Error Bounds For Gaussian Process Regression Under Bounded Support Noise With Applications To Safety Certification

Robert Reed[1], **Luca Laurenti**[2], **Morteza Lahijanian**[1]

[1]Deptartment of Aerospace Engineering Sciences, University of Colorado Boulder, USA
[2]Delft Center for Systems and Control, Delft University of Technology, The Netherlands
Robert.Reed-1@colorado.edu, L.Laurenti@tudelft.nl, Morteza.Lahijanian@colorado.edu

Abstract

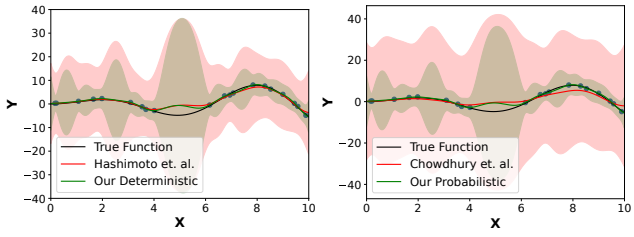
Gaussian Process Regression (GPR) is a powerful and elegant method for learning complex functions from noisy data with a wide range of applications, including in safety-critical domains. Such applications have two key features: (i) they require rigorous error quantification, and (ii) the noise is often bounded and non-Gaussian due to, e.g., physical constraints. While error bounds for applying GPR in the presence of non-Gaussian noise exist, they tend to be overly restrictive and conservative in practice. In this paper, we provide novel error bounds for GPR under bounded support noise. Specifically, by relying on concentration inequalities and assuming that the latent function has low complexity in the reproducing kernel Hilbert space (RKHS) corresponding to the GP kernel, we derive both probabilistic and deterministic bounds on the error of the GPR. We show that these errors are substantially tighter than existing state-of-the-art bounds and are particularly well-suited for GPR with neural network kernels, i.e., Deep Kernel Learning (DKL). Furthermore, motivated by applications in safety-critical domains, we illustrate how these bounds can be combined with stochastic barrier functions to successfully quantify the safety probability of an unknown dynamical system from finite data. We validate the efficacy of our approach through several benchmarks and comparisons against existing bounds. The results show that our bounds are consistently smaller, and that DKLs can produce error bounds tighter than sample noise, significantly improving the safety probability of control systems.

1 Introduction

Gaussian Process Regression (GPR) is a powerful and elegant method for learning complex functions from noisy data, renowned for its rigorous uncertainty quantification (Rasmussen, Williams et al. 2006). This makes GPR particularly valuable in the control and analysis of *safety-critical* systems (Berkenkamp et al. 2017; Lederer and Hirche 2019; Lederer, Umlauft, and Hirche 2019; Jagtap, Soudjani, and Zamani 2020; Jackson et al. 2021b,a; Griffioen, Devonport, and Arcak 2023; Wajid, Awan, and Zamani 2022). However, in such systems, the underlying assumption of Gaussian measurement noise often does not hold. Measurements are typically filtered to reject outliers, and physical systems cannot traverse infinite distances in a single time step. Consequently, bounded support noise distributions provide a more accurate representation for many cyber-physical systems. But, without the Gaussian assumption, the mean and variance predictions of GPs cannot be directly used to represent confidence in the underlying system. To address this, recent works (Hashimoto et al. 2022; Srinivas et al. 2012; Chowdhury and Gopalan 2017; Jackson et al. 2021a) have diverged from a fully Bayesian approach and developed GPR error bounds under mild assumptions on the sample noise distribution, specifically sub-Gaussian and bounded noise (see Figure 1). While these bounds are useful, they tend to be overly restrictive and conservative, relying on parameters that must be over-approximated. This work aims to overcome these limitations by providing novel, tighter error bounds for GPR under bounded support noise, enhancing their applicability in safety-critical domains.

In this paper, we present novel error bounds for GPR under bounded support noise. Our key insight is that two factors contribute to the regression error: the error due to GP mean prediction at a point without noise, and the error due to noise perturbing the noise-free prediction by a factor proportional to the noise's support. Then, by relying on concentration inequalities and assuming the latent function has low complexity in the reproducing kernel Hilbert space (RKHS) corresponding to the GP kernel, we derive both probabilistic (i.e., with some confidence) and deterministic bounds on these error terms. Specifically, we use convex optimization and Hoeffding's Inequality to obtain accurate bounds. We demonstrate that these errors are substantially tighter than existing bounds, as depicted in Figure 1. Furthermore, our bounds are particularly well-suited for GPs with neural network kernels, such as Deep Kernel Learning (DKL). We also illustrate how these bounds can be combined with stochastic barrier functions (Kushner 1967; Santoyo, Dutreix, and Coogan 2021), which are a generalization of Lyapunov functions, to quantify the safety probability of an unknown control system from finite data. We validate our approach through several benchmarks and comparisons, showing that our bounds are consistently smaller than the state-of-the-art, resulting in more accurate safety certification of control systems.

The key contributions of this work are three-fold: (i) derivation of novel probabilistic and deterministic error bounds for GP regression with bounded support noise, (ii) demonstration of the application of these bounds with stochastic barrier



(a) Deterministic Bounds (b) Probabilistic Bounds

Figure 1: Predictive mean and error bounds when learning from 20 samples of $y = x\sin(x) + v$ with $|v| < 0.5$. In (a) we plot the comparison of deterministic error bounds and mean predictions, and in (b) we show the probabilistic bounds that hold with 95% probability. We set $\sigma_n = 0.1$ for our predictions, $\sigma_n = 0.5$ for Hashimoto et al. as per Lemma 2, and $\sigma_n^2 = 1 + 2/20$ for Chowdhury et al. as per Lemma 1. Note bounds for Abbasi-Yadkori (2013) are nearly identical to Chowdhury et al. in this example, see Appendix of (Reed, Laurenti, and Lahijanian 2024).

functions to effectively quantify the safety probability of unknown control systems by solely using data, and (iii) validation of the approach through extensive benchmarks and comparisons, showing consistently tighter error bounds than state-of-the-art methods and enhanced accuracy in determining safety probabilities for barrier certificates.

