{f}(\boldsymbol{x}), \boldsymbol{f}(\boldsymbol{x}'))$ which

denotes the multi-task kernel. In this setting, it is reasonable to assume positive correlation between tasks solely. Moreover, it is assumed that for each task there exists a data set $\mathcal{D}_i$ which is stacked into a global set $\boldsymbol{\mathcal{D}} \coloneqq \{\boldsymbol{X}, \tilde{\boldsymbol{y}}\}$, where $\boldsymbol{X} = [X_1^T, \dots, X_u^T]^T$ and $\tilde{\boldsymbol{y}} = [\boldsymbol{y}_1^T, \dots, \boldsymbol{y}_u^T]^T$. Then, the Gram matrix is given by

$$\boldsymbol{K}_\Sigma = \boldsymbol{K}_\Sigma(\boldsymbol{X}, \boldsymbol{X}) = \begin{bmatrix} \sigma_{1,1}^2 K_{1,1} & \dots & \sigma_{1,u}^2 K_{1,u} \\ \vdots & \ddots & \vdots \\ \sigma_{u,1}^2 K_{u,1} & \dots & \sigma_{u,u}^2 K_{u,u} \end{bmatrix},$$

where $K_{t,t'}$ are Gram matrices using data sources t and t'. For notational reasons, we neglect $\boldsymbol{\theta}$ to denote the dependency of the Gram matrices on the remaining hyperparameters, which are equivalent for all tasks. Note that if the covariance entries are zero, i.e., $\sigma_{t,t'} = 0, \forall t \neq t'$, the off-diagonal blocks of $\boldsymbol{K}_\Sigma$ are zero which means that all information sources are independent and can be divided into u separate Gaussian processes. To perform inference in the multi-task setting, one simply needs to substitute the single-task Gram matrix K and measurements $\boldsymbol{y}$ by their multi-task equivalents $\boldsymbol{K}_\Sigma$ and $\tilde{\boldsymbol{y}}$.

3. Extension of Uniform Error Bounds for Unknown Source Correlation

Now, we introduce an extension of the uniform error bounds for unknown hyperparameters. We consider the correlation matrix $\Sigma \in \mathcal{C}$ to be the only uncertain hyperparameter and $\mathcal{C} \subset \mathbb{R}^{u \times u}$ to be the set of all positive definite correlation matrices, where u denotes the number of sources. By definition, this type of matrices are real and symmetric. In addition, we introduce $\mathbb{P}(\cdot) : \mathbb{R}^u \to [0, 1]$ to denote the Gaussian measure. The results presented in this section and in Capone et al. (2022) can be combined to consider also the uncertainty of the lengthscales.

In this work, we operate in the Bayesian setting, which is manifested in the following assumption:

Assumption 1 *The function $\boldsymbol{f}(\cdot)$ is a sample of a Gaussian process with multi-task kernel $\Sigma\, k(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}^{u \times u}$ and hyperprior $p(\Sigma)$ of positive definite correlation matrices.*

If a base kernel $k(\cdot, \cdot)$ is selected that satisfy the universal approximation property Micchelli et al. (2006), then the multi-task kernel $\Sigma\, k(\cdot, \cdot)$ also satisfies the universal approximation property Caponnetto et al. (2008) which makes the assumption non-restrictive. In addition, we assume that there exists a known scaling function $\beta : \mathcal{C} \to \mathbb{R}^+$ such that

$$\mathbb{P}\left(|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{\mu}_\Sigma(\boldsymbol{x})| \leq \beta^{\frac{1}{2}}(\Sigma)\boldsymbol{\sigma}_\Sigma(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in \mathcal{X}\right) \geq 1 - \rho, \tag{6}$$

