Gaussian Process Uniform Error Bounds with Unknown Hyperparameters for Safety-Critical Applications

Alexandre Capone [1]   **Armin Lederer** [1]   **Sandra Hirche** [1]

Abstract

Gaussian processes have become a promising tool for various safety-critical settings, since the posterior variance can be used to directly estimate the model error and quantify risk. However, state-of-the-art techniques for safety-critical settings hinge on the assumption that the kernel hyperparameters are known, which does not apply in general. To mitigate this, we introduce robust Gaussian process uniform error bounds in settings with unknown hyperparameters. Our approach computes a confidence region in the space of hyperparameters, which enables us to obtain a probabilistic upper bound for the model error of a Gaussian process with arbitrary hyperparameters. We do not require to know any bounds for the hyperparameters a priori, which is an assumption commonly found in related work. Instead, we are able to derive bounds from data in an intuitive fashion. We additionally employ the proposed technique to derive performance guarantees for a class of learning-based control problems. Experiments show that the bound performs significantly better than vanilla and fully Bayesian Gaussian processes.

1. Introduction

Gaussian processes (GPs) have become an often-used tool for regression due to their flexibility and good generalization properties. In addition to being successful at approximating unknown functions, GPs also come equipped with a measure of model uncertainty in the form of the posterior variance (Rasmussen & Williams, 2006). This quantity has shown promising results for estimating the error between the posterior mean and the underlying function (Srinivas et al., 2012; Chowdhury & Gopalan, 2017; Lederer et al.,
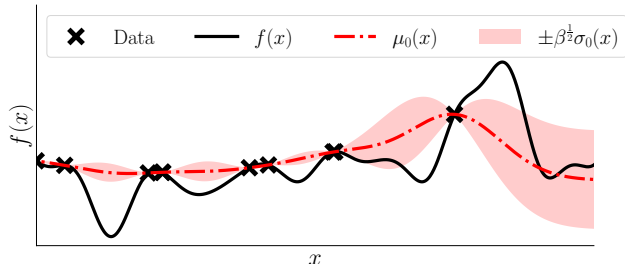
[1]TUM School of Computation, Information and Technology, Technical University of Munich, Munich, Germany. Correspondence to: Alexandre Capone <alexandre.capone@tum.de>.

Figure 1. Overconfident uniform error bounds. The solid black line represents the unknown function, the dash-dotted line the GP mean. Crosses correspond to measurement data, the pink shaded area represents the estimated regression error. The estimated regression error increases too slowly away from the data due to the high lengthscale value and low signal variance. As a result, most of $f(x)$ in regions with no data is not captured by the estimated regression error.

2019; Maddalena et al., 2021; Sun et al., 2021). Due to this characteristic, GPs have become particularly interesting for safety-critical settings, i.e., whenever safety or performance constraints need to be considered during decision-making. Some examples include controller tuning (Capone & Hirche, 2019; Lederer et al., 2020), estimating safe operating regions (Berkenkamp et al., 2017), as well as recommendation engines (Sui et al., 2015).

Uniform error bounds for Gaussian process regression are typically obtained by scaling the posterior standard deviation, where the scaling factor depends on the number of data, the input domain, as well as regularity assumptions on the unknown function (Srinivas et al., 2012; Chowdhury & Gopalan, 2017; Maddalena et al., 2021; Lederer et al., 2019). The resulting error bound is generally of a probabilistic nature, i.e., it holds with high probability, with few exceptions (Maddalena et al., 2021; Wu & Schaback, 1993). In Lederer et al. (2019), an error bound is derived under the assumption that the true function is drawn from a GP. The bound is achieved by estimating the error on a finite grid over the input space, and then extending the bound to the whole input space using Lipschitz properties of the GP. In Srinivas et al. (2012); Chowdhury & Gopalan (2017);

Maddalena et al. (2021), error bounds are derived under the assumption that the underlying function belongs to a reproducing kernel Hilbert space. Additionally, error bounds from methods related to GPs can be directly exploited under weak restrictions. Since radial basis function interpolation yields regressors that are identical to the GP posterior mean for noise-free training data (Kanagawa et al., 2018), deterministic error bounds can be proven using the theory of reproducing kernel Hilbert spaces (Wu & Schaback, 1993; Wendland, 2004). Moreover, error bounds from regularized kernel regression, as derived in Mendelson (2002); Shi (2013); Dicker et al. (2017), can be straightforwardly extended to GPs due to the equivalence of regularized kernel and GP regression under weak assumptions (Rasmussen & Williams, 2006).

While using the posterior variance to estimate the model error can prove efficient in practice, this type of approach hinges on a critical assumption that seldom holds in practice, namely that the GP hyperparameters have been specified accurately. This is because the error bounds are derived by using smoothness assumptions about the GP kernel to bound the model error. Hence, if the choice of hyperparameters is poor, then the posterior variance is typically not a good measure of the model error. This potentially leads to overconfident error bounds, rendering the model inapplicable in safety-critical settings, as illustrated in Figure 1.

Related work. This paper addresses GP uniform error bounds under misspecified kernel hyperparameters. To mitigate misspecified error bounds and guarantee no regret in a Bayesian optimization setting, Berkenkamp et al. (2019) developed an approach that gradually decreases the length-scales of a GP, yielding convergence of the optimization algorithm towards the maximum of an unknown function. However, this is not useful for safety-critical settings, where the posterior variance, which varies strongly with the hyperparameters, is used to estimate the model error and determine the risk associated with decisions. For Matérn kernels, Tuo & Wang (2020); Wang et al. (2020) have developed robust error bounds based on the fill distance under a misspecified smoothness parameter. Furthermore, Tuo & Wang (2020) provides a more general bound based on upper and lower bounds for the decay rate of the true function. Fiedler et al. (2021) have derived error bounds for settings where the norm of the difference between the kernel used for regression and that of the reproducing kernel Hilbert space (RKHS) of the true function can be bounded. Similarly, (Beckers et al., 2018) provides an error bound when choosing a covariance function from a predefined set, under the assumption that the unknown function corresponds to a GP with a kernel from the same set. A drawback of these approaches is that they do not provide a principled or intuitive approach for obtaining hyperparameter bounds or candidate kernels.

Our contribution. In this work, we mitigate the risk of making a poor choice of hyperparameters by equipping a Gaussian process with an error bound that accounts for the lack of prior knowledge with regard to the hyperparameters. We additionally present a principled way of choosing robust uniform error bounds in a Bayesian setting without requiring any prior upper and lower bounds for the hyperparameters. The corresponding theoretical guarantees hold for a class of frequently used kernels, extending the applicability of Gaussian processes in safety-critical settings.

2. Gaussian Processes

In this section, we briefly review GPs and then discuss how error bounds and hyperparameters are chosen in practice.

We use GPs for regression, where we aim to infer an unknown function $f : \mathcal{X} \to \mathbb{R}$ with $\mathcal{X} \subseteq \mathbb{R}^d$. To this end, we treat function values $f(\boldsymbol{x})$ as random variables, of which any subset is jointly normally distributed (Rasmussen & Williams, 2006). A Gaussian process is fully specified by a mean function $m : \mathcal{X} \to \mathbb{R}$ and a positive-definite kernel $k_{\boldsymbol{\vartheta}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. In this paper, we set $m = 0$ without loss of generality, and restrict ourselves to stationary kernels with radially non-increasing Fourier transform. This is specified by the following assumption.

Assumption 2.1. The kernel $k_{\boldsymbol{\vartheta}}(\cdot, \cdot)$ is of the form

$$k_{\boldsymbol{\vartheta}}(\boldsymbol{x}, \boldsymbol{x}') \coloneqq k\left(\left(\frac{x_1 - x_1'}{\vartheta_1}, \ldots, \frac{x_d - x_d'}{\vartheta_d} \right)^{\top} \right),$$

for some $k : \mathbb{R}^d \to \mathbb{R}$ with $k(\boldsymbol{0}) = 1$. Furthermore, $k(\cdot)$ has Fourier transform $\hat{k}(\boldsymbol{\omega}) \coloneqq \kappa(\|\boldsymbol{\omega}\|_2)$ for some non-increasing non-negative function $\kappa : \mathbb{R}_+ \to \mathbb{R}_+$.

Here $\boldsymbol{\vartheta} \coloneqq (\vartheta_1, \ldots, \vartheta_d)^{\top} \in \boldsymbol{\Theta}$ denotes the kernel length-scales, which scale the kernel inputs and specify the variability of the underlying function. A similar assumption can be found in other papers that analyze algorithms based on stationary kernels (Bull, 2011; Berkenkamp et al., 2019). When using this type of kernel, the similarity between any two function evaluations decreases with the weighted distance between the corresponding inputs. Many frequently encountered kernels satisfy these properties. Examples include the Gaussian and Matérn kernels, which are often employed because they satisfy the universal approximation property (Micchelli et al., 2006). Note that we assume $k(\boldsymbol{0}, \boldsymbol{0}) = 1$ only for the sake of simplicity, and all the results and tools presented in this paper can be straightforwardly extended to the more general case where the scaling factor of $k(\cdot, \cdot)$, i.e., the signal variance, is also a hyperparameter. This is discussed in Section 3.1.