Related Works Several studies consider relaxations of the Gaussian assumption in the GP regression setting. In particular, Snelson, Ghahramani, and Rasmussen propose an approach that automatically generates a monotonic, non-linear warping function to map the observations into a latent space where it can be best modelled as a GP. However there is no guarantee that the warped data follows a GP. Also, the method requires an inverse of the warping function, but it is generally only available in an approximate form. Hence, the model cannot be used to generate formal guarantees. A closer work to ours is (Jensen, Nielsen, and Larsen 2013), which derives posteriors for GP regression when the observation space has bounded support, and hence the noise has bounded support. However, similar to (Snelson, Ghahramani, and Rasmussen 2003), the presented derivations provide approximate posteriors for the GP models, which limits their viability in safety-critical settings.

Works by Srinivas et al.; Abbasi-Yadkori; Chowdhury and Gopalan; Hashimoto et al. provide formal regression error bounds for GP models under non-Gaussian noise. Specifically, work (Srinivas et al. 2012) first develops a framework for GP regression under the assumption that noise is R-sub-Gaussian and then formally quantifies probabilistic bounds on the GP prediction errors. Using a similar framework, (Chowdhury and Gopalan 2017) derives tighter probabilistic bounds. Work (Abbasi-Yadkori 2013) derives similar bounds for Kernel Ridge Regression, which uses the same mean prediction as a GPR without a posterior variance. While applicable, the setting in each of these works is more general than ours, which results in larger error terms when assessed with bounded support noise. Work (Hashimoto et al. 2022) specifically focuses on bounded noise and derives a deterministic error bound. This bound is generally tighter than the prior probabilistic bounds; nevertheless, it becomes loose as the size of the support increases. Since both error bounds in (Chowdhury and Gopalan 2017; Hashimoto et al. 2022) are directly applicable to our scenario, we show their derivation in Section 2 and compare our results against them. In the experiments, we also compare against (Srinivas et al. 2012; Abbasi-Yadkori 2013).

Finally, recent work (Maddalena, Scharnhorst, and Jones 2021) derives a tighter deterministic error bound when compared to the results of (Hashimoto et al. 2022) by using a Kernel Ridge Regression approach. However, it requires the kernel to be strictly positive definite. While the squared exponential kernel, which is a popular choice in the controls application, can satisfy this requirement, the kernel matrix can be ill-conditioned resulting in inaccurate inversions. This restriction also limits the use of more expressive kernels that are positive semi-definite which may innately improve the error bounds, such as deep kernels. We show that our bounds are well-suited for deep kernels.

2 Setup and Background

We consider a stochastic system of the following form:

$$\mathbf{y} = f(\mathbf{x}) + \mathbf{v}, \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^d$ is the input, $\mathbf{y} \in \mathbb{R}$ is the output, and $\mathbf{v} \in V \subset \mathbb{R}$ is an additive zero-mean[1] random variable (noise) with bounded support $\sigma_v \in \mathbb{R}_{\geq 0}$, i.e., $V = \{v \in \mathbb{R} \mid |v| \leq \sigma_v\}$, and unknown distribution. Function $f : \mathbb{R}^d \to \mathbb{R}$ is *unknown*. The main goal of the paper is to regress f from a dataset of input-output pairs $D = \{(x_i, y_i)\}_{i=1}^m$ and derive bounds on the error between the regressed model predictions and the true function f. However, taking f as entirely unknown can lead to an ill-posed problem; hence, we impose a standard smoothness (well-behaved) assumption (Srinivas et al. 2012; Jackson et al. 2021b) on f.

Assumption 1 (RKHS Continuity)**.** *For a compact set $X \subset \mathbb{R}^d$, let $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ be a given continuously differentiable kernel and $\mathcal{H}_\kappa(X)$ the reproducing kernel Hilbert space (RKHS) corresponding to κ over X with induced norm $\|\cdot\|_\kappa$. Let $f(\cdot) \in \mathcal{H}_\kappa(X)$. Moreover, there exists a constant $B \geq 0$ s.t. $\|f(\cdot)\|_\kappa \leq B$.*

Assumption 1 implies that $f(x) = \sum_{n=1}^{\infty} \alpha_n \kappa(x, x_n)$ for representer points $\alpha_n \in \mathbb{R}$ and $x_n \in \mathbb{R}^d$. Note that for most kernels κ used in practice, such as the squared exponential kernel, $\mathcal{H}_\kappa$ is dense in the space of continuous functions (Steinwart 2001).

We stress that, in this paper, we do not assume that f is a sample from a Gaussian process prior. Moreover, since the noise is non-Gaussian, the likelihood model is also not Gaussian. Nevertheless, similar to the *agnostic* setting described in (Srinivas et al. 2012; Chowdhury and Gopalan

[1]The assumption that $\mathbb{E}[\mathbf{v}] = 0$ is without loss of generality. In fact, if $\mathbb{E}[\mathbf{v}] \neq 0$, we can add a bias term to f to shift the expectation to 0.

2017; Hashimoto et al. 2022), we would still like to use the GP regression framework to regress f with misspecified noise and prior models. In the remainder of this section, we provide a brief background on GP regression and state the existing error bounds on the learning error. In Section 3, we derive new bounds and discuss why they are tighter than the existing ones. We also illustrate the practical utility of these bounds in a control application in Section 4. Finally, we provide empirical evaluations of the new errors bounds in Section 5.

Existing Error Bounds Using GPs and RKHS

Let $D = \{(x_i, y_i)\}_{i=1}^m$ be a dataset consisting of m input-output observations pairs from System (1). A popular method to predict the output of f at a new input x^* with confidence on the predicted value is Gaussian Process regression (GPR). A Gaussian Process (GP) is a collection of random variables, such that any finite collection of those random variables is jointly Gaussian (Rasmussen, Williams et al. 2006). GPR starts by placing a Gaussian prior over f through use of a kernel $\kappa : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$. Then, under the assumption that $\mathbf{v}$ is a zero-mean Gaussian independent random variable with variance σ_n^2 for any new input x^*, the predictive mean $\mu_D(x^*)$ and variance $\sigma_D^2(x^*)$ can be derived as:

$$\mu_D(x^*) = K_{x^*,\mathcal{X}}(K_{\mathcal{X},\mathcal{X}} + \sigma_n^2 I)^{-1} Y, \quad (2)$$

$$\sigma_D^2(x^*) = K_{x^*,x^*} - K_{x^*,\mathcal{X}}(K_{\mathcal{X},\mathcal{X}} + \sigma_n^2 I)^{-1} K_{\mathcal{X},x^*}, \quad (3)$$

where $\mathcal{X} = [x_1, \ldots, x_m]^T$ and $Y = [y_1, \ldots, y_m]^T$ are respectively the vectors of input and output data, $K_{\mathcal{X},\mathcal{X}}$ is a matrix whose i-th row and j-th column is $\kappa(x_i, x_j)$, and $K_{x,\mathcal{X}} = K_{\mathcal{X},x}^T$ are vectors whose i-th entry is $\kappa(x, x_i)$.