equivalently to Assumption 3.2 by Capone et al. (2022). There exist different methods to define the scaling functions, e.g., Lederer et al. (2019). If the input space $\mathcal{X}$ has finite cardinality, β is constant and independent of the hyperparameters Srinivas et al. (2010). In (4), the posterior probability density function is integrated over the interval from $\boldsymbol{\vartheta}'$ to $\boldsymbol{\vartheta}''$ which is selected such that some confidence is included. First, we need to define a relation for $\mathcal{C}$ to be able to order the members of the set. Therefore, the function $h(\Sigma_1, \Sigma_2) = \max \mathrm{eig} \Sigma_1^{-1} \Sigma_2$ is introduced which maps two correlation matrices into the positive reals. The reason for selecting this function follows from the definition of the error bounds in Lemma 2. Using $h(\cdot, \cdot)$ we define the set

$$\mathcal{C}_{\Sigma', \Sigma''} = \{\Sigma \in \mathcal{C} | h(\Sigma', \Sigma) \leq h(\Sigma', \Sigma'')\}, \tag{7}$$

which is a subset of $\mathcal{C}$ and comprises all correlation matrices that have a smaller value in h than two selected (Σ', Σ''). Now we are able to define a set of bounding correlation matrices by

$$\mathcal{C}_\delta = \left\{ (\Sigma', \Sigma'') \in \mathcal{C}^2 \middle| \int_{\mathcal{C}_{\Sigma',\Sigma''}} p(\Sigma|\boldsymbol{X}, \tilde{\boldsymbol{y}}) \mathrm{d}\Sigma \geq 1 - \delta \right\}.$$

In words, we are searching for a bound Σ', Σ'' such that their corresponding set $\mathcal{C}_{\Sigma',\Sigma''}$ comprises enough members that establish a predefined confidence region on the posterior. After identifying a matching pair of bounds, we are able to modify the scaling function $\beta(\cdot)$ such that safeness for the uncertain hyperparameters can be guaranteed. The extension is summarized in the following results.

Lemma 2 *Let $\boldsymbol{\sigma}_{\Sigma'}(\boldsymbol{x}), \boldsymbol{\sigma}_{\Sigma}(\boldsymbol{x})$ be the posterior variance conditioned on the data $\boldsymbol{\mathcal{D}}$ with different correlation matrices $\Sigma', \Sigma'' \in \mathcal{C}_\delta$, $\Sigma \in \mathcal{C}_{\Sigma',\Sigma''}$, and let $\gamma^2 \geq h(\Sigma', \Sigma'')$. Then*

$$\gamma^2 \boldsymbol{\sigma}_{\Sigma'}(\boldsymbol{x}) \quad \geq \quad \boldsymbol{\sigma}_{\Sigma}(\boldsymbol{x}), \quad \forall \boldsymbol{x} \in \mathcal{X}, \forall \Sigma \in \mathcal{C}_{\Sigma',\Sigma''}.$$

Proof See Lübsen et al. (2024). ■

Lemma 2 bounds the posterior variance for all correlation matrices in the set $\mathcal{C}_{\Sigma',\Sigma''}$. However, bounding the posterior variance solely is not sufficient since the error bounds also depend on the posterior mean. This is addressed in the following Lemma 3.

Lemma 3 *Let $\boldsymbol{\mu}$ be a member of the reproducing kernel Hilbert space (RKHS) $\mathcal{H}''$ with kernel $K''(\boldsymbol{x}, \boldsymbol{x}') = \Sigma'' \, k(\boldsymbol{x}, \boldsymbol{x}')$, and let $\mathcal{H}'$ be an RKHS with kernel $K'(\boldsymbol{x}, \boldsymbol{x}') = \Sigma' \, k(\boldsymbol{x}, \boldsymbol{x}')$. Then with $\lambda^2 \geq h(\Sigma'', \Sigma')$ it follows*

$$\lambda^2 ||\boldsymbol{\mu}||^2_{K'} \quad \geq \quad ||\boldsymbol{\mu}||^2_{K''}.$$

Proof See Lübsen et al. (2024). ■

Lemma 3 is used to bound the posterior mean in different RKHS. In contrast to the unknown function $\boldsymbol{f}$ which is not restricted to the RKHS as stated in Assumption 1, the posterior mean is a member of the RKHS since $\boldsymbol{\mu}(\boldsymbol{x}_*) = \langle \boldsymbol{\alpha}, K(\boldsymbol{x}_*, \cdot) \rangle_K$ with $\boldsymbol{\alpha} = (\boldsymbol{K}_\Sigma + \sigma_n^2 \, I)^{-1} \tilde{\boldsymbol{y}}$. With the two intermediate results robust error bounds in the multi-task setting can be specified as summarized in Theorem 4.