Given a set of N (potentially noisy) measurements $\mathcal{D} = \{\boldsymbol{x}_i, y_i\}_{i=1}^N \coloneqq \{\boldsymbol{X}, \boldsymbol{y}\}$, where $y_i = f(\boldsymbol{x}_i) + \varepsilon_i$, $\varepsilon_i \sim$

$\mathcal{N}(0, \sigma_n^2)$ and $\boldsymbol{x}_i \in \mathcal{X}$, we are able to condition a Gaussian process on $\mathcal{D}$ to obtain the posterior distribution of $f(\boldsymbol{x}^*)$ at an arbitrary test point $\boldsymbol{x}^* \in \mathcal{X}$. The corresponding distribution is normal $f(\boldsymbol{x}^*) \sim \mathcal{N}(\mu_{\boldsymbol{\vartheta}}(\boldsymbol{x}^*), \sigma_{\boldsymbol{\vartheta}}^2(\boldsymbol{x}^*))$ with mean and variance

$$\mu_{\boldsymbol{\vartheta}}(\boldsymbol{x}^*) := \boldsymbol{k}_{\boldsymbol{\vartheta}}(\boldsymbol{x}^*)^\top \left(\boldsymbol{K}_{\boldsymbol{\vartheta}} + \sigma_n^2 \boldsymbol{I} \right)^{-1} \boldsymbol{y}$$

$$\sigma_{\boldsymbol{\vartheta}}^2(\boldsymbol{x}^*) := k_{\boldsymbol{\vartheta}}(\boldsymbol{x}^*, \boldsymbol{x}^*) - \boldsymbol{k}_{\boldsymbol{\vartheta}}(\boldsymbol{x}^*)^\top \left(\boldsymbol{K}_{\boldsymbol{\vartheta}} + \sigma_n^2 \boldsymbol{I} \right)^{-1} \boldsymbol{k}_{\boldsymbol{\vartheta}}(\boldsymbol{x}^*),$$

where $\boldsymbol{k}_{\boldsymbol{\vartheta}}(\boldsymbol{x}^*) = (k_{\boldsymbol{\vartheta}}(\boldsymbol{x}_1, \boldsymbol{x}^*), \ldots, k_{\boldsymbol{\vartheta}}(\boldsymbol{x}_N, \boldsymbol{x}^*))^\top$, σ_n^2 denotes the noise variance, and the entries of the covariance matrix $\boldsymbol{K}_{\boldsymbol{\vartheta}}$ are given by $(\boldsymbol{K}_{\boldsymbol{\vartheta}})_{i,j} = k_{\boldsymbol{\vartheta}}(\boldsymbol{x}_i, \boldsymbol{x}_j)$. In the case of stationary kernels, the posterior variance $\sigma_{\boldsymbol{\vartheta}}^2(\boldsymbol{x}^*)$ is typically low for test inputs $\boldsymbol{x}^*$ that are close to the measurement data $\mathcal{D}$ and vice versa.

2.1. Choosing Hyperparameters

By far the most common technique for choosing the hyperparameters of a Gaussian process is maximizing the log marginal likelihood of the measurements given the hyperparameters

$$\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\vartheta}) = -\frac{1}{2}\log|\boldsymbol{K}_{\boldsymbol{\vartheta}} + \sigma_n^2\boldsymbol{I}| - \frac{N}{2}\log(2\pi) \\ -\frac{1}{2}\boldsymbol{y}^\top \left(\boldsymbol{K}_{\boldsymbol{\vartheta}} + \sigma_n^2\boldsymbol{I} \right)^{-1} \boldsymbol{y}. \tag{1}$$

By maximizing (1), we obtain a trade-off between model complexity and data fit. Coupled with the fact that the gradient of (1) can be computed analytically, choosing hyperparameters in this manner yields many practical benefits. Furthermore, if the marginal likelihood is well peaked, then the true hyperparameters are likely to be situated near the selected ones. However, if this is not the case, then this approach can lead to overconfident hyperparameters, as the marginal likelihood can decrease slowly away from the maximum, implying that the lengthscales are potentially smaller than the ones obtained. This type of behavior is particularly frequent if little data has been observed, as both short and long lengthscales explain the data consistently (Rasmussen & Williams, 2006). Less common approaches for choosing the Gaussian process hyperparameters include the cross validation (Cressie, 2015) and log-pseudo likelihood maximization (Sundararajan & Keerthi, 2001). In some settings, the hyperparameters are chosen based on prior knowledge about the system (Kirschner et al., 2019).

3. Uniform Error Bounds for Unknown Hyperparameters

We now introduce a modified version of standard GP error bounds that aims to overcome the limitations mentioned in the previous chapter. The proofs of all results stated here can be found in the appendix.
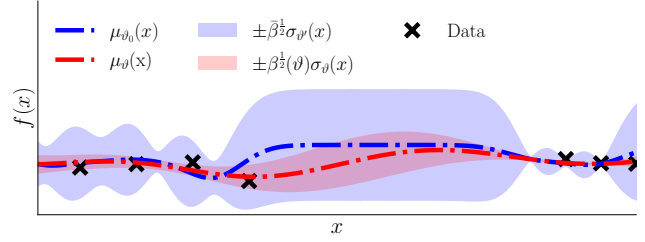

Figure 2. The uniform error bound obtained with bounding vectors $\boldsymbol{\vartheta}', \boldsymbol{\vartheta}''$ for a posterior mean with working hyperparameters $\boldsymbol{\vartheta}_0$, in blue/light blue, fully contains that of the Gaussian process with lengthscales $\boldsymbol{\vartheta}$, shown in red/light red, provided that $\boldsymbol{\vartheta}' \leq \boldsymbol{\vartheta}, \boldsymbol{\vartheta}_0 \leq \boldsymbol{\vartheta}''$, where $\leq$ denotes component-wise inequality. This yields Theorem 3.5.

In the Gaussian process regression literature, there are typically two different types of assumptions that can be made for analysis. On the one hand, the Bayesian case can be considered, where we assume the unknown function $f(\cdot)$ to be sampled from a Gaussian process (Sun et al., 2021). On the other hand, the frequentist setting can be considered, where $f(\cdot)$ is assumed to have a bounded RKHS norm (Srinivas et al., 2012; Chowdhury & Gopalan, 2017) with respect to the kernel $k_{\boldsymbol{\vartheta}}(\cdot, \cdot)$. While some of the techniques presented in this paper can be useful both in a Bayesian and frequentist setting, the Bayesian paradigm provides a straightforward way of narrowing the set of candidate hyperparameters used to derive an error bound, whereas no such path is evident in the frequentist setting. Hence, we henceforth assume a Bayesian setting, as described in the following assumption.

Assumption 3.1 (Bayesian setting). The unknown function $f(\cdot)$ corresponds to a sample from a Gaussian process with kernel $k_{\boldsymbol{\vartheta}}(\cdot, \cdot)$, where the hyperparameters $\boldsymbol{\vartheta}$ are drawn from a hyperprior $p(\boldsymbol{\vartheta})$.

This assumption is not very restrictive, and is often encountered in control and reinforcement learning settings (Deisenroth et al., 2015; Kocijan et al., 2005; Hewing et al., 2019). The choice of prior $p(\boldsymbol{\vartheta})$ can be based on any prior beliefs about the underlying function, e.g., Lipschitz continuity, which can be encoded into the prior using a chi-squared or uniform distribution.

We also assume to have a scaling function that specifies a uniform error bound for an arbitrary fixed vector of hyperparameters $\boldsymbol{\vartheta}$ with a prespecified probability $1 - \rho$.

Assumption 3.2 (Scaling function). For an arbitrary ρ with $0 < \rho < 1$, there exists a known positive function $\beta : \mathbb{R}^d \to \mathbb{R}_+$, such that for any vector of lengthscales $\boldsymbol{\vartheta} \in \boldsymbol{\Theta}$,

$$\mathrm{P}\Big(|f(\boldsymbol{x}) - \mu_{\boldsymbol{\vartheta}}(\boldsymbol{x})| \leq \beta^{\frac{1}{2}}(\boldsymbol{\vartheta})\sigma_{\boldsymbol{\vartheta}}(\boldsymbol{x}) \ \forall \boldsymbol{x} \in \mathcal{X}\Big) \geq 1 - \rho$$

holds, where $f(\boldsymbol{x})$ denotes a sample from a Gaussian process with prior mean zero and kernel $k_{\boldsymbol{\vartheta}}(\cdot, \cdot)$, conditioned on measurement data $\mathcal{D}$.

Error bounds for fixed hyperparameters $\boldsymbol{\vartheta}$ in the Bayesian case can be derived by assuming that the input space is compact and that the kernel satisfies some regularity requirements (Srinivas et al., 2012). Improved error bounds can be obtained if we are only interested in samples $f(\boldsymbol{x})$ that satisfy a predefined Lipschitz continuity requirement (Lederer et al., 2019; Sun et al., 2021). Additionally, if we assume the input space to be finite, then $\beta(\cdot)$ is a constant, i.e., does not depend on $\boldsymbol{\vartheta}$ (Srinivas et al., 2012).

Our approach is then based on the following result.

Lemma 3.3. *Let $k(\cdot)$ be a kernel, $\boldsymbol{\vartheta}', \boldsymbol{\vartheta}'', \boldsymbol{\vartheta} \in \boldsymbol{\Theta}$ vectors of lengthscales with $\boldsymbol{\vartheta}' \leq \boldsymbol{\vartheta} \leq \boldsymbol{\vartheta}''$, where $\leq$ denotes component-wise inequality, and let $\mathcal{D}$ denote a measurement data set of an unknown function $f(\cdot)$. Furthermore, choose $\gamma > 0$ with $\gamma^2 = \prod_{i=1}^d \frac{\vartheta_i''}{\vartheta_i'}$. Then*

$$\sigma_{\boldsymbol{\vartheta}}(\boldsymbol{x}) \leq \gamma \sigma_{\boldsymbol{\vartheta}'}(\boldsymbol{x})$$

holds for all $\boldsymbol{x} \in \mathcal{X}$.

Remark 3.4. Lemma 3.3 is very general, which might lead to conservative scaling factors γ. In some settings, however, the inequality in Lemma 3.3 also holds for smaller values of γ. This is the case, e.g., for the commonly employed squared-exponential kernel, as illustrated in Section 5.

In essence, Lemma 3.3 states that the scaled posterior variance decreases with the lengthscales. As a direct consequence, if a pair of bounding hyperparameters $\boldsymbol{\vartheta}', \boldsymbol{\vartheta}''$ is available, we can use Lemma 3.3 to construct an error bound that contains all error bounds corresponding to lengthscales within then interval $\boldsymbol{\vartheta}' \leq \boldsymbol{\vartheta} \leq \boldsymbol{\vartheta}''$. This is specified in the following result.