As already mentioned, in our setting in System (1), we do not assume f is a sample from a GP and the noise $\mathbf{v}$ is strictly bounded (hence, non-Gaussian). Fortunately, recent works show that, even under such settings, as long as Assumption 1 holds, the GP analytical posterior prediction can still be used outside of the Bayesian framework, i.e., even with misspecified noise and priors. Specifically, Work (Chowdhury and Gopalan 2017) shows that when the measurement noise is conditionally R-sub-Gaussian, which includes bounded support distributions, probabilistic bounds on the error between the prediction μ_D and f can be derived as follows.

Lemma 1 ((Chowdhury and Gopalan 2017, Theorem 2))**.** *Let $X \subset \mathbb{R}^d$ be a compact set, $B > 0$ be the bound on $\|f\|_\kappa \leq B$, and $\Gamma \in \mathbb{R}_{>0}$ be the maximum information gain of κ. If noise $\mathbf{v}$ has a R-sub-Gaussian distribution and μ_D and σ_D are obtained via Equations* (2)-(3) *on a dataset D of size m with $\sigma_n^2 = 1 + 2/m$, then it holds that, for every $\delta \in (0, 1]$,*

$$\mathbb{P}\Big(\forall x \in X,\ |\mu_D(x) - f(x)| \leq \beta(\delta)\sigma_D(x)\Big) \geq 1 - \delta, \quad (4)$$

where $\beta(\delta) = B + R\sqrt{2(\Gamma + 1 + \log(1/\delta))}$.

Lemma 1 provides probabilistic bounds on the GP regression error $|\mu_D(x) - f(x)|$ by relying on the kernel parameters such as RKHS norm bound B and information gain Γ, which are often difficult to obtain accurately. They can however be (conservatively) bounded using techniques introduced in (Srinivas et al. 2012; Jackson et al. 2021a; Hashimoto et al. 2022). Another important observation in Lemma 1 is that, when noise $\mathbf{v}$ has a R-sub-Gaussian distribution, variance σ_n^2 is set to $1 + 2/m$, which adds conservatism to the bounds by substantially increasing the variance of GP predictions. Note that to obtain estimates with confidence 1, i.e., $\delta = 0$, Lemma 1 requires infinitely-many samples. Alternative probabilistic bounds in this setting are considered by Srinivas et al., which makes use of similar parameters, and by Abbasi-Yadkori which leaves σ_n as a decision variable (see the extended version Appendix (Reed, Laurenti, and Lahijanian 2024)). As shown in our experiments, they are also conservative.

In the case that noise $\mathbf{v}$ has bounded support (i.e., $|\mathbf{v}| \leq \sigma_v$), deterministic bounds on the prediction error are provided by Hashimoto et al..

Lemma 2 ((Hashimoto et al. 2022, Lemma 2))**.** *Let X, κ, and B be as in Lemma 1. If noise $\mathbf{v} \in V$ has a finite support σ_v (i.e., $|\mathbf{v}| \leq \sigma_v$) and μ_D and σ_D are obtained via Equations* (2)-(3) *on a dataset D of size m with $\sigma_n = \sigma_v$, then it holds that, for every $x \in X$,*

$$|\mu_D(x) - f(x)| \leq \beta_T \sigma_D(x), \quad (5)$$

where $\beta_T = \sqrt{B^2 - Y^T(K_{\mathcal{X},\mathcal{X}} + \sigma_v^2 I)^{-1} Y + m}$, and Y and $K_{\mathcal{X},\mathcal{X}}$ are as in Equations (2)-(3).

By restricting noise to a bounded set, Lemma 2 is able to provide a deterministic bound on the GP regression error $|\mu_D(x) - f(x)|$. Unlike the probabilistic error bounds in Lemma 1, the deterministic bound in Lemma 2 does not rely on the information gain Γ, removing a source of conservatism. However, it is generally conservative when σ_v is not small, due to placing $\sigma_n = \sigma_v$ in Equations (2)-(3).

Note that the bounds in Lemma 1 employ assumptions that are too general for our problem, i.e., R-sub-Gaussian (conditioned on the filtration) vs bounded support noise independent of filtration, and rely on parameters that must be approximated. Similarly, the bounds in Lemma 2 do not allow probabilistic reasoning and are restrictive when the support on noise is large. Hence, there is a need for probabilistic error bounds derived specifically for bounded support noise, as well as a need for an improved deterministic error bound that does not grow conservatively with the size of the support.

3 Bounded Support Error Bounds

In this section, we introduce novel probabilistic and deterministic error bounds for GPR of System (1), where noise has a bounded support distribution. We show that in this setting, the results of Lemmas 1 and 2 can be substantially improved. All the proofs can be found in the Appendix of the extended version (Reed, Laurenti, and Lahijanian 2024).

Probabilistic Error Bounds

We begin with the probabilistic bounds. With an abuse of notation, for the vector of input samples $\mathcal{X} = [x_1, \ldots, x_m]^T$, let $f(\mathcal{X}) = [f(x_1), \ldots, f(x_m)]^T$. Then, we observe that the noise output vector Y is such that $Y = f(\mathcal{X}) + \mathcal{V}$, where $\mathcal{V} = [v_1, \ldots, v_m]^T$ is a vector of i.i.d. samples of the noise,

each of which is bounded by σ_v. Consequently, from Eqn (2) and by denoting $G = (K_{\mathcal{X},\mathcal{X}} + \sigma_n^2 I)^{-1}$ and $W_x = K_{x,\mathcal{X}} G$, we can bound the GP regression learning error as

$$|\mu_D(x) - f(x)| = |W_x(f(\mathcal{X}) + \mathcal{V}) - f(x)| \quad (6)$$
$$\leq |W_x f(\mathcal{X}) - f(x)| + |W_x \mathcal{V}|. \quad (7)$$

Hence, the error is bounded by a sum of two terms: $|W_x f(\mathcal{X}) - f(x)|$, which is the error due to mean prediction at x with no noise, and $|W_x \mathcal{V}|$, which is the error due to the noise with a value at most proportional to σ_v. The following lemma, which extends results in (Hashimoto et al. 2022), bounds the first term.