Theorem 4 *Let Assumption 1 hold, and assume there exists a scaling function $\beta(\cdot)$ and a failure probability ρ which are known such that (6) holds. Let $(\Sigma', \Sigma'') \in \mathcal{C}_\delta$ with posterior of hyperparameters $p(\Sigma|\boldsymbol{X}, \tilde{\boldsymbol{y}})$, let $\boldsymbol{\sigma}^2_{\Sigma'}$ be the posterior variance and $\boldsymbol{\mu}_{\Sigma_0}$ the posterior mean obtained with correlation matrices Σ' and $\Sigma_0 \in \mathcal{C}_{\Sigma',\Sigma''}$, respectively, and select $\gamma^2 \geq h(\Sigma', \Sigma'')$ and $\lambda^2 \geq \max_{\Sigma \in \mathcal{C}_{\Sigma',\Sigma''}} h(\Sigma'', \Sigma)$. Then with*

$$\bar{\beta} = \left(\lambda \frac{2 \, \|\tilde{\boldsymbol{y}}\|_2}{\sigma_n} + \gamma \max_{\Sigma \in \mathcal{C}_{\Sigma',\Sigma''}} \beta^{\frac{1}{2}}(\Sigma) \right)^2$$

we have

$$\mathbb{P}\left(|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{\mu}_{\Sigma_0}(\boldsymbol{x})| \leq \bar{\beta}^{\frac{1}{2}} \boldsymbol{\sigma}_{\Sigma'}, \quad \forall \boldsymbol{x} \in \mathcal{X} \right) \geq (1 - \delta)(1 - \rho).$$

Proof The result can be obtained by following the same steps as in Theorem 3.7 by Capone et al. (2022) where the scaling parameter γ from Lemma 2 replaces Lemma 3.3, and λ from Lemma 3 replaces Lemma A.7 in Capone et al. (2022). We take the maximum λ over the set $\mathcal{C}_{\Sigma',\Sigma''}$ to ensure safeness for every $\Sigma_0 \in \mathcal{C}_{\Sigma',\Sigma''}$. For γ this follows immediately from the definition of the set $\mathcal{C}_{\Sigma',\Sigma''}$ in (7). ■

The correlation bounds (Σ', Σ'') and scaling parameters γ, λ can be computed numerically, for example, via comparing samples from the hyperparameter posterior which can be generated with MCMC methods. In general, the computational cost of inverting correlation matrices is lower when compared to computing the posterior distribution of a Gaussian process, as these matrices are notably smaller than the Gram matrix. The highest time consumption comes from computing the posterior distribution and generating posterior samples. Regarding the computation of posterior distributions, there are several approximation methods that can be employed to enhance the inference speed when dealing with a substantial amount of data points Snelson and Ghahramani (2005); Titsias (2009).