Theorem 3.5. *Let Assumption 3.2 and Assumption 2.1 hold, and let ρ and $\beta(\cdot)$ be as in Assumption 3.2. Let $\boldsymbol{\vartheta}', \boldsymbol{\vartheta}'', \boldsymbol{\vartheta} \in \boldsymbol{\Theta}$ be vectors of lengthscales with $\boldsymbol{\vartheta}' \leq \boldsymbol{\vartheta} \leq \boldsymbol{\vartheta}''$, where $\leq$ denotes component-wise inequality, let $\mathcal{D}$ be a measurement data set, and let $f(\boldsymbol{x})$ be a sample from a Gaussian process with mean zero and kernel $k_{\boldsymbol{\vartheta}}(\cdot, \cdot)$, conditioned on $\mathcal{D}$. Furthermore, let $\mu_{\boldsymbol{\vartheta}_0}(\cdot)$ denote the posterior mean for arbitrary lengthscales $\boldsymbol{\vartheta}_0$ with $\boldsymbol{\vartheta}' \leq \boldsymbol{\vartheta}_0 \leq \boldsymbol{\vartheta}''$, and define*

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

It then holds that

$$P\left(|f(\boldsymbol{x}) - \mu_{\boldsymbol{\vartheta}_0}(\boldsymbol{x})| \leq \bar{\beta}^{\frac{1}{2}} \sigma_{\boldsymbol{\vartheta}'}(\boldsymbol{x}) \quad \forall \boldsymbol{x} \in \mathcal{X} \right) \geq 1 - \rho.$$

Theorem 3.5 implies that we can obtain a uniform error bound using the scaled posterior variance corresponding to the smallest vector of lengthscales $\boldsymbol{\vartheta}'$.

Remark 3.6. The term $2\sigma_n^{-1}\|\boldsymbol{y}\|_2$ in (2) originates from the discrepancy $|\mu_{\boldsymbol{\vartheta}}(\boldsymbol{x}) - \mu_{\boldsymbol{\vartheta}_0}(\boldsymbol{x})|$ between the posterior means using the working lengthscales $\boldsymbol{\vartheta}_0$ and $\boldsymbol{\vartheta}$ within the bounding lengthscales $\boldsymbol{\vartheta}' \leq \boldsymbol{\vartheta} \leq \boldsymbol{\vartheta}''$. In practice, if this discrepancy is determined to be small, e.g., by evaluating the discrepancy $|\mu_{\boldsymbol{\vartheta}}(\boldsymbol{x}) - \mu_{\boldsymbol{\vartheta}_0}(\boldsymbol{x})|$ for different $\boldsymbol{\vartheta}$ and test inputs $\boldsymbol{x}$, then $2\sigma_n^{-1}\|\boldsymbol{y}\|_2$ can be substituted by smaller values.

Note that the probabilistic nature of Theorem 3.5 stems from Assumption 3.2 also being probabilistic, and no additional uncertainty is introduced when deriving Theorem 3.5. In other words, the confidence region generated by the robust uniform error bound fully contains all confidence regions corresponding to $\boldsymbol{\vartheta}$ with $\boldsymbol{\vartheta}' \leq \boldsymbol{\vartheta} \leq \boldsymbol{\vartheta}''$. This is illustrated in Figure 2.

In order to be able to apply Theorem 3.5 in the more general setting where the bounding vectors $\boldsymbol{\vartheta}', \boldsymbol{\vartheta}''$ are not given a priori, the next step is to determine permissible $\boldsymbol{\vartheta}', \boldsymbol{\vartheta}''$. Since we are in a Bayesian scenario, an intuitive approach is to choose the bounding lengthscales $\boldsymbol{\vartheta}', \boldsymbol{\vartheta}''$ such that they form a $1 - \delta$ confidence interval for some $\delta \in (0, 1)$. Note that deriving a similar technique in the frequentist setting, e.g., as in Chowdhury & Gopalan (2017) or Srinivas et al. (2012), is not straightforward since the corresponding tools are not directly compatible with a distribution over the hyperparameters.

Formally, our approach consists of obtaining a pair of hyperparameters $\boldsymbol{\vartheta}', \boldsymbol{\vartheta}''$ that lies within the set

$$\mathcal{P}_\delta = \left\{ (\boldsymbol{\vartheta}', \boldsymbol{\vartheta}'') \in \boldsymbol{\Theta}^2 \, \middle| \, \int_{\boldsymbol{\vartheta}' \leq \boldsymbol{\vartheta} \leq \boldsymbol{\vartheta}''} p(\boldsymbol{\vartheta}|\mathcal{D})d\boldsymbol{\vartheta} \geq 1 - \delta \right\},$$
$$(3)$$

where the posterior over the lengthscales $\boldsymbol{\vartheta}$ given the data $\mathcal{D} = \{\boldsymbol{X}, \boldsymbol{y}\}$ is computed using Bayes' rule as

$$p(\boldsymbol{\vartheta}|\boldsymbol{y}, \boldsymbol{X}) = \frac{p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\vartheta})p(\boldsymbol{\vartheta})}{p(\boldsymbol{y}|\boldsymbol{X})}, \quad (4)$$

and the posterior $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\vartheta})$ is computed similarly to (1). The normalizing factor $p(\boldsymbol{y}|\boldsymbol{X})$ is then given by

$$p(\boldsymbol{y}|\boldsymbol{X}) = \int_{\boldsymbol{\Theta}} p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\vartheta})p(\boldsymbol{\vartheta})d\boldsymbol{\vartheta}. \quad (5)$$

By choosing the bounding lengthscales $\boldsymbol{\vartheta}', \boldsymbol{\vartheta}''$ in this fashion, a vector of lengthscales $\boldsymbol{\vartheta}$ sampled from the posterior distribution $p(\boldsymbol{\vartheta}|\mathcal{D})$ lies within the interval $\boldsymbol{\vartheta}' \leq \boldsymbol{\vartheta} \leq \boldsymbol{\vartheta}''$ with high probability. We are then able to estimate the error in the setting with unknown hyperparameters by applying Theorem 3.5, obtaining the following result.

Theorem 3.7. *Let Assumption 3.2 and Assumption 2.1 hold. Let $\boldsymbol{\vartheta}', \boldsymbol{\vartheta}_0, \boldsymbol{\vartheta}'' \in \boldsymbol{\Theta}$ be vectors of lengthscales with*
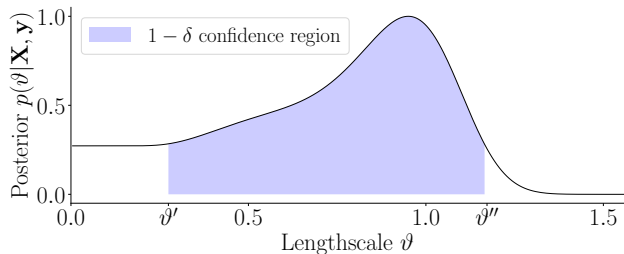
Figure 3. Posterior probability distribution $p(\vartheta|\boldsymbol{X}, \boldsymbol{y})$ given data $\mathcal{D} = \{\boldsymbol{X}, \boldsymbol{y}\}$ and confidence region of $1 - \delta$ generated by bounding lengthscales $(\vartheta', \vartheta'')$. By applying Theorem 3.5, we can use the pair $(\vartheta', \vartheta'')$ to obtain a robust error bound. This yields Theorem 3.7. Note that the posterior is poorly peaked, which indicates that the lengthscales obtained via log-likelihood maximization are potentially too high, leading to overconfident error bounds.

$(\boldsymbol{\vartheta}', \boldsymbol{\vartheta}'') \in \mathcal{P}_\delta$ *and choose* $\bar{\beta}$ *as in Theorem* 3.5. *Furthermore, let* $\boldsymbol{\vartheta} \sim p(\boldsymbol{\vartheta}|\mathcal{D})$, *and let* $f(\cdot)$ *denote a sample from a Gaussian process with kernel* $k_{\boldsymbol{\vartheta}}(\cdot, \cdot)$ *as specified by Assumption* 3.1. *Then*

$$|f(\boldsymbol{x}) - \mu_{\boldsymbol{\vartheta}_0}(\boldsymbol{x})| \le \bar{\beta}^{\frac{1}{2}} \sigma_{\boldsymbol{\vartheta}'}(\boldsymbol{x}) \tag{6}$$

holds for all $\boldsymbol{x} \in \mathcal{X}$ *with probability* $(1 - \delta)(1 - \rho)$.

3.1. Extension to Noise and Signal Variance

So far, we discussed the setting where the signal and noise variance are constant and known. However, the proposed tools extend straightforwardly to the setting where they are also unknown, provided that a corresponding prior distribution is also available. This is because the posterior variance $\sigma_{\boldsymbol{\vartheta}}^2(\cdot)$ increases with the noise and signal variance, hence a result of the form of Theorem 3.7 can be easily obtained.

3.2. Choosing Bounding Hyperparameters

Typically, the set of bounding pairs $\mathcal{P}_\delta$ contains more than one element. Hence, Theorem 3.7 allows us some flexibility when deriving the uniform error bound, since we can choose any pair $(\boldsymbol{\vartheta}', \boldsymbol{\vartheta}'')$ that lies within $\mathcal{P}_\delta$. In the following, we derive a heuristic that aims to approximate the smallest possible $1 - \delta$ confidence region around the working hyperparameters $\boldsymbol{\vartheta}_0$, which would automatically meet the requirements to apply Theorem 3.7. Formally, this is achieved by solving the optimization problem

$$\min_{\boldsymbol{\vartheta}', \boldsymbol{\vartheta}'' \in \boldsymbol{\Theta}} \quad \|\boldsymbol{\vartheta}'' - \boldsymbol{\vartheta}'\|_2$$
$$\text{s.t.} \int_{\boldsymbol{\vartheta}' \le \boldsymbol{\vartheta} \le \boldsymbol{\vartheta}''} p(\boldsymbol{\vartheta}|\mathcal{D}) d\boldsymbol{\vartheta} \ge 1 - \delta. \tag{7}$$

In other words, we aim to choose the pair of bounding hyperparameters $(\boldsymbol{\vartheta}', \boldsymbol{\vartheta}'') \in \mathcal{P}_\delta$ that deviate the least from $\boldsymbol{\vartheta}_0$.

3.3. Discussion

Although the integrals in Equation (3) and (5) generally cannot be computed analytically, we can resort to different approximations. In low-dimensional settings, we can solve the integral expression using numerical integration. A further option is to employ Markov chain Monte Carlo (MCMC) methods (Rasmussen & Williams, 2006). Alternatively, we can apply approximate inference methods, such as Laplace's method or expectation propagation, which aim to approximate the posterior $p(\boldsymbol{\vartheta}|\boldsymbol{y}, \boldsymbol{X})$ with a normal distribution. The latter methods are particularly well suited for settings with a high number of data points, as they do not require a high number of evaluations of the posterior distribution. Note that this is an advantage of the proposed technique compared to fully Bayesian GPs, which typically require some form of MCMC approach. Moreover, after computing the bounding hyperparameters $\boldsymbol{\vartheta}', \boldsymbol{\vartheta}''$, the computational complexity of evaluating the error bound is only twice that of a standard GP, whereas in a fully Bayesian GP a potentially high number of GPs has to be evaluated per prediction. A further advantage compared to fully Bayesian GPs is that, in practice, it is reasonable to expect better error estimates, particularly if we choose a low risk parameter δ and a sufficiently smooth prior $p(\boldsymbol{\vartheta})$. This is because the marginal likelihood is typically well behaved in the hyperparameter space, and the fully Bayesian posterior mean and variance will lie within the uniform error bound computed with Theorem 3.7. This is supported by the experimental results in Section 5. We note, however, that the proposed approach can also be combined with a fully Bayesian GP, i.e., by using the fully Bayesian posterior mean only for regression, and the presented techniques to bound the regression error.