Lemma 3. *Let X, κ, B, and D be as in Lemma 1, and $G = (K_{\mathcal{X},\mathcal{X}} + \sigma_n^2 I)^{-1}$, $W_x = K_{x,\mathcal{X}} G$, and $c^* \leq f(\mathcal{X})^T G f(\mathcal{X})$. Then, it holds that, for every $x \in X$,*

$$|W_x f(\mathcal{X}) - f(x)| \leq \sigma_D(x)\sqrt{B^2 - c^*}. \quad (8)$$

In Theorem 1, we rely on Hoeffding's Inequality (Mohri, Rostamizadeh, and Talwalkar 2018) to bound the second term in Eqn (7), which in turn provides a probabilistic bound on the GP regression error when combined with Lemma 3.

Theorem 1 (Bounded Support Probabilistic RKHS Error)**.** *Let X, κ, B, D, G, W_x, and c^* be as in Lemma 3, and define $\lambda_x = 4\sigma_v^2 K_{x,\mathcal{X}} G^2 K_{\mathcal{X},x}$. If noise $\mathbf{v} \in V$ is zero-mean and has a finite support σ_v (i.e., $|\mathbf{v}| \leq \sigma_v$) and μ_D and σ_D are obtained via Eqns* (2)-(3) *on dataset D with any choice of $\sigma_n > 0$, then it holds that, for every $x \in X$ and $\delta \in (0, 1]$,*

$$\mathbb{P}\Big(|\mu_D(x) - f(x)| \leq \epsilon(x, \delta)\Big) \geq 1 - \delta, \quad (9)$$

where $\epsilon(x, \delta) = \sigma_D(x)\sqrt{B^2 - c^*} + \sqrt{\frac{\lambda_x}{2} ln \frac{2}{\delta}}$.

The proof uses Lemma 3 for the first term of $\epsilon(x, \delta)$ and derives the second term by applying Hoeffding's Inequality to $|W_x \mathcal{V}|$ by noting that $\mathbb{E}[W_x \mathcal{V}] = 0$ and each $-\sigma_v \leq v_i \leq +\sigma_v$, enabling the random variable to maintain bounded support on each term. Note that Theorem 1 only requires two parameters in its probabilistic bound: c^* and B. In fact, c^* can be found by solving the following quadratic optimization problem, where we rely on the boundedness of the support of the distribution of $\mathcal{V} = [v_1, \ldots, v_m]^T$:

$$c^* = \min_{-\sigma_v \leq v_i \leq \sigma_v, i=1,\ldots,m} (Y - \mathcal{V})^T G (Y - \mathcal{V}). \quad (10)$$

The other parameter, B, can be formally bounded by the technique introduced in (Jackson et al. 2021a). This is in contrast with Lemma 1, which also requires an approximation on the information gain of the kernel. We should also emphasize that Theorem 1 allows σ_n to remain as a decision variable, enabling an optimization over σ_n that can further minimize the error bounds as compared to Lemmas 1 and 2.

We also note as $m \to \infty$ then λ_x tends toward 0, which implies that we can set δ arbitrarily close to 0, reducing the error bound to the result of Lemma 3, which decreases with the number of samples as $\sigma_D(x)$ decreases and c^* increases. This demonstrates an $O(\sqrt{m})$ improvement over the results of Lemma 2. A detailed discussion is provided in the Appendix of (Reed, Laurenti, and Lahijanian 2024).

We extend our point-wise probabilistic error bounds to a uniform bound with the following corollary.

Corollary 1 (Uniform Error Bounds)**.** *For a given compact set $X' \subseteq X$,*

$$\mathbb{P}\Big(|\mu_D(x) - f(x)| \leq \bar{\epsilon}_{X'}(\delta)\Big) \geq 1 - \delta \quad (11)$$

$\forall x \in X'$, where $\bar{\epsilon}_{X'}(\delta) = \sup_{x \in X'} \epsilon(\delta, x)$.

Deterministic Error Bounds

If error bounds with confidence one ($\delta = 0$) are desired, Theorem 1 would require infinite samples. Here, we show how confidence one results can be derived with finite samples, producing an alternative deterministic error bound to the one in Lemma 2.

Theorem 2 (Bounded Support Deterministic RKHS Error)**.** *Let X, B, D, G, c^* be as in Theorem 1, and $\Lambda_x = \sum_{i=1}^m \sigma_v |[K_{x,\mathcal{X}} G]_i|$. If noise $\mathbf{v} \in V$ has a finite support σ_v (i.e., $|\mathbf{v}| \leq \sigma_v$) and μ_D and σ_D are obtained via Eqns* (2)-(3) *on dataset D with any choice of $\sigma_n > 0$, then it holds that, for every $x \in X$,*

$$|\mu_D(x) - f(x)| \leq \epsilon_d(x) \quad (12)$$

where $\epsilon_d(x) = \sigma_D(x)\sqrt{B^2 - c^*} + \Lambda_x$.

Proof. Consider the term $|W_x \mathcal{V}|$ in Eqn (7). The noise distribution that maximizes $|W_x \mathcal{V}|$ is found by setting $v_i = \text{sign}([W_x]_i)\sigma_v$, where $\text{sign}([W_x]_i) = 1$ if the i-th term of W_x is ≥ 0 and 0 otherwise. Using this bound and Lemma 3 on the RHS of Eqn (7), we conclude the proof. □