4. Safe Multi-Task Bayesian Optimization

In safe Bayesian optimization, the goal is to minimize a multi-output objective $\boldsymbol{f}(\boldsymbol{x})$ while including constraints $\boldsymbol{g}(\boldsymbol{x})$ for each $f_i(\boldsymbol{x})$, i.e., $\min_{\boldsymbol{x}\in\mathcal{X}} \boldsymbol{f}(\boldsymbol{x}) \quad \text{s.t.} \quad \boldsymbol{g}(\boldsymbol{x}) \geq 0$. Frequently $\boldsymbol{g}(\boldsymbol{x}) = T - \boldsymbol{f}(\boldsymbol{x})$, where T is a safety threshold. Nonetheless, all the results are applicable to arbitrary constraints that fulfill Assumption 1. Moreover, it is necessary to possess information regarding an initial safe set, denoted as $\mathcal{S}_0$ which comprises at least one input that fulfills the constraints. This assumption is widespread, as asserted by Kirschner et al. (2019); Sui et al. (2015), even in practical scenarios where an initial estimation is typically provided. Note that since $\boldsymbol{f}(\boldsymbol{x})$ is unknown also $\boldsymbol{g}(\boldsymbol{x})$ is unknown according to the definition of the constraints and needs to be learned online. This means that only stochastic statements can be made about satisfying the constraints. Similarly to `SafeOpt-MC` in Berkenkamp et al. (2021), the safe set is defined to be $\mathcal{S} \coloneqq \cap_{i=1}^{u} \{\boldsymbol{x} \in \mathcal{X} | \mu_{\Sigma',i}(\boldsymbol{x}) + \bar{\beta}^{\frac{1}{2}} \sigma_{\Sigma',i}(\boldsymbol{x}) \leq T\}$, which includes inputs where the constraints are fulfilled with high probability according to Theorem 4. Starting from the initial safe set, the domain is explored to find the minimum. There exist various approaches to conduct this optimization, in Berkenkamp et al. (2016) the most uncertain points in a subset of $\mathcal{S}$ are repetitively evaluated. In Sui et al. (2018) and Lübsen et al. (2023), the exploration and exploitation phases are separated, meaning that the algorithm first focuses on expanding $\mathcal{S}$ and then on optimization. The commonality among all methods is their requirement for numerous evaluations of the objective function. This can be reduced by including simulations of the objective in the optimization.

Figure 1 compares the safe region of the naive single-task optimization (a) with the multi-task setting where a low fidelity (b) and a high fidelity model (c) are considered. The fidelity of a model is reflected in the correlation matrix; the higher the agreement of model and the ground truth, the higher the information quality of the function value. Figure 1 (b) shows that even small correlations extend the safe set.

The proposed method is summarized in Algorithm 1. We start with an initial safe set, a scaling function $\beta(\cdot)$ and a GP equipped with an appropriate hyperprior that reflects the initial guess about the hyperparameter. In the loop, posterior samples are generated and collected in the set $\mathcal{P}$. The samples allow to approximate the robust scaling parameter from Theorem 4. In addition, Σ' is

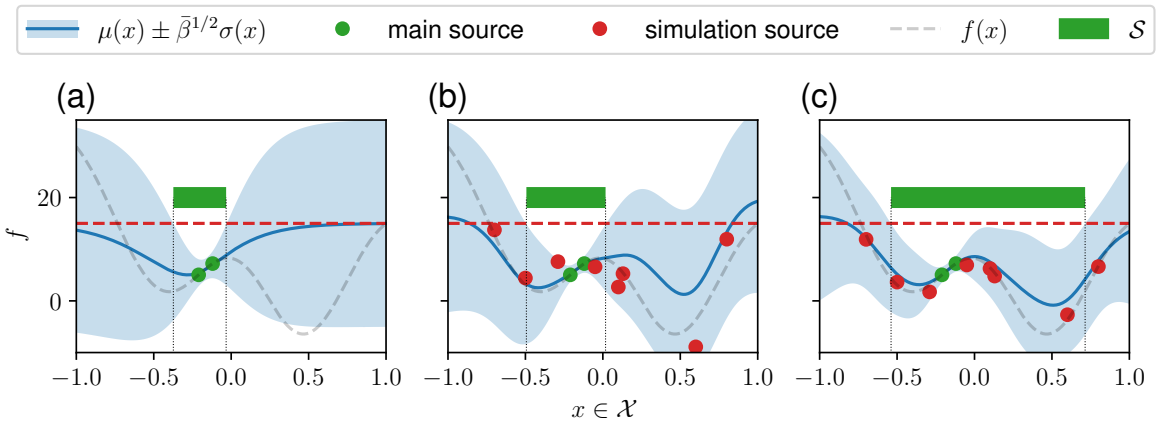

Figure 1: Overview of different safe Bayesian optimization settings with safety threshold T denoted by "- - -". (a) shows the single-task setting, where no simulation samples are considered, and the safe region is the smallest. (b) visualizes the multi-task setting with slight correlation and (c) with high correlation. In both cases, The multi-task setting increases the safe region.