It is also worth noting that there is no direct analogy between the approach presented in this paper and robust uniform error bounds in the frequentist setting, i.e., where the unknown function $f(\cdot)$ is assumed to be fixed with a bounded RKHS norm. We believe this to be a strong argument in favor of employing a Bayesian perspective instead of a frequentist one when performing regression.

4. Control with Performance Guarantees

Theorem 3.7 can be employed straightforwardly to derive safety guarantees in different settings. In the following, we show how to apply it in a learning-based control setting with a commonly encountered system structure.

4.1. Control Problem

We consider an m-dimensional dynamical system that obeys the frequently encountered strict-feedback form (Sabanovic et al., 1993; Krstic et al., 1995b; Kwan & Lewis, 2000):

$$
\begin{aligned}
\dot{x}_1 &= f_1(x_1) + g(x_1)x_2 \\
\dot{x}_2 &= f_2(x_1, x_2) + g(x_1, x_2)x_3 \\
&\vdots \\
\dot{x}_m &= f_m(x_1, \ldots, x_m) + g(x_1, \ldots, x_m)u,
\end{aligned}
\tag{8}
$$

where $x_1, \ldots, x_m \in \mathbb{R}$ and $u \in \mathbb{R}$ denote the system's states and control input, respectively. The functions $f_i : \mathbb{R}^i \to \mathbb{R}$ are unknown and modeled using GPs, whereas we assume to know $g_i : \mathbb{R}^i \to \mathbb{R}$. These assumptions are common in this setting (Kwan & Lewis, 2000; Capone & Hirche, 2019).

Our goal is to design a control law u that steers the subsystem $\dot{x}_1 = f_1(x_1) + g(x_1)x_2$ towards a desired time-dependent trajectory $x_d(t)$. In other words, we aim to reduce the norm of the time-dependent error $e_1(t) = x_1(t) - x_d(t)$. In the following, we assume that $x_d(t)$ is m times continuously differentiable, and that all its derivatives are bounded, which is not a restrictive assumption. For the sake of simplicity, we henceforth use the notation $f_i := f_i(x_1, \ldots, x_i)$ and $g_i := g_i(x_1, \ldots, x_i)$.

In order to provide performance guarantees for (8), we require some additional smoothness assumptions with respect to the kernels used to model the functions f_i. This is expressed formally in the following.

Assumption 4.1. Each function f_i is drawn from a Gaussian process with $m - i$ times differentiable kernel $k_{\vartheta_i}(\cdot, \cdot)$.

Assumption 4.1 implies that the functions f_i are $m - i$ times differentiable, which is not a restrictive assumption, as it applies for many systems of the form given by (8), e.g., robotic manipulators, jet engines, and induction motors (Kwan & Lewis, 2000).

4.2. Backstepping Control

In order to track the desired trajectory $x_d(t)$, we employ a backstepping technique similar to the one proposed in Capone & Hirche (2019). The idea behind backstepping is to recursively design fictitious control inputs $x_{i+1,d}$ for each subsystem $\dot{x}_i = f_i + g_i x_{i+1,d}$, which we then aim to track by means of the true control input u. For more details, the reader is referred to (Krstic et al., 1995a).

To model the unknown functions $f_i(\cdot)$, we assume to have N noisy measurements $\{x_{1 \ldots i, j}, f_{i,j}\}_{j=1}^N$. The control input

u can then be computed recursively using

$$
\begin{aligned}
x_{2,d} &= g_1^{-1}\left(-\mu_1 + \dot{x}_d - C_1 e_1\right) \\
&\vdots \\
x_{i,d} &= g_{i-1}^{-1}\left(-\mu_{i-1} + \dot{x}_{i-1,d} - C_{i-2}e_{i-2}\right) \\
&\vdots \\
u &= g_m^{-1}\left(-\mu_m + \dot{x}_{m,d} - e_m - g_{m-1}e_{m-1}\right),
\end{aligned}
\tag{9}
$$

where μ_i denotes the posterior mean of the i-th Gaussian process, $e_i = x_i - x_{i,d}$ the tracking error of the i-th subsystem, and $C_i : \mathbb{R}^i \to \mathbb{R}$ are state-dependent control gains.

The stability analysis of the closed-loop system is carried out using Lyapunov theory. By using a quadratic Lyapunov function, it is straightforward to show that there exists an ultimate upper bound ξ for the tracking error $e_d(t)$, i.e., $e_d(t) \le \xi$ holds for large enough t, where $\xi > 0$ depends on the control gains C_i (Capone & Hirche, 2019, Lemma 2). By using this result and Theorem 3.7, we are able to determine adaptive control gains C_i that yield a tracking error smaller or equal than a predetermined value ξ_{des}, which is crucial in safety-critical applications. This is formally stated in the following result.

Theorem 4.2. *Let Assumption 2.1 and Assumption 4.1 hold, and let the control input u be given by (9). Furthermore, let $(\vartheta_1', \vartheta_1''), \ldots, (\vartheta_m', \vartheta_m'')$ be bounding hyperparameters for the subsystems $1, \ldots, m$, obtained with (7) and risk parameter $\delta \in (0, 1)$, and let $\beta_i(\cdot)$ be the corresponding scaling functions. Choose $\xi_{des} > 0$ and let $\bar{\beta}_i$ be as in Theorem 3.5, and choose each state-dependent control gain as*

$$
C_i(\boldsymbol{x}) := \frac{1}{\xi_{des}}\sqrt{\sum_{j=1}^m \bar{\beta}_j \sigma_{\vartheta_j'}^2(\boldsymbol{x})},
\tag{10}
$$

where γ_j are chosen as in Lemma 3.3. Then, with probability $(1-\rho)^m(1-\delta)^m$, there exists a $T > 0$, such that

$$
\|\boldsymbol{x}(t) - \boldsymbol{x}_d(t)\|_2 \le \xi_{des}
\tag{11}
$$

holds for all $t > T$.

Hence, we can use (10) to obtain the control gains required for the control performance specification ξ_{des}.

5. Results

We now present experimental results where the performance of the proposed error bound is compared to that of vanilla and fully Bayesian GPs[1]. We first showcase the prediction error on regression benchmarks, then apply the proposed

---

[1]The corresponding code can be found at https://github.com/aCapone1/gauss_proc_unknown_hyp.

(a) $N = 2$ data points.

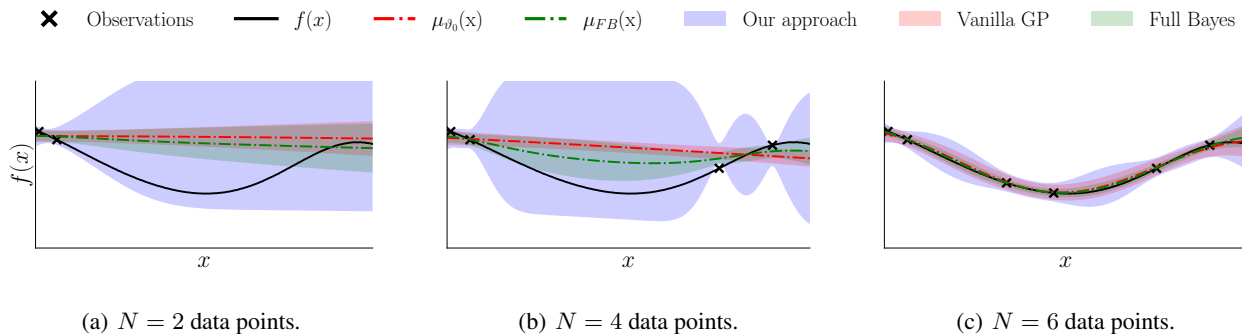(b) $N = 4$ data points.

(c) $N = 6$ data points.

Figure 4. Function $f(\cdot)$ sampled from a GP. Mean and estimated regression error when using our approach (blue), a vanilla GP (red), and a fully Bayesian GP (green). We used scaling parameters of $\beta = \bar{\beta} = 2$. Both the vanilla GP and the fully Bayesian GP fail to correctly estimate most of the regression error unless the input space is sufficiently covered by the data.

DATA SET	BSTN$_{50}$	BSTN$_{450}$	ML$_{50}$	ML$_{300}$	WINE$_{200}$	WINE$_{1000}$	SRCS$_{850}$	SRCS$_{5000}$	SRCS$_{10000}$
DIMENSION	$d = 13$	$d = 13$	$d = 1$	$d = 1$	$d = 11$	$d = 11$	$d = 21$	$d = 21$	$d = 21$
OUR APPROACH	**0.19**	**0.35**	**0.00**	**0.00**	**0.01**	**0.01**	**0.08**	**0.02**	**0.01**
VANILLA GP	0.41	0.48	0.11	0.01	0.04	0.04	0.98	0.98	0.97
FULL BAYES	0.36	0.44	**0.00**	**0.00**	0.04	0.04	0.76	-	-

Table 1. Average rate of error bound violations using the proposed approach, vanilla GPs, and fully Bayesian GPs. Lower is better. We set $\bar{\beta} = \beta = 2$ to avoid overly conservative error estimates while enabling a fair comparison with high practical relevance. BSTN stands for Boston (house prices), ML for Mauna Loa, and SRCS for Sarcos. The subscripts in the data set names indicate the number of training data points used. For the two largest Sarcos data sets, the number of samples required to get reliable fully Bayesian models is prohibitive, hence we do not provide corresponding predictions. Best-performing methods are in boldface. Our approach performs best across all scenarios.

technique to design a learning-based control law. A Gaussian kernel is used in all cases except the Mauna Loa experiments, and we also consider uncertainty in the signal and noise variances, as discussed in Section 3. In all except the Sarcos experiments, the fully Bayesian GPs and integral expression in (7) are approximated using the No-U Turn Sampler algorithm for MCMC (Hoffman et al., 2014). The implementations are carried out using GPyTorch (Gardner et al., 2018).