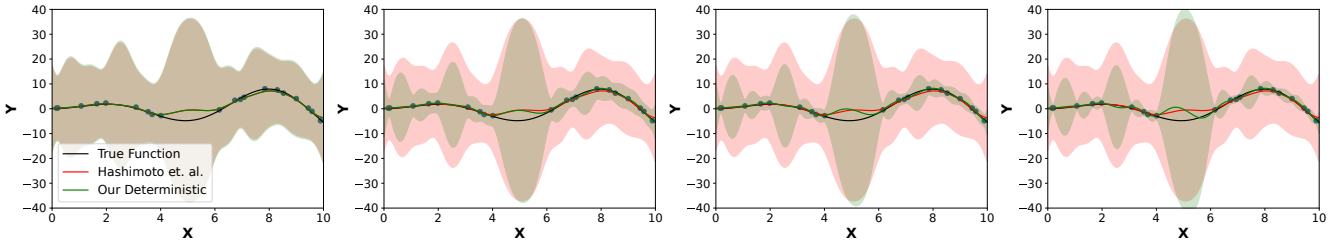
Remark 1. *We stress that while Lemmas 1 and 2 define strict requirements on the value of σ_n used in the posterior prediction, i.e., $1+2/m$ and σ_v respectively, both Theorems 1 and 2 allow for any value to be used for σ_n. In particular, $\sigma_n \ll \sigma_v$ is a valid selection which can reduce the predictive variance, enabling much tighter deterministic and probabilistic bounds. Similarly, when σ_v is very small, we can choose $\sigma_n > \sigma_v$ to avoid numerical instabilities in inverting $(K_{\mathcal{X},\mathcal{X}} + \sigma_n^2 I)$.*

We demonstrate the trend of the error bounds as we adjust σ_n used for the posterior predictions of a GP in Figure 2. It is immediately clear that using a smaller σ_n when there is more data results in tight error bounds across the domain.

Remark 2. *In this paper, we consider the offline setting, where we are given a set of i.i.d. data. In this setting, the bounds proposed in (Abbasi-Yadkori 2013; Srinivas et al. 2012; Chowdhury and Gopalan 2017) are still valid. To extend to online setting, our bounds can be updated as data is gathered, and the noise must remain independent of the filtration, which is typical in robotics applications.*

Extension to Deep Kernel Learning

Deep Kernel Learning (DKL) is an extension of GPR (Ober, Rasmussen, and van der Wilk 2021). In DKL, we compose a base kernel, e.g., the squared-exponential $\kappa_{se} = \sigma_s \exp\left(-\|x - x'\|/2l^2\right)$, with a neural network as $\kappa_{DKL}(x, x') = \kappa_{se}(\psi_w(x), \psi_w(x'))$, where $\psi_w : \mathbb{R}^d \to \mathbb{R}^s$ is a neural network parameterized by weights w. Then posterior predictions still use analytical Eqns (2)-(3), but the kernel now includes significantly more parameters which has been



(a) $\sigma_n = \sigma_v$ (b) $\sigma_n = \sigma_v/5$ (c) $\sigma_n = \sigma_v/10$ (d) $\sigma_n = \sigma_v/20$

Figure 2: Predictive mean and deterministic error bounds when learning from 20 samples of $y = x\sin(x) + v$ with $|v| < 0.5$ as σ_n varies. We plot the mean and bounds from Lemma (2) in red, with our mean and bounds in green. In all cases our bounds remain valid, but demonstrate optimal performance when $\sigma_n \in [\sigma_v/10, \sigma_v/5]$.

shown to significantly improve the representational power of GPs without needing more data in the kernel (Reed, Laurenti, and Lahijanian 2023).

Remark 3. *For DKL to satisfy Assumption 1, the kernel must remain continuously differentiable and positive semi-definite. Using the* GeLU *or* Tanh *activation functions and learning $\psi_w(x)$ as a model of $f(x)$ using stochastic mini-batching can prove sufficient. Then, under the assumption that ψ_w is well-behaved over a compact set X, the RKHS norm $\|f\|_{\kappa_{DKL}} \leq \|f\|_{\kappa}$ is feasible. This can be inferred directly from (Jackson et al. 2021a, Proposition 2), as DKL tends to correlate data more effectively over space, it is reasonable to expect that $\inf_{x,x' \in X} \kappa_{DKL}(x, x') \geq \inf_{x,x' \in X} \kappa(x, x')$.*

DKL can reduce the posterior variance of a GP and use the same RKHS norms as the base kernel, i.e., decreasing σ_D without increasing bound B. Therefore, DKL can enable tighter GP regression error bounds with the same number of predictive samples than standard GP. Similarly, in the event that many samples are available, network ψ_w can be pre-trained over a large set of samples and the kernel can use a small subset of the data for posterior predictions, enabling computationally efficient calculations with increased accuracy. Furthermore, compared to Lemma 1, our error bounds utilize significantly more information about the kernel. This allows an informed kernel, such as DKL, to further reduce the bound beyond just the value of σ_D.

4 Application to Safety of Control Systems

The bounds we derive in Theorems 1 and 2 can be particularly important to provide safety guarantees for dynamical systems learned from data. Specifically, consider the following model of a discrete-time stochastic system with additive noise

$$\mathbf{x}(k+1) = f(\mathbf{x}(k)) + \mathbf{v} \qquad k \in \mathbb{N}, \tag{13}$$

where $\mathbf{x} \in \mathbb{R}^d$ is the state, $\mathbf{v} \in V \subset \mathbb{R}^d$ is a random variable independent and identically distributed at each time step representing an additive noise term, and $f : \mathbb{R}^d \to \mathbb{R}^d$ is an unknown, possibly non-linear, function (vector field). Intuitively, $\mathbf{x}(k)$ represents a general model of a stochastic system taking values in $\mathbb{R}^d$.

A key challenge in many applications is to guarantee that System (13) will stay within a given safe set $X_s \subset \mathbb{R}^d$, i.e., it avoids obstacles, for a given time horizon $[0, N]$. A common technique to obtain such guarantees is to rely on stochastic barrier functions (Kushner 1967; Santoyo, Dutreix, and Coogan 2021), which represents a generalization of Lyapunov functions to provide set invariance. The intuition behind barriers is to build a function $\mathcal{B} : \mathbb{R}^d \to \mathbb{R}$ that when composed with System (13) forms a supermartingale and consequently, martingale inequalities can be used to study the systems' evolution, as shown in the following Proposition.