Algorithm 1: Safe Multi-Source Bayesian Optimization (`SaMSBO`)

Input: Initial safe set $\mathcal{S}_0$, multi-task Gaussian process model $\mathcal{GP}$ with hyperpriors $p(\Sigma)$, scaling function $\beta(\cdot)$

Output: $\boldsymbol{x}_{\text{opt}}$

while *termination condition not true* **do**

- $\mathcal{P} \leftarrow \text{MCMC}(\mathcal{GP}, p(\Sigma))$ `// generate samples via MCMC methods`
- $\bar{\beta}^{1/2} \leftarrow \mathcal{P}, \mathcal{GP} \leftarrow \Sigma'$ `// Determine` $\bar{\beta}$ `from samples according to Theorem 4`
- `// load` Σ' `into GP`
- $\mathcal{GP}, \boldsymbol{\mathcal{D}} \leftarrow \text{BO}(\mathcal{GP})$ `// Perform BO step for all tasks and update GP model`

end

$\boldsymbol{x}_{\text{opt}} \leftarrow \arg\min \boldsymbol{\mathcal{D}}$ `// return best main task solution`

loaded into the GP model. Then, the predictive distribution is used by an acquisition function to identify promising inputs at which the objective functions are evaluated, and both the GP model and the data set are updated. Finally, after the loop terminates, the best input of the main task is returned.

5. Simulation Results

In this section we demonstrate the proposed algorithm in simulation. The considered plant models the laser-based synchronization (LbSync) system Schulz et al. (2015) at European XFEL, akin to the system employed in Lübsen et al. (2023) and depicted in Figure 2. The difference lies in the fact that all models G represent the same laser models in this scenario, in other words, $G_i = G_j, \forall i, j = 1, \ldots, N$. The filter models F_r and $F_{1:N}$ colorize the white Gaussian noise inputs

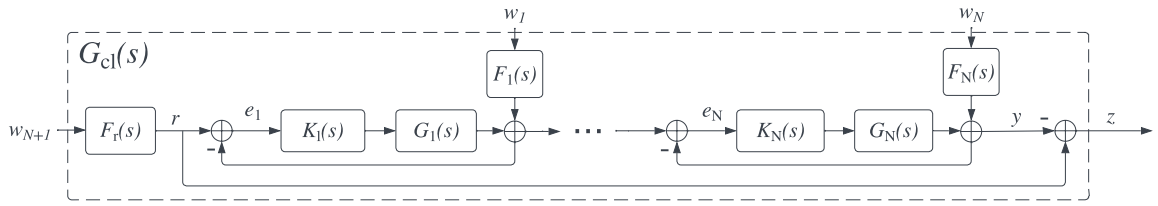

Figure 2: Illustration of the interconnected system. The blocks F_r and $F_i, i = 1, \ldots, N$ denote disturbance filters which colorize the white noise inputs $w_j, j = 1, \ldots, N+1$. G_i denote the laser plants and K_i PI controllers for each subsystem.