The values for β proposed in theory are typically very conservative (Srinivas et al., 2012), which makes their use impractical. Similarly, the scaling factor $\bar{\beta}$ proposed by Theorem 3.7 assumes very high values. In many safety-critical applications it is common to choose the scaling factor as $\beta = 2$, independently of the hyperparameters ϑ (Berkenkamp et al., 2017; Umlauft et al., 2017). Moreover, empirical results indicate that Lemma 3.3 holds with $\gamma = 1$ in the case of a Gaussian kernel, i.e., the error bound will be at least as robust as that of the vanilla GP for $\beta = \bar{\beta}$. Hence, to obtain a fair, practically relevant and not overly conservative comparison, we set $\bar{\beta} = \beta = 2$ in the following experiments. Note that we still apply the results from Theorem 3.7 to compute the posterior variance $\sigma_{\vartheta'}(\cdot)$. The

working hyperparameters ϑ_0 are chosen by maximizing the log marginal likelihood (1). We employ a confidence parameter of $\delta = 0.05$ for (7) and assume uniform distributions as hyperpriors $p(\vartheta)$. Note that even though this imposes hard bounds on the hyperparameters, these are still very large.

5.1. Regression - Toy Problem and Benchmarks

In the regression experiments, to additionally illustrate the behavior of the proposed technique as more data becomes available, we train the GPs with data sets of varying sizes N.

We first investigate how the proposed technique performs when estimating the regression error of a function $f(\cdot)$ that is sampled from a GP. The results can be seen in Figure 4. Our approach always captures the behavior of the underlying sample $f(\cdot)$. Both the vanilla GP error bound and the fully Bayesian error bound fail to do so if little data is available (Figure 4(a)) or if the data is sparse (Figure 4(b)). The estimated error is only accurate for the vanilla and fully Bayesian GPs if the data covers the state space sufficiently well (Figure 4(c)).

We now apply the proposed technique to estimate the regression error of the Boston house prices data set (Pedregosa

et al., 2011), the UCI wine quality data set (Cortez et al., 2009), the Mauna Loa CO_2 time series, and the Sarcos data set. For the Mauna Loa data set, we employ a spectral mixture kernel with 20 mixtures (Wilson & Adams, 2013), which consists of a sum of 20 Gaussian kernels multiplied with sinusoidal kernels. It is straightforward to show that its Fourier transform increases with the lengthscales, hence we are able to apply Theorem 3.7. The mixture means, which specify the frequency of the periodic components, are assumed to be fixed except for the fully Bayesian case. For the two largest Sarcos data sets, we perform a Laplace approximation around the maximum of the posterior, which yields a normal distribution (MacKay, 2002). To ensure that the corresponding covariance matrix is positive definite, we employ an empirical Bayes approach and specify a quadratic hyperprior around the log likelihood maximum. The bounding hyperparameters are then obtained by computing the corresponding rectangular confidence region as in Šidák (1967).

We run each scenario multiple times and select the N training and N_{test} test points randomly every time. The Sarcos experiments are repeated 10 times due to high computational requirements, all other experiments are repeated 100 times. To evaluate performance, we check how often the error bound is violated, which corresponds to measuring the quantity

$$\frac{1}{N_{\text{test}}} \sum_{m=1}^{N_{\text{test}}} \mathbb{I}\left(|y_m^{\text{test}} - \mu_{\boldsymbol{\vartheta}_0}(\boldsymbol{x}_m^{\text{test}})| - \beta^{\frac{1}{2}} \sigma_{\boldsymbol{\vartheta}'}(\boldsymbol{x}_m^{\text{test}}) \right), \quad (12)$$

where the superscript test denotes test inputs/outputs, and

$$\mathbb{I}(z) = \begin{cases} 0, & \text{if } z \leq 0 \\ 1, & \text{otherwise} \end{cases}$$

is the indicator function. This metric[2] is highly relevant, since the theoretical guarantees of many GP-based Bayesian optimization and safe control algorithms hinge on the assumption that (12) is equal to zero with high probability (Srinivas et al., 2012; Berkenkamp et al., 2017; Umlauft et al., 2017; Capone & Hirche, 2019; Lederer et al., 2019). In practice, most such algorithms have been shown to provide satisfactory results even if this requirement is not strictly enforced. However, it is still highly desirable for (12) to be as small as possible, as this reduces the amount of potentially unsafe choices.

The results are summarized in Table 1. As can be seen, our bound is always more accurate than that of vanilla or fully Bayesian GPs, particularly in low-data regimes. This is to be expected, as the marginal likelihood function is known to be poorly peaked for small N (MacKay, 1999).

---

[2] Kuleshov et al. (2018) employ a similar metric to determine the calibration error in regression settings.
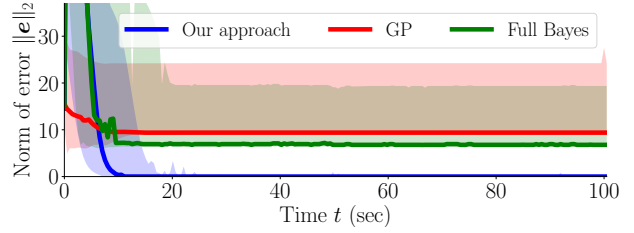

Figure 5. Tracking error norm $\|e\|_2$ over time t. Solid lines and shaded regions correspond to the median and lower/upper deciles, respectively. Blue corresponds to our method, red to vanilla GPs, green to fully Bayesian GPs. Our method yields the desired tracking error, as expected from Theorem 4.2, whereas the vanilla and fully Bayesian GPs do not.

5.2. Control Design with Little Data

We now apply the proposed technique to design a control law for a safety-critical setting, where a one-link planar manipulator with motor dynamics is to be steered towards the origin using the method presented in Section 4 and $N = 10$ training data points. Note that $N = 10$ is not very small for control purposes in the proposed setting, since good performance can already be achieved with as little as $N = 50$ data points (Capone & Hirche, 2019). The data is obtained using a low-gain sinusoidal input.

The manipulator dynamics are given by

$$D\ddot{\varphi} + B\dot{\varphi} + G\sin(\varphi) = \tau$$
$$M + \dot{\tau} + H\tau + Z\dot{\varphi} = u,$$

where φ and τ are the system's angle and torque, respectively, u is the motor voltage, which we can control directly, and D, B, G, M, H, Z are system parameters. By approximating the differential equation of each state $\varphi, \dot{\varphi}, \tau$ using a separate Gaussian process, we are able to employ a backstepping technique to track a desired trajectory, as described in Section 4. We then use Theorem 4.2 to choose the control gains $C_i(\boldsymbol{x})$, $i = 1, \ldots, 3$, by setting the desired error to $\epsilon_{\max} = 1$.

To evaluate the control performance, we run 100 simulations for each setting, where the initial conditions of the manipulator are randomly sampled from a normal distribution. The norm of the resulting tracking error is displayed in Figure 5. As can be seen, the gains obtained using the proposed approach perform considerably better than the ones obtained with vanilla and fully Bayesian GPs. This is because the posterior variance obtained with our approach increases more rapidly away from the collected data points than in the other settings, expressing less confidence in the posterior mean.

6. Conclusion

We have presented robust uniform error bounds for Gaussian processes with unknown hyperparameters. Our approach is applicable for stationary radially non-decreasing kernels, which are commonly employed in practice. It hinges on computing a confidence region for the hyperparameters, which does not require pre-specified bounds for the hyperparameters. The presented theoretical results make them flexible and easily applicable to safety-critical scenarios, where theoretical guarantees are often required. In numerical regression benchmarks, the proposed error bound was shown to outperform the error bound obtained with standard and fully Bayesian Gaussian processes. Furthermore, the presented tool resulted in better performance in a control problem, indicating better suitability for safety-critical settings.

Acknowledgements

References

Beckers, T., Umlauft, J., and Hirche, S. Mean square prediction error of misspecified Gaussian process models. In *2018 IEEE Conference on Decision and Control*, pp. 1162–1167, 2018.

Berkenkamp, F., Turchetta, M., Schoellig, A., and Krause, A. Safe model-based reinforcement learning with stability guarantees. In *Advances in Neural Information Processing Systems*, pp. 908–919, 2017.

Berkenkamp, F., Schoellig, A. P., and Krause, A. No-regret Bayesian optimization with unknown hyperparameters. *Journal of Machine Learning Research*, 20:1–24, 2019.

Bull, A. D. Convergence rates of efficient global optimization algorithms. *Journal of Machine Learning Research*, 12(10), 2011.

Capone, A. and Hirche, S. Backstepping for partially unknown nonlinear systems using Gaussian processes. *IEEE Control Systems Letters*, 3:416–421, 2019.

Chowdhury, S. R. and Gopalan, A. On kernelized multi-armed bandits. In , *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 844–853. PMLR, 06–11 Aug 2017.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. Modeling wine preferences by data mining from physicochemical properties. *Decision support systems*, 47(4): 547–553, 2009.

Cressie, N. *Statistics for spatial data*. John Wiley & Sons, 2015.

Deisenroth, M. P., Fox, D., and Rasmussen, C. E. Gaussian processes for data-efficient learning in robotics and control. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):408–423, 2015.

Dicker, L. H., Foster, D. P., and Hsu, D. Kernel Ridge vs. Principal Component Regression: Minimax Bounds and the Qualification of Regularization Operators. *Electronic Journal of Statistics*, 11(1):1022–1047, 2017.

Farrell, J. A., Polycarpou, M., Sharma, M., and Dong, W. Command filtered backstepping. *IEEE Transactions on Automatic Control*, 54(6):1391–1395, 2009.

Fiedler, C., Scherer, C. W., and Trimpe, S. Practical and rigorous uncertainty bounds for Gaussian process regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7439–7447, 2021.

Gardner, J., Pleiss, G., Weinberger, K. Q., Bindel, D., and Wilson, A. G. Gpytorch: Blackbox matrix-matrix Gaussian process inference with GPU acceleration. In , *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Hewing, L., Kabzan, J., and Zeilinger, M. N. Cautious model predictive control using Gaussian process regression. *IEEE Transactions on Control Systems Technology*, 2019.

Hoffman, M. D., Gelman, A., et al. The no-u-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

Horn, R. A. and Johnson, C. R. *Matrix analysis*. Cambridge university press, 2012.