Proposition 1 ((Mazouz et al. 2022), Section 3.1)**.** *Given an initial set X_0, a safe set X_s, and an unsafe set $X_u = \mathbb{R}^d \setminus X_s$, a twice differentiable function $\mathcal{B} : \mathbb{R}^d \to \mathbb{R}$ is a barrier function if the following conditions hold for $\beta, \eta \in [0, 1]$: $\mathcal{B}(x) \geq 0$ for all $x \in \mathbb{R}^d$, $\mathcal{B}(x) \leq \eta$ for all $x \in X_0$, $\mathcal{B}(x) \geq 1$ for all $\forall x \in X_u$, and*

$$\mathbb{E}[\mathcal{B}(f(x) + \mathbf{v}) \mid x] \leq \mathcal{B}(x) + \beta \quad \forall x \in X_s. \tag{14}$$

Then, it holds that

$$P\big(\forall k \in [0, N], \mathbf{x}(k) \in X_s \mid \mathbf{x}(0) \in X_0\big) \geq 1 - (\beta N + \eta).$$

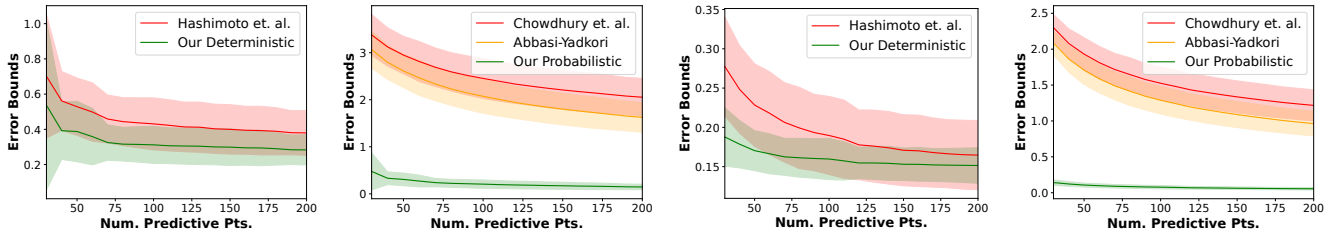
Hence, by finding an appropriate $\mathcal{B}$, safety of System (13) can be guaranteed. However, how to construct such $\mathcal{B}$ when f is unknown is still an open problem (Jagtap, Soudjani, and Zamani 2020; Wajid, Awan, and Zamani 2022; Mazouz et al. 2022, 2024). Here, we show that, under the assumption that f lies in the RKHS of κ and $\|f^{(i)}\|_\kappa < B_i$, where $f^{(i)}$ denotes the i-th output of f, we can employ GPR along with Theorems 1 and 2 to construct $\mathcal{B}$.

In particular, by partitioning X_s into a set of convex regions $Q = \{q_1, ..., q_{|Q|}\}$ s.t. $X_s = \cup_{q \in Q} q$, it is sufficient to simply replace the condition in (14) with the following constraints for each $q \in Q$:

$$\mathbb{E}[\mathcal{B}(z_q + \mathbf{v}) \mid x \in q] \leq \mathcal{B}(x) + \beta \tag{15a}$$

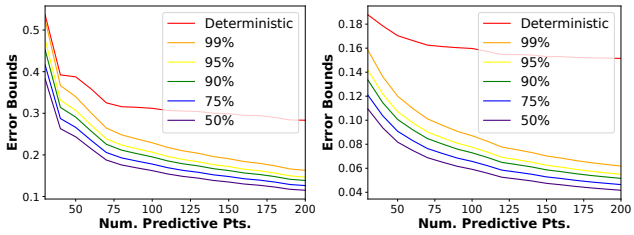
$$z_q^{(i)} \in [\mu_D^{(i)}(x) \pm \bar{\epsilon}_q^{(i)}(\delta)] \qquad \forall x \in q, \tag{15b}$$

where $\mu_D^{(i)}$ denotes the mean prediction for $f^{(i)}$ and $\bar{\epsilon}_q^{(i)}(\delta)$ can be computed by combining Corollary 1 with either Theorem 1 or 2. Work (Mazouz et al. 2024) shows how to calculate a barrier with such constraints. In the former case, the safety probability in Proposition 1 holds with the confidence resulting from Theorem 1, while in the latter case results hold with confidence 1. In Section 5, we consider both cases.



(a) GP - det. error comparison (b) GP - prob. error comparison (c) DKL - det. error comparison (d) DKL - prob. error comparison

Figure 3: Trends of error bounds with increasing data used for the posterior prediction for a 2D system when $|v| \leq 0.1$ using Left: the *squared exponential kernel* and Right: *DKL*. In (a) and (c) we compare our deterministic error bound (green) with results from Lemma 2 (red). In (b) and (d) we compare our probabilistic error bound (green) to results from Lemma 1 (red) and Abbasi-Yadkori (orange) with $\delta = 0.05$. In (c) and (f) we compare our deterministic and probabilistic bounds with $\delta \in [0.01, 0.5]$.



(a) GP - vary δ comparison (b) DKL - vary δ comparison

Figure 4: Trends of our error bounds with increasing data used for the posterior prediction for a 2D system when $|v| \leq 0.1$ using (a) the *SE kernel* and (b) *DKL*. We compare our deterministic (Theorem 2) and probabilistic (Theorem 1) bounds with $\delta \in [0.01, 0.5]$.

5 Experiments

In this section, we demonstrate the effectiveness of our bounds in comparison to the existing state-of-the-art bounds. We first provide a visual of the trends of our bounds as more data is utilized for posterior predictions, as well as demonstrating the improvements that DKL can provide over standard GPs. We then demonstrate how our bounds perform in comparison to the state-of-the-art across several dynamical systems of interest. Finally, we show how our bounds can be used in safety certification, through the use of stochastic barrier functions, as described in Section 4. We note that deep kernels in general are only positive semi-definite, which prohibits the use of techniques in (Maddalena, Scharnhorst, and Jones 2021) for generating a deterministic bound. Further detail on the models considered can be seen in the Appendix of (Reed, Laurenti, and Lahijanian 2024).

2D Visual Example We consider the case of GP regression for a linear system with $\mathbf{x} \in \mathbb{R}^2$ and $|v| \leq 0.1$, defined as

$$\mathbf{y} = 0.4x_1 + 0.1x_2 + v. \tag{16}$$

We first consider a squared exponential kernel and using the methods proposed in (Hashimoto et al. 2022) we set $B = 10$. For our predictions, we set $\sigma_n = \sigma_v/5 = 0.02$. We illustrate the trend of the bounds as the number of data points in the kernel increases in Figures 3a-3b, showing the averaged mean bound and one standard deviation over 10^4 test points. We note that our deterministic bound is comparable or better than the existing approach and our probabilistic bound significantly outperforms existing bounds.