$w_{1:N+1}$ to model environmental disturbances, e.g., vibrations, temperature changes and humidity. Throughout the optimization procedures, we operate under the assumption that, alongside the main task, there exist two supplementary tasks of information that simulate the main task, i.e., $\boldsymbol{f}(\boldsymbol{x}) = [f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), f_3(\boldsymbol{x})]^T$ where $f_1(\boldsymbol{x})$ is the main task. TIn order to emulate the discrepancy between simulation and reality, the primary task employs nominal models, while supplementary tasks are subjected to disturbances. It is assumed that uncertainty resides in the filter models, given that the laser model can be accurately identified and exhibits minimal variation over time, while disturbance sources change more frequently. Furthermore, it is essential to ensure the safety of the primary task i.e., $\boldsymbol{g}(\boldsymbol{x}) = [T - f_1(\boldsymbol{x}), 0, 0]^T$, as the supplementary tasks are simulations, where $T = 30$. Hence, evaluating unsafe regions will not damage the system, rather, this can help to estimate the safe region. The goal is to minimize the root-mean-square seminorm of the performance output z by tuning the PI parameters of the controllers $K_{i:N}$. Following Parseval's Theorem this corresponds to a H_2 minimization of the closed-loop system Heuer (2018). Implementations are carried out using GPyTorch Gardner et al. (2018) and BoTorch Balandat et al. (2020), and MCMC samples are generated with the No-U Turn Sampler algorithm Hoffman and Gelman (2014). In addition, a Lewandowski-Kurowicka-Joe distribution Lewandowski et al. (2009) is used for prior distribution over correlation matrices, because it easily allows including prior knowledge about the expected correlation by adjusting the shape factor $\eta \in (0, \infty)$. Setting $\eta < 1$ matrices with higher correlation are favored, while for $\eta > 1$ low correlation matrices are favored. Note that, the definitions in Lemma 2, 3 provide error bounds to ensure safeness along all tasks, while in this setting, only the main task needs to be safe. Therefore, we neglect the influence of the uncertainty of the posterior mean in Theorem 4, i.e., $\lambda^2 = 0$, to avoid overly conservative error bounds in the optimization. In addition, we select a constant scaling factor $\beta = 4$ and adjust the remaining hyperparameters such that the single-task optimization is safe. Per BO step, one evaluation of the main task and 15 evaluations of the supplementary tasks are taken.

Figure 3 summarizes the benchmark results. In (a), the robustness of the algorithm is investigated, where the supplementary tasks are constructed with disturbed filter transfer functions. The disturbances are generated by sampling from a uniform distribution with magnitude according to the line color in the legend. A $\pm 10\%$ variation signifies that the state space model values can fluctuate by up to 10% (in both directions) from their nominal values. (a) displays the average of the best observation from 20 instances of the optimization. For each instance, the filter models' disturbances are resampled. Clearly, the number of observations increases with the uncertainty of the supplementary tasks because the optimal solutions among the tasks may not perfectly match. Nevertheless,

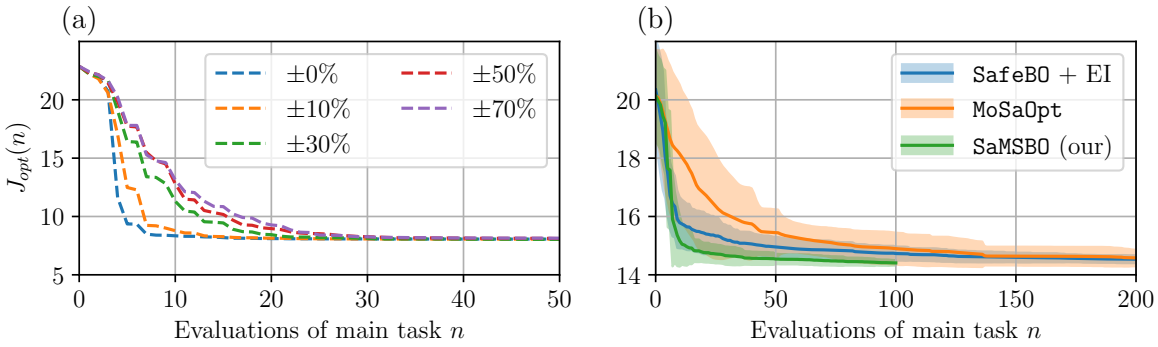

Figure 3: (a) shows the performance of SaMSBO by tuning a chain for $N = 2$ lasers, where the extra tasks have disturbed filter transfer functions. The line colors indicate the range of the disturbance. (b) shows the performance of different BO algorithms applied on a chain with $N = 5$ lasers. In this trial the filter disturbance lies in the range $\pm 10\%$.

the optimal solutions are found for all uncertainties. In (b) the initial points are changing, and the models are fixed across the iterations. Additionally, the outcomes obtained by applying a Safe BO algorithm, similar to SafeOpt Sui et al. (2015) but with EI acquisition function, and MoSaOpt, a line search method from Lübsen et al. (2023), are plotted. The mean values are denoted by the lines and the shaded area represents the standard deviation. SaMSBO outperforms both regarding solution quality and sample efficiency of the main task.