Kanagawa, M., Hennig, P., Sejdinovic, D., and Sriperumbudur, B. K. Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences. *arXiv preprint arXiv:1807.02582*, 2018.

Khalil, H. K. Noninear systems. *Prentice-Hall, New Jersey*, 2(5):5–1, 1996.

Kirschner, J., Mutny, M., Hiller, N., Ischebeck, R., and Krause, A. Adaptive and safe Bayesian optimization in high dimensions via one-dimensional subspaces. In *International Conference on Machine Learning*, pp. 3429–3438. PMLR, 2019.

Kocijan, J., Girard, A., Banko, B., and Murray-Smith, R. Dynamic systems identification with Gaussian processes. *Mathematical and Computer Modelling of Dynamical Systems*, 11(4):411–424, 2005.

Krstic, M., Kokotovic, P. V., and Kanellakopoulos, I. *Nonlinear and adaptive control design*. John Wiley & Sons, Inc., 1995a.

Krstic, M., Protz, J., Paduano, J., and Kokotovic, P. Backstepping designs for jet engine stall and surge control. In *Proceedings of 1995 34th IEEE Conference on Decision and Control*, volume 3, pp. 3049–3055, 1995b.

Kuleshov, V., Fenner, N., and Ermon, S. Accurate uncertainties for deep learning using calibrated regression. In *International Conference on Machine Learning*, pp. 2796–2804. PMLR, 2018.

Kwan, C. and Lewis, F. L. Robust backstepping control of nonlinear systems using neural networks. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 30(6):753–766, 2000.

Lederer, A., Umlauft, J., and Hirche, S. Uniform error bounds for Gaussian process regression with application to safe control. In *Advances in Neural Information Processing Systems*, pp. 659–669, 2019.

Lederer, A., Capone, A., and Hirche, S. Parameter optimization for learning-based control of control-affine systems. In , *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pp. 465–475. PMLR, 10–11 Jun 2020.

MacKay, D. J. C. Comparison of approximate methods for handling hyperparameters. *Neural Computation*, 11(5):1035–1068, 1999.

MacKay, D. J. C. *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, USA, 2002.

Maddalena, E. T., Scharnhorst, P., and Jones, C. N. Deterministic error bounds for kernel-based learning techniques under bounded noise. *Automatica*, 134:109896, 2021.

Mendelson, S. Improving the Sample Complexity using Global Data. *IEEE Transactions on Information Theory*, 48(7):1977–1991, 2002.

Micchelli, C. A., Xu, Y., and Zhang, H. Universal kernels. *Journal of Machine Learning Research*, 7(Dec):2651–2667, 2006.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P.,

Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Rao, C. R. *Linear statistical inference and its applications*, volume 2. Wiley New York, 1973.

Rasmussen, C. E. and Williams, C. K. Gaussian processes for machine learning. 2006. *The MIT Press, Cambridge, MA, USA*, 2006.

Robbins, H. *An Empirical Bayes Approach to Statistics*, pp. 157–164. University of California Press, 2020.

Sabanovic, A., Sabanovic, N., and Ohnishi, K. Sliding modes in power converters and motion control systems. *International Journal of Control*, 57(5):1237–1259, 1993.

Shi, L. Learning Theory Estimates for Coefficient-based Regularized Regression. *Applied and Computational Harmonic Analysis*, 34(2):252–265, 2013.

Šidák, Z. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. W. Information-theoretic regret bounds for Gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 58(5):3250–3265, 2012.

Sui, Y., Gotovos, A., Burdick, J., and Krause, A. Safe exploration for optimization with Gaussian processes. In , *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 997–1005, Lille, France, 07–09 Jul 2015. PMLR.

Sun, D., Khojasteh, M. J., Shekhar, S., and Fan, C. 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*, pp. 728–741. PMLR, 07 – 08 June 2021.

Sundararajan, S. and Keerthi, S. S. Predictive approaches for choosing hyperparameters in Gaussian processes. *Neural computation*, 13(5):1103–1118, 2001.

Tuo, R. and Wang, W. Kriging prediction with isotropic Matérn correlations: Robustness and experimental designs. *J. Mach. Learn. Res.*, 21:187–1, 2020.

Umlauft, J., Beckers, T., Kimmel, M., and Hirche, S. Feedback linearization using Gaussian processes. In *2017 IEEE 56th Annual Conference on Decision and Control*, pp. 5249–5255, 2017.

Wang, W., Tuo, R., and Jeff Wu, C. On prediction properties of kriging: Uniform error bounds and robustness. *Journal of the American Statistical Association*, 115(530):920–930, 2020.

Wendland, H. *Scattered Data Approximation*. Cambridge University Press, 2004.

Wilson, A. and Adams, R. Gaussian process kernels for pattern discovery and extrapolation. In *International conference on machine learning*, pp. 1067–1075. PMLR, 2013.

Wu, Z. M. and Schaback, R. Local Error Estimates for Radial Basis Function Interpolation of Scattered Data. *IMA Journal of Numerical Analysis*, 13(1):13–27, 1993.

A. Proofs

We begin by listing some well-known results required to prove the main statements from the paper.

Lemma A.1 (Wendland, 2004, Theorem 6.11). *Suppose that $k \in L_1(\mathbb{R}^d)$ is a continuous function. Then $k(\cdot)$ is positive definite if and only if its Fourier transform is nonnegative and nonvanishing.*

Lemma A.2 (Rao, 1973, p.70). *Let $A, B \in \mathbb{R}^{d \times d}$ be two symmetric positive-definite matrices, such that $A - B$ is symmetric positive-definite. Then $B^{-1} - A^{-1}$ is symmetric positive-definite.*

The following result is a direct consequence of the block matrix inversion formula, which can be found, e.g., in Horn & Johnson, 2012, p.18.

Lemma A.3 (Horn & Johnson, 2012, p.18). *Let*

$$K = \begin{pmatrix} \tilde{K} & k \\ k^\top & k \end{pmatrix}$$

be an $N \times N$ non-singular partitioned matrix. Then the lower-right entry of its inverse is given by

$$[K^{-1}]_{N,N} = \left(k - k^\top \tilde{K}^{-1} k \right)^{-1}.$$

By putting Lemma A.2 and Lemma A.3 together, we obtain the following statement.

Lemma A.4. *Let*

$$K_1 = \begin{pmatrix} \tilde{K}_1 & k_1 \\ k_1^\top & k_1 \end{pmatrix}, \quad K_2 = \begin{pmatrix} \tilde{K}_2 & k_2 \\ k_2^\top & k_2 \end{pmatrix}$$

be $N \times N$ symmetric positive definite partitioned matrices, such that $K_1 - K_2$ is also positive definite. Then

$$k_1 - k_1^\top \tilde{K}_1^{-1} k_1 \geq k_2 - k_2^\top \tilde{K}_2^{-1} k_2.$$

Proof. From Lemma A.2, we obtain that $q^\top K_2^{-1} q \geq q^\top K_1^{-1} q$ holds for any $q \in \mathbb{R}^N$. In particular, for $q = (0, \cdots 0, 1)^\top$, this implies, together with Lemma A.3,

$$(k_2 - k_2^\top \tilde{K}_2^{-1} k_2)^{-1} = q^\top K_2^{-1} q \geq q^\top K_1^{-1} q = (k_1 - k_1^\top \tilde{K}_1^{-1} k_1)^{-1},$$

i.e., $k_1 - k_1^\top \tilde{K}_1^{-1} k_1 \geq k_2 - k_2^\top \tilde{K}_2^{-1} k_2$ holds. $\qquad\square$

Using Lemma A.1, we then obtain the following.

Lemma A.5. *Let $\vartheta'', \vartheta, \vartheta'$ be lengthscales with $\vartheta'' > \vartheta > \vartheta'$, and let $\gamma = \left(\prod_{i=1}^d \frac{\vartheta_i''}{\vartheta_i'} \right)^{\frac{1}{2}}$. Furthermore, for an arbitrary measurement data set $\mathcal{D}$, let K_ϑ and $K_{\vartheta'}$ denote the corresponding covariance matrices computed using the kernels $k_\vartheta(\cdot, \cdot)$ and $k_{\vartheta'}(\cdot, \cdot)$, respectively. Then the matrix*

$$\gamma^2 K_{\vartheta'} - K_\vartheta$$

is positive semi-definite.

Proof. Recall that, by Assumption 2.1, the kernels $k_\vartheta(\cdot, \cdot)$ and $k_{\vartheta'}(\cdot, \cdot)$ are of the form

$$k_\vartheta(x, x') = k \left(\left(\frac{x_1 - x_1'}{\vartheta_1}, \ldots, \frac{x_d - x_d'}{\vartheta_d} \right)^\top \right), \qquad k_{\vartheta'}(x, x') = k \left(\left(\frac{x_1 - x_1'}{\vartheta_1'}, \ldots, \frac{x_d - x_d'}{\vartheta_d'} \right)^\top \right),$$

respectively, where $k(\cdot)$ has Fourier transform $\hat{k}(\omega) := \kappa(\|\omega\|_2)$. Define the matrices $T := \text{diag}(\frac{1}{\vartheta_1}, \ldots, \frac{1}{\vartheta_d})$ and $T' := \text{diag}(\frac{1}{\vartheta_1'}, \ldots, \frac{1}{\vartheta_d'})$, and the functions $k_T : \mathbb{R}^d \to \mathbb{R}$ and $k_{T'} : \mathbb{R}^d \to \mathbb{R}$, such that $k_T(x) := k(Tx)$ and $k_{T'}(x) := k(T'x)$

holds. Let $\hat{k}_{\boldsymbol{T}}(\cdot)$ and $\hat{k}_{\boldsymbol{T}'}(\cdot)$ denote the Fourier transform of of $k_{\boldsymbol{T}}(\cdot)$ and $k_{\boldsymbol{T}'}(\cdot)$, respectively. We then have

$$
\begin{aligned}
\hat{k}_{\boldsymbol{T}}(\boldsymbol{\omega}) &= \int\limits_{\mathbb{R}^d} k(\boldsymbol{T}\boldsymbol{x})e^{-i2\pi\boldsymbol{x}^\top\boldsymbol{\omega}}d\boldsymbol{x} \\
&= \int\limits_{\mathbb{R}^d} k(\boldsymbol{z})e^{-i2\pi\boldsymbol{z}^\top\boldsymbol{T}^{-\top}\boldsymbol{\omega}}\det(\boldsymbol{T}^{-1})d\boldsymbol{z} \\
&= \left(\prod_{i=1}^d \vartheta_i\right)\int\limits_{\mathbb{R}^d} k(\boldsymbol{z})e^{-i2\pi\boldsymbol{z}^\top\boldsymbol{T}^{-\top}\boldsymbol{\omega}}d\boldsymbol{z} \\
&= \left(\prod_{i=1}^d \vartheta_i\right)\hat{k}(\boldsymbol{T}^{-1}\boldsymbol{\omega}) = \left(\prod_{i=1}^d \vartheta_i\right)\kappa(\|\boldsymbol{T}^{-1}\boldsymbol{\omega}\|_2).
\end{aligned}
\tag{13}
$$