Next, we consider the same scenario but with a DKL kernel. Results are shown in Figures 3c-3d. Here, we assume a fixed neural network that is pre-trained to model $f(x)$ from data. We consider a network with two hidden layers of 16 neurons each with the GeLU activation function. The network is trained over 1000 sample points through stochastic mini-batching. Figure 4 compares averages of our deterministic bound to our probabilistic bound. Interestingly, DKL generates probabilistic errors lower than the bound on sample noise, enabling highly accurate predictions even with very noisy samples. We note that, for each model, our probabilistic bound remain accurate for the desired confidence level.

Examples for Multiple Dynamical Systems We compare our bounds to existing bounds (Lemma 1-2) and bounds proposed by Abbasi-Yadkori; Srinivas et al. for a variety of dynamical systems from literature (Jackson et al. 2021b; Adams, Lahijanian, and Laurenti 2022; Reed, Laurenti, and Lahijanian 2023), namely a 2D linear model, a 2D non-linear model, a 3D Dubin's car model, and a 5D second order car model. We consider identical noise distributions in all dimensions; for the 2D and 3D systems we consider $\mathbf{v} \sim \text{Uniform}(-0.1, 0.1)$ and for the 5D system $\mathbf{v} \sim \text{Uniform}(-0.2, 0.2)$. Results are in Table 1.

We note that modifying the value of σ_n used for posterior predictions does not impact the accuracy of the model significantly but can be optimized to improve the error bounds. In all cases, we see that DKL (with a well trained neural network prior) significantly outperforms standard GP models when computing error bounds. We emphasize that the probabilistic bounds from (Abbasi-Yadkori 2013), Lemma 1, and (Srinivas et al. 2012) are significantly larger than ours for two primary reasons: (i) those bounds generalize to any conditional sub-Gaussian distribution while ours is specific to bounded support, and (ii) our bound incorporates significantly more information about the kernel, allowing an informed kernel to reduce the bound further than just the value of $\sigma_D(x)$. This results in our probabilistic bound being at least an order of magnitude smaller than prior works.

Safety Certification of Unknown Stochastic Systems We consider the Inverted Pendulum (2D) agent from the Gym-

System	κ	σ_n	Our Det.		Lem 2 Det.		Our Prob.		AY Prob.		Lem 1 Prob.		SKKS Prob.	
			true	$\|\epsilon\|_1$	true	$\|\epsilon\|_1$	true	$\|\epsilon\|_1$	true	$\|\epsilon\|_1$	true	$\|\epsilon\|_1$	true	$\|\epsilon\|_1$
2D Lin	SE	$\sigma_v/5$	0.06	0.86	0.04	1.45	0.06	0.57	0.10	5.72	0.10	7.53	0.04	350
2D Lin	SE	$\sigma_v/10$	0.07	0.84	0.04	1.45	0.07	0.50	0.10	5.72	0.10	7.53	0.04	350
2D Lin	DKL	$\sigma_v/5$	0.04	0.36	0.04	0.45	0.04	0.15	0.08	2.72	0.08	3.66	0.04	141
2D Lin	DKL	$\sigma_v/10$	0.04	**0.33**	0.04	0.45	0.04	**0.12**	0.08	2.72	0.08	3.66	0.04	141
2D NL	SE	$\sigma_v/5$	0.09	1.23	0.08	1.68	0.09	0.94	0.32	7.47	0.32	9.29	0.08	350
2D NL	SE	$\sigma_v/10$	0.11	1.31	0.08	1.68	0.11	0.90	0.32	7.47	0.32	9.29	0.08	350
2D NL	DKL	$\sigma_v/5$	0.04	0.39	0.04	0.53	0.04	0.22	0.19	3.79	0.19	4.75	0.04	130
2D NL	DKL	$\sigma_v/10$	0.04	**0.35**	0.04	0.53	0.04	**0.18**	0.19	3.79	0.19	4.75	0.04	130
3D DC	SE	$\sigma_v/5$	0.10	1.96	0.09	2.34	0.10	1.13	0.41	9.61	0.41	13.67	0.09	1071
3D DC	SE	$\sigma_v/10$	0.12	2.13	0.09	2.34	0.12	1.06	0.41	9.61	0.41	13.67	0.09	1071
3D DC	DKL	$\sigma_v/5$	0.03	0.44	0.03	0.49	0.03	0.12	0.29	2.82	0.29	4.32	0.03	195
3D DC	DKL	$\sigma_v/10$	0.03	**0.43**	0.03	0.49	0.03	**0.11**	0.29	2.82	0.29	4.32	0.03	195
5D Car	SE	$\sigma_v/5$	0.51	10.33	0.30	11.7	0.51	5.60	0.36	14.0	0.36	22.63	0.30	3475
5D Car	SE	$\sigma_v/10$	0.68	13.0	0.30	11.7	0.68	6.16	0.36	14.0	0.36	22.63	0.30	3475
5D Car	DKL	$\sigma_v/5$	0.06	1.48	0.06	1.54	0.06	0.46	0.25	4.34	0.25	6.82	0.06	581
5D Car	DKL	$\sigma_v/10$	0.06	**1.43**	0.06	1.54	0.06	**0.40**	0.25	4.34	0.25	6.82	0.06	581

Table 1: Average L_1 error bounds ($\|\epsilon\|_1$) over 10000 test points. We report the value for various values of σ_n and for the squared exponential kernel (SE) and for DKL. We report both the true error ($\|\mu - f\|_1$) induced by the model estimated empirically over 10^4 samples and the bounds produced by Theorem 2 (Our Det.), Lemma 2 (Lem 2 Det.), and probabilistic bounds from Theorem 1 (Our Prob.), Abbasi-Yadkori (AY Prob.), Lemma 1 (Lem 1 Prob.), and Srinivas et al. (SKKS Prob.) in order setting $\delta = 0.05$. Lemma 2 and (Srinivas et al. 2012) set $\sigma_n = \sigma_v$ and Lemma 1 and (Abbasi-Yadkori 2013) set $\sigma_n^2 = 1 + 2/m$.