6. Conclusion and Outlook

We proposed the first robustly safe Bayesian optimization algorithm in a multi-task setting. We theoretically derived bounds that guarantee safeness for unknown correlation hyperparameters with high probability. Moreover, the proposed algorithm was benchmarked against other state-of-the-art methods in simulation. SaMSBO demonstrated superior solution quality and sample efficiency, ultimately achieving the most favorable convergence rate. However, one drawback is the doubled computation time per iteration compared to naive multi-task BO, due to the MCMC approximation of the hyper posterior, especially in later stages of the algorithm. Nevertheless, based on the simulation results, the total optimization time is still significantly reduced overall.

In the future, numerous aspects can be enhanced. Primarily, the scaling bounds specified in Lemma 2, 3 will become less stringent if safety is required for the main task solely, which is in real applications usually the case. Furthermore, including an efficient exploration technique of the safe set is crucial, particularly in high-dimensional scenarios. An approach to tackle this challenge has been introduced by Zagorowska et al. (2023), wherein exploration is reformulated as an optimization problem which can be efficiently solved. Finally, our plan involves testing the algorithm in a real environment by optimizing the LbSync system at European XFEL.

Acknowledgments

We acknowledge support from Deutsches Elektronen-Synchrotron DESY Hamburg, Germany, a member of the Helmholtz Association HGF.

References

Mauricio A. Álvarez and Neil D. Lawrence. Computationally efficient convolved multiple output Gaussian processes. *Journal of Machine Learning Research*, 12(41):1459–1500, 2011.

Maximilian Balandat, Brian Karrer, Daniel R. Jiang, Samuel Daulton, Benjamin Letham, Andrew Gordon Wilson, and Eytan Bakshy. Botorch: A framework for efficient monte-carlo Bayesian optimization. In *Advances in Neural Information Processing Systems 33*, 2020.

Felix Berkenkamp, Angela P. Schoellig, and Andreas Krause. Safe controller optimization for quadrotors with Gaussian processes. In *IEEE Int. Conf. Robot. Autom (ICRA)*, pages 491–496, 2016.

Felix Berkenkamp, Andreas Krause, and Angela P. Schoellig. Bayesian optimization with safety constraints: safe and automatic parameter tuning in robotics. *Machine Learning*, 2021. ISSN 1573-0565.

Edwin V Bonilla, Kian Chai, and Christopher Williams. Multi-task Gaussian process prediction. In *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

Alexandre Capone, Armin Lederer, and Sandra Hirche. Gaussian process uniform error bounds with unknown hyperparameters for safety-critical applications. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

Andrea Caponnetto, Charles A. Micchelli, Massimiliano Pontil, and Yiming Ying. Universal multi-task kernels. *Journal of Machine Learning Research*, 9(52):1615–1646, 2008.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 844–853. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/chowdhury17a.html.

A. Ferran Pousa, S. Jalas, M. Kirchen, A. Martinez de la Ossa, M. Thévenet, S. Hudson, J. Larson, A. Huebl, J.-L. Vay, and R. Lehe. Bayesian optimization of laser-plasma accelerators assisted by reduced physical models. *Phys. Rev. Accel. Beams*, 26:084601, Aug 2023. doi: 10.1103/PhysRevAccelBeams.26.084601.

Christian Fiedler, Carsten Scherer, and Sebastian Trimpe. Practical and rigorous uncertainty bounds for gaussian process regression. 05 2021.

Jacob R Gardner, Geoff Pleiss, David Bindel, Kilian Q Weinberger, and Andrew Gordon Wilson. Gpytorch: Blackbox matrix-matrix Gaussian process inference with gpu acceleration. In *Advances in Neural Information Processing Systems*, 2018.

Michael Heuer. *Identification and control of the laser-based synchronization system for the European X-ray Free Electron Laser*. Doctoral dissertation, Technische Universität Hamburg-Harburg, 2018.