Since $\kappa(\cdot)$ is non-increasing and $\boldsymbol{\vartheta} > \boldsymbol{\vartheta}'$, we have that

$$
\kappa(\|(\boldsymbol{T}')^{-1}\boldsymbol{\omega}\|_2) \geq \kappa(\|\boldsymbol{T}^{-1}\boldsymbol{\omega}\|_2).
$$

Hence, (13) together with $\gamma^2 > \prod_{i=1}^d \frac{\vartheta_i}{\vartheta_i'}$ implies

$$
\gamma^2\hat{k}_{\boldsymbol{T}'}(\boldsymbol{\omega}) - \hat{k}_{\boldsymbol{T}}(\boldsymbol{\omega}) \geq \prod_{i=1}^d \frac{\vartheta_i}{\vartheta_i'}\hat{k}_{\boldsymbol{T}'}(\boldsymbol{\omega}) - \hat{k}_{\boldsymbol{T}}(\boldsymbol{\omega}) \geq 0
$$

for all $\boldsymbol{\omega} \in \mathbb{R}^d$. It then follows from Lemma A.1 that $\gamma^2 k_{\boldsymbol{T}'}(\cdot) - k_{\boldsymbol{T}}(\cdot)$ is a positive-definite function. Since

$$
\gamma^2 k_{\boldsymbol{T}'}(\boldsymbol{x} - \boldsymbol{x}') - k_{\boldsymbol{T}}(\boldsymbol{x} - \boldsymbol{x}') = \gamma^2 k_{\boldsymbol{\vartheta}'}(\boldsymbol{x}, \boldsymbol{x}') - k_{\boldsymbol{\vartheta}}(\boldsymbol{x}, \boldsymbol{x}')
$$

this implies the desired result. □

Proof of Lemma 3.3. The result follows directly from Lemma A.4 and Lemma A.5. □

In order to prove Theorem 3.5, we aim to bound the difference between posterior means $|\mu_{\boldsymbol{\vartheta}}(\boldsymbol{x}) - \mu_{\boldsymbol{\vartheta}'}(\boldsymbol{x})|$. To this end, we employ the two following results.

Lemma A.6. *Let $\|\cdot\|_{k_{\boldsymbol{\vartheta}}}$ denote the RKHS norm with respect to a kernel $k_{\boldsymbol{\vartheta}}$, and consider a function $\mu(\cdot)$ with bounded RKHS norm $\|\mu\|_{k_{\boldsymbol{\vartheta}}} < \infty$. Then, for all $\boldsymbol{x} \in \mathcal{X}$,*

$$
|\mu(\boldsymbol{x})|^2 \leq \sigma_{\boldsymbol{\vartheta}}^2(\boldsymbol{x})\left(\|\mu\|_{k_{\boldsymbol{\vartheta}}}^2 + \sum_{i=1}^N \left(\frac{\mu(\boldsymbol{x}_i)}{\sigma_n}\right)^2\right),
$$

holds, where σ_n is the noise variance and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ are the measurement inputs.

Proof. See Srinivas et al., 2012, Appendix B, eq. (11). □

Lemma A.7 (Bull, 2011)**.** *Let $\mu(\cdot)$ be a function with bounded reproducing kernel Hilbert space norm $\|\mu\|_{k_{\boldsymbol{\vartheta}}}$ with respect to the kernel $k_{\boldsymbol{\vartheta}}$. Then, for all $\boldsymbol{\vartheta}' \leq \boldsymbol{\vartheta}$,*

$$
\|\mu\|_{k_{\boldsymbol{\vartheta}'}}^2 \leq \prod_{i=1}^d \frac{\vartheta_i}{\vartheta_i'}\|\mu\|_{k_{\boldsymbol{\vartheta}}}^2.
$$

We are now able to bound the difference between the means of two Gaussian processes conditioned on the same data, as described in the next statement.

Lemma A.8. *Let Assumption 2.1 hold, let $\vartheta', \vartheta'', \vartheta, \vartheta_0 \in \Theta$ be vectors of lengthscales with $\vartheta' \leq \vartheta_0, \vartheta \leq \vartheta''$, and let $\mu_{\vartheta_0}(x), \mu_\vartheta(x)$ denote GP posterior means conditioned on some measurement data set $\mathcal{D} = (X, y)$. Then*

$$|\mu_{\vartheta_0}(x) - \mu_\vartheta(x)|^2 \leq \sigma_{\vartheta'}^2(x) \gamma^2 4 \frac{\|y\|_2^2}{\sigma_n^2}$$

holds for all $x \in \mathcal{X}$, where σ_n is the noise variance, and y is the measurement data.

Proof. Define $\tilde{K}_\vartheta := K_\vartheta + \sigma_n I$. From Lemma A.6, we have

$$|\mu_{\vartheta_0}(x) - \mu_\vartheta(x)|^2 \leq \sigma_{\vartheta'}^2(x) \left(\|\mu_{\vartheta_0} - \mu_\vartheta\|_{k_{\vartheta'}}^2 + \sum_{i=1}^N \frac{(\mu_\vartheta(x_i) - \mu_{\vartheta_0}(x_i))^2}{\sigma_n^2} \right) \tag{14}$$

Recall that, for a function of the form $\mu(\cdot) = \sum_{i=1}^N \alpha_i k_\vartheta(\cdot, x_i)$, its RKHS norm with respect to the kernel $k_\vartheta(\cdot, \cdot)$ is given by

$$\|\mu_\vartheta\|_{k_\vartheta}^2 = \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j k_\vartheta(x_i, x_j).$$

Hence, the RKHS norm with respect to $k_\vartheta(\cdot, \cdot)$ of a posterior mean function $\mu_\vartheta(\cdot)$ can be upper-bounded as

$$\|\mu_\vartheta\|_{k_\vartheta}^2 = y^\top \tilde{K}_\vartheta^{-1} K_\vartheta \tilde{K}_\vartheta^{-1} y \leq y^\top \tilde{K}_\vartheta^{-1} y \leq \frac{\|y\|_2^2}{\sigma_n^2},$$

where the first inequality is due to the eigenvalues of $\tilde{K}_\vartheta^{-\frac{1}{2}} K_\vartheta \tilde{K}_\vartheta^{-\frac{1}{2}}$ being smaller or equal to one, whereas the last inequality holds because the eigenvalues of $\tilde{K}_\vartheta$ are greater or equal to σ_n^2. Furthermore, the summation in (14) can be bounded as

$$\sum_{i=1}^N \frac{(\mu_\vartheta(x_i) - \mu_{\vartheta_0}(x_i))^2}{\sigma_n^2} = \frac{1}{\sigma_n^2} \|K_\vartheta \tilde{K}_\vartheta^{-1} y - K_{\vartheta_0} \tilde{K}_{\vartheta_0}^{-1} y\|_2^2$$

$$\leq \frac{1}{\sigma_n^2} \|K_\vartheta \tilde{K}_\vartheta^{-1} y\|_2^2 + \frac{1}{\sigma_n^2} \|K_{\vartheta_0} \tilde{K}_{\vartheta_0}^{-1} y\|_2^2 \leq \frac{2\|y\|_2^2}{\sigma_n^2}.$$

Plugging this into (14) together with Lemma A.7 and $\gamma > 1$ yields

$$|\mu_{\vartheta_0}(x) - \mu_\vartheta(x)|^2$$

$$\leq \sigma_{\vartheta'}^2(x) \left(\|\mu_{\vartheta_0} - \mu_\vartheta\|_{k_{\vartheta'}}^2 + \sum_{i=1}^N \frac{(\mu_\vartheta(x_i) - \mu_{\vartheta_0}(x_i))^2}{\sigma_n^2} \right)$$

$$\leq \sigma_{\vartheta'}^2(x) \left(\|\mu_{\vartheta_0}\|_{k_{\vartheta'}}^2 + \|\mu_\vartheta\|_{k_{\vartheta'}}^2 + \frac{2\|y\|_2^2}{\sigma_n^2} \right)$$

$$\leq \sigma_{\vartheta'}^2(x) \left(\prod_{i=1}^d \frac{\vartheta_{0,i}}{\vartheta_i'} \|\mu_{\vartheta_0}\|_{k_{\vartheta_0}}^2 + \prod_{i=1}^d \frac{\vartheta_i}{\vartheta_i'} \|\mu_\vartheta\|_{k_\vartheta}^2 + \frac{2\|y\|_2^2}{\sigma_n^2} \right)$$

$$\leq \sigma_{\vartheta'}^2(x) \left(\gamma^2 \|\mu_{\vartheta_0}\|_{k_{\vartheta_0}}^2 + \gamma^2 \|\mu_\vartheta\|_{k_\vartheta}^2 + \frac{2\|y\|_2^2}{\sigma_n^2} \right)$$

$$\leq \sigma_{\vartheta'}^2(x) \gamma^2 4 \frac{\|y\|_2^2}{\sigma_n^2}.$$

$\square$

EXPERIMENT	BSTN	ML	WINE	SRCS	CONTROL
SIGNAL VARIANCE	$[1, 50]$	$[10^{-10}, 10^5]$	$[10^{-2}, 10^2]$	$[10^{-1}, 10^3]$	$[10^{-6}, 10]$
LENGTHSCALES	$[10^{-1}, 10^2]$	$[10^{-10}, 5 \times 10^{12}]$	$[10^{-2}, 10]$	$[10^{-1}, 50]$	$[10^{-15}, 10^{-2}]$
NOISE VARIANCE	$[10^{-1}, 10^2]$	$[10^{-5}, 10^2]$	$[10^{-2}, 1]$	$[10^{-2}, 80]$	$[10^{-5}, 10^{-1}]$

Table 2. Lower/upper bounds of uniform distributions used as hyperpriors. BSTN stands for Boston (house prices), ML for Mauna Loa, and SRCS for Sarcos. Control refers to the backstepping control experiment in Section 5.