Model	Our Prob.				Our Det.				Lem 2 Det.			
	η	β	P_s	t	η	β	P_s	t	η	β	P_s	t
Pendulum	1e-6	1e-6	0.999	136	1e-6	0.077	0.923	147	1e-6	0.499	0.499	2979
4D Linear	1e-6	1e-6	0.999	451	1e-6	0.172	0.827	2880	1e-6	0.249	0.749	13233

Table 2: Barrier results, where t is time in seconds taken to synthesize a barrier, $P_s = 1 - (\eta + N\beta)$, and $N = 1$.

nasium environment and a contractive linear 4D system for data-driven safety certification using the formulation in Section 4. We collected data for the Pendulum model under the best controller available from the OpenAI Leaderboard (OpenAI 2024) and perturb the systems with PERT distributions (a transformation of a Beta distribution with a specified mean). We construct a barrier using the Dual-Approach suggested in (Mazouz et al. 2024) and compare results when generating interval bounds on f with Corollary 1 using Theorem 2, Lemma 2, and Theorem 1 with $\delta = 0.05$ using DKL models.

The barriers identify a lower bound on the probability of the system remaining in a predefined safe set for a given horizon when initialized with $\theta, \dot{\theta} \in [-0.025, 0.025] \times [-0.055, 0.055]$ for the Pendulum and each state in $[-0.1, 0.1]$ for the 4D system. Barriers are synthesized on an Intel Core i7-12700K CPU at 3.60GHz with 32 GB of RAM. Results are reported in Table 2. Safety probabilities using bounds of Lemma 1 and (Abbasi-Yadkori 2013) are not reported in the table because they result in $P_s = 0$ after three time steps for both models.

Barrier certificates using Theorem 1 are based on a 95% confidence. We see that our bounds allow for a significant improvement in certification results as compared to existing bounds. For instance, for a time horizon $N = 10$, applying Proposition 1 with our bounds results in guarantees of 99.9% safety probability with 95% confidence, whereas using Lemma 2 results in a 0% safety probability for the Pendulum model. We see similar results for the 4D system, where the deterministic bounds remain too conservative to identify the contractive nature of the system yet our probabilistic bounds enable guarantees of safety with high confidence. We also note that the reduced conservatism of our approach enable barrier synthesis to be significantly faster.

6 Conclusion

In this paper, we derive novel error bounds (both probabilistic and deterministic) for GP regression under bounded support noise. We demonstrate that by assuming sample noise has zero mean, error bounds that are tighter than sample noise can be achieved. We show that our error bounds utilize significantly more information about the kernel than existing probabilistic bounds (i.e. each term informed by $K_{\mathcal{X},x}$); hence, they are well-suited for regression techniques based on informed kernels such as DKL that correlate x with the entire dataset. We also show the improvements that our bounds can provide over the state-of-the-art in safety certification of control systems through the use of stochastic barrier functions, generating certificates with significantly larger safety probabilities. Future directions include application of these bounds in safe control and shield design in reinforcement learning.

Acknowledgments

This work was supported by the Air Force Research Lab (AFRL) under agreement number FA9453-22-2-0050.

References

Abbasi-Yadkori, Y. 2013. Online learning for linearly parametrized control problems.

Adams, S. A.; Lahijanian, M.; and Laurenti, L. 2022. Formal Control Synthesis for Stochastic Neural Network Dynamic Models. *IEEE Control Systems Letters*.

Berkenkamp, F.; Turchetta, M.; Schoellig, A.; and Krause, A. 2017. Safe model-based reinforcement learning with stability guarantees. *Advances in neural information processing systems*, 30.

Chowdhury, S. R.; and Gopalan, A. 2017. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, 844–853. PMLR.

Griffioen, P.; Devonport, A.; and Arcak, M. 2023. Probabilistic Invariance for Gaussian Process State Space Models. In *Learning for Dynamics and Control Conference*, 458–468. PMLR.

Hashimoto, K.; Saoud, A.; Kishida, M.; Ushio, T.; and Dimarogonas, D. V. 2022. Learning-based symbolic abstractions for nonlinear control systems. *Automatica*, 146: 110646.

Jackson, J.; Laurenti, L.; Frew, E.; and Lahijanian, M. 2021a. Formal verification of unknown dynamical systems via Gaussian process regression. *arXiv preprint arXiv:2201.00655*.

Jackson, J.; Laurenti, L.; Frew, E.; and Lahijanian, M. 2021b. Strategy Synthesis for Partially-Known Switched Stochastic Systems. In *International Conference on Hybrid Systems: Computation and Control*, HSCC '21. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383394.

Jagtap, P.; Soudjani, S.; and Zamani, M. 2020. Formal synthesis of stochastic systems via control barrier certificates. *IEEE Transactions on Automatic Control*, 66(7): 3097–3110.

Jensen, B. S.; Nielsen, J. B.; and Larsen, J. 2013. Bounded gaussian process regression. In *2013 IEEE international workshop on machine learning for signal processing (MLSP)*, 1–6. IEEE.

Kushner, H. J. 1967. Stochastic stability and control.

Lederer, A.; and Hirche, S. 2019. Local asymptotic stability analysis and region of attraction estimation with gaussian processes. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, 1766–1771. IEEE.

Lederer, A.; Umlauft, J.; and Hirche, S. 2019. Uniform error bounds for gaussian process regression with application to safe control. *Advances in Neural Information Processing Systems*, 32.

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

Mazouz, R.; Baymler Mathiesen, F.; Laurenti, L.; and Lahijanian, M. 2024. Piecewise Stochastic Barrier Functions. *arXiv preprint arXiv:2404.16986*.

Mazouz, R.; Muvvala, K.; Ratheesh Babu, A.; Laurenti, L.; and Lahijanian, M. 2022. Safety guarantees for neural network dynamic systems via stochastic barrier functions. *Advances in Neural Information Processing Systems*, 35: 9672–9686.

Mohri, M.; Rostamizadeh, A.; and Talwalkar, A. 2018. *Foundations of machine learning*. MIT press.

Ober, S. W.; Rasmussen, C. E.; and van der Wilk, M. 2021. The promises and pitfalls of deep kernel learning. In *Uncertainty in Artificial Intelligence*, 1206–1216. PMLR.

OpenAI. 2024. Gym Leaderboard. https://github.com/openai/gym/wiki/Leaderboard. Accessed: 2024-05-10.

Rasmussen, C. E.; Williams, C. K.; et al. 2006. *Gaussian processes for machine learning*, volume 1. Springer.

Reed, R.; Laurenti, L.; and Lahijanian, M. 2023. Promises of deep kernel learning for control synthesis. *IEEE Control Systems Letters*