Matthew D. Hoffman and Andrew Gelman. The no-u-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, pages 1–30, 2014.

Donald R. Jones, Matthias Schonlau, and William J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, Dec 1998. ISSN 1573-2916.

Johannes Kirschner, Mojmir Mutny, Nicole Hiller, Rasmus Ischebeck, and Andreas Krause. Adaptive and safe Bayesian optimization in high dimensions via one-dimensional subspaces. In *36th Int. Conf. Mach. Learn. (ICML)*, pages 3429–3438, 2019.

Armin Lederer, Jonas Umlauft, and Sandra Hirche. Uniform error bounds for Gaussian process regression with application to safe control. In *Advances in Neural Information Pro-cessing Systems*, page 659–669, June 2019.

Benjamin Letham and Eytan Bakshy. Bayesian optimization for policy search via online-offlineexperimentation. *J. Mach. Learn. Res.*, pages 593–1623, 2019.

Daniel Lewandowski, Dorota Kurowicka, and Harry Joe. Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100(9):1989–2001, 2009. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2009.04.008.

Jannis O. Lübsen, Maximilian Schütte, Sebastian Schulz, and Annika Eichler. A safe Bayesian optimization algorithm for tuning the optical synchronization system at European XFEL. In *22th World Congr. Int. Fed. Autom. Control (IFAC)*, 2023.

Jannis O. Lübsen, Christian Hespe, and Annika Eichler. Safe Multi-Task Bayesian Optimization. 2024. URL https://arxiv.org/abs/2312.07281.

Charles A. Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 7(95):2651–2667, 2006.

Kevin P Murphy. *Machine learning: a probabilistic perspective*. Cambridge, MA, 2012.

S Schulz, Ivanka Grguras, C Behrens, Hubertus Bromberger, John Costello, Marie Czwalinna, Matthias Felber, M Hoffmann, M Ilchen, H Liu, Tommaso Mazza, M Meyer, Sven Pfeiffer, Pawel Predki, S Schefer, Christian Schmidt, U Wegner, H. Schlarb, and A Cavalieri. Femtosecond all-optical synchronization of an X-ray free-electron laser. *Nature communications*, 6:5938, 01 2015.

Edward Snelson and Zoubin Ghahramani. Sparse gaussian processes using pseudo-inputs. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005. URL https://proceedings.neurips.cc/paper_files/paper/2005/file/4491777b1aa8b5b32c2e8666dbe1a495-Paper.pdf.

Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. pages 1015–1022, 07 2010.

Yanan Sui, Alkis Gotovos, Joel Burdick, and Andreas Krause. Safe exploration for optimization with Gaussian processes. In *32nd Int. Conf. Mach. Learn. (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, pages 997–1005, Lille, France, 07–09 Jul 2015. PMLR.

Yanan Sui, Vincent Zhuang, Joel W. Burdick, and Yisong Yue. Stagewise safe Bayesian optimization with Gaussian processes. In *35th Int. Conf. Mach. Learn. (ICML)*, 2018.

Dawei Sun, Mohammad Javad Khojasteh, Shubhanshu Shekhar, and Chuchu Fan. Uncertain-aware safe exploratory planning using Gaussian process and neural control contraction metric. In *Proceedings of the 3rd Conference on Learning for Dynamics and Control*, volume 144 of *Proceedings of Machine Learning Research*, pages 728–741. PMLR, 07 – 08 June 2021.

Kevin Swersky, Jasper Snoek, and Ryan P. Adams. Multi-task Bayesian optimization. In *Advances in Neural Information Processing Systems*, 2013.

Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In David van Dyk and Max Welling, editors, *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 567–574, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR. URL https://proceedings.mlr.press/v5/titsias09a.html.

Christopher K Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

Marta Zagorowska, Efe C. Balta, Varsha Behrunani, Alisa Rupenyan, and John Lygeros. Efficient sample selection for safe learning. In *22th World Congr. Int. Fed. Autom. Control (IFAC)*, 2023.