Proof of Theorem 3.5. By applying Assumption 3.2 and Lemmas 3.3 and A.8, we obtain

$$
\begin{aligned}
&|f(\boldsymbol{x}) - \mu_{\boldsymbol{\vartheta}_0}(\boldsymbol{x})| \\
\leq &|f(\boldsymbol{x}) - \mu_{\boldsymbol{\vartheta}}(\boldsymbol{x})| + |\mu_{\boldsymbol{\vartheta}}(\boldsymbol{x}) - \mu_{\boldsymbol{\vartheta}_0}(\boldsymbol{x})| \\
\leq &\left(\max_{\boldsymbol{\vartheta}' \leq \boldsymbol{\vartheta} \leq \boldsymbol{\vartheta}''} \beta^{\frac{1}{2}}(\boldsymbol{\vartheta}) \right) \gamma \sigma_{\boldsymbol{\vartheta}'}(\boldsymbol{x}) + \frac{2\|\boldsymbol{y}\|_2}{\sigma_n} \gamma \sigma_{\boldsymbol{\vartheta}'}(\boldsymbol{x}) \\
= &\bar{\beta}^{\frac{1}{2}} \sigma_{\boldsymbol{\vartheta}'}(\boldsymbol{x}).
\end{aligned}
$$

. $\qquad\qquad\qquad\square$

Proof of Theorem 3.7. It follows from the definition of $\mathcal{P}_\delta$ that $\boldsymbol{\vartheta}' \leq \boldsymbol{\vartheta} \leq \boldsymbol{\vartheta}''$ holds with probability at least $(1 - \delta)$. Theorem 3.7 is then a direct consequence of Theorem 3.5. $\qquad\square$

Proof of Theorem 4.2. The dynamics of the control error can be written in the compact form (Capone & Hirche, 2019)

$$
\dot{\boldsymbol{e}} = \dot{\boldsymbol{x}} - \dot{\boldsymbol{x}}_d = \Delta \boldsymbol{f} - (\boldsymbol{C} + \boldsymbol{G})\, \boldsymbol{e}, \tag{15}
$$

where $\Delta \boldsymbol{f} := (f_1 - \mu_1, \dots, f_d - \mu_d)$, $\boldsymbol{C} := \mathrm{diag}(C_1(\boldsymbol{x}, \dots, C_d(\boldsymbol{x})))$, and

$$
\boldsymbol{G} :=
\begin{pmatrix}
0 & g_1 & 0 & \cdots & & 0 \\
-g_1 & 0 & g_2 & & & \vdots \\
0 & -g_2 & 0 & & & \\
\vdots & & & \ddots & & g_{d-1} \\
0 & \cdots & & & -g_{d-1} & 0
\end{pmatrix}.
$$

Consider then the Lyapunov function $V(\boldsymbol{e}) = \frac{1}{2}\boldsymbol{e}^\top \boldsymbol{e}$. Due to Theorem 3.7, the corresponding time derivative is bounded with probability $(1 - \delta)^m (1 - \rho)^m$ as

$$
\dot{V}(\boldsymbol{e}) = \boldsymbol{e}^\top \dot{\boldsymbol{e}} = \boldsymbol{e}^\top \Delta \boldsymbol{f} - \boldsymbol{e}^\top \boldsymbol{C} \boldsymbol{e} \tag{16}
$$

$$
\leq \|\boldsymbol{e}\|_2 \left(\|\Delta \boldsymbol{f}\|_2 - \|\boldsymbol{e}\|_2 \min_i C_i(\boldsymbol{x}) \right) \tag{17}
$$

$$
\leq \|\boldsymbol{e}\|_2 C_i(\boldsymbol{x}) \left(\xi_{\mathrm{des}} - \|\boldsymbol{e}\|_2 \right), \tag{18}
$$

where we employ $\boldsymbol{e}^\top \boldsymbol{G} \boldsymbol{e} = \boldsymbol{0}$. The term $\|\boldsymbol{e}\|_2 C_i(\boldsymbol{x}) \left(\xi_{\mathrm{des}} - \|\boldsymbol{e}\|_2 \right)$ is negative for $\|\boldsymbol{e}\|_2 < \xi_{\mathrm{des}}$, hence (16) implies that the norm of $\boldsymbol{e}$ does not grow beyond ξ_{des} after a specific time T (Khalil, 1996) with probability $(1 - \delta)^m (1 - \rho)^m$. $\qquad\square$

B. Experiments

All experiments were carried out using an AMD Ryzen Threadripper 2990WX with 32 cores.

The hyperpriors in all cases except the Sarcos experiments with more than $N = 850$ training inputs are uniform distributions. The corresponding upper and lower bounds are shown in Table 2.

B.1. Setting $\bar{\beta} = \beta$

In the experimental section, we employ $\bar{\beta} = \beta$, as opposed to explicitly computing the values suggested by Theorem 3.7. We now provide some justification for this choice.

Recall that $\bar{\beta}^{\frac{1}{2}}$ in Theorem 3.7 is computed as

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

As can be seen in the proof of Theorem 3.5, the term $2\sigma_n^{-1}\|\boldsymbol{y}\|_2$ is used to upper-bound the discrepancy $|\mu_{\boldsymbol{\vartheta}}(\boldsymbol{x}) - \mu_{\boldsymbol{\vartheta}_0}(\boldsymbol{x})|$ between the posterior mean given lengthscales $\boldsymbol{\vartheta}$ sampled from the posterior and that of the working hyperparameters $\boldsymbol{\vartheta}_0$. Hence, if less conservative bounds for the difference $|\mu_{\boldsymbol{\vartheta}}(\boldsymbol{x}) - \mu_{\boldsymbol{\vartheta}_0}(\boldsymbol{x})|$ are available, they can be used to replace the estimate $2\sigma_n^{-1}\|\boldsymbol{y}\|_2$ without losing theoretical guarantees. In our experiments, we observed that the difference $|\mu_{\boldsymbol{\vartheta}}(\boldsymbol{x}) - \mu_{\boldsymbol{\vartheta}_0}(\boldsymbol{x})|$ was often small, and that Lemma A.4 always held with $\gamma = 1$, which in turn suggests that, in the case of Gaussian kernels, the posterior variance is decreasing with respect to the lengthscales. Furthermore, we have

$$\max_{\boldsymbol{\vartheta}' \leq \boldsymbol{\vartheta} \leq \boldsymbol{\vartheta}''} \beta^{\frac{1}{2}}(\boldsymbol{\vartheta}) = \sqrt{2}$$

for all $\boldsymbol{\vartheta}$ by assumption. Hence, setting $\bar{\beta} = \beta$, corresponds to ignoring the discrepancy term $|\mu_{\boldsymbol{\vartheta}}(\boldsymbol{x}) - \mu_{\boldsymbol{\vartheta}_0}(\boldsymbol{x})|$ and choosing $\gamma = 1$. We thus obtain the uniform error bound

$$|f(\boldsymbol{x}) - \mu_{\boldsymbol{\vartheta}_0}(\boldsymbol{x})| \leq \beta^{\frac{1}{2}} \sigma_{\boldsymbol{\vartheta}'}(\boldsymbol{x}).$$

We stress that this upper bound will still be more conservative than that obtained with the working hyperparameters $\boldsymbol{\vartheta}_0$.

B.2. Laplace Approximation and Empirical Bayes

We now detail the Laplace approximation and empirical Bayes approach used for the two largest Sarcos data sets ($N = 5000$ and $N = 10000$).

Typically, the Laplace approximation of the posterior is obtained by computing the Hessian of the posterior around its maximum, and then treating the negative Hessian, which should be positive definite, as the covariance matrix of the corresponding Gaussian approximation (MacKay, 2002).

In our setting, using uniform hyperpriors would lead to the Hessian of the posterior being the Hessian of the log likelihood, since the derivatives of the uniform hyperprior are zero almost everywhere. However, the Hessian of the log likelihood is not always negative definite due to numerical issues or premature termination of the optimization algorithm. Moreover, in high-dimensional settings such as the Sarcos data set, it is often poorly peaked, even for large N. Since the resulting normal distribution extends to the domain with negative lengthscales, as opposed to the exact posterior, the corresponding confidence region can be exceedingly conservative. For these reasons, we consider an empirical Bayes approach, where the prior is chosen based on the data (Robbins, 2020). In our case, this is achieved by specifying a normal hyperprior around the (estimated) maximum of the log likelihood $\boldsymbol{\vartheta}_0$, i.e.,

$$\log p(\boldsymbol{\vartheta}) \hat{=} - (\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0)^\top \boldsymbol{H}_p (\boldsymbol{\vartheta} - \boldsymbol{\vartheta}_0),$$

where $\hat{=}$ denotes equality up to a constant, and $\boldsymbol{H}_p = h_p \boldsymbol{I}$ is a diagonal matrix such that the Hessian of the log posterior

$$\boldsymbol{H}_{\boldsymbol{\vartheta}} \left(\log(p(\boldsymbol{\vartheta}|\boldsymbol{y}, \boldsymbol{X})) \right) = -\boldsymbol{H}_p + \boldsymbol{H}_{\boldsymbol{\vartheta}} \left(\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\vartheta}) \right)$$

is negative definite. Here $\boldsymbol{H}_{\boldsymbol{\vartheta}}(\cdot)$ denotes the Hessian operator with respect to $\boldsymbol{\vartheta}$. Note that increasing h_p results in a less conservative set of bounding hyperparameters $\boldsymbol{\vartheta}', \boldsymbol{\vartheta}''$, since it implies more confidence in the estimated maximum $\boldsymbol{\vartheta}_0$. In the Sarcos experiments, we set h_p to ten times the largest nonnegative eigenvalue of $\boldsymbol{H}_{\boldsymbol{\vartheta}} \left(\log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\vartheta}) \right)$.

B.3. Control Problem

The dimensionless parameters of the dynamical system are given by $M = 0.05$, $B = D = 1$, $G = Z = 10$, $H = 0.5$. We employ a command-filtered backstepping approach to bypass the computation of the control input, which has no practical impact on the results if a corresponding filtering parameter is chosen high enough. For more details, see (Farrell et al., 2009; Capone & Hirche, 2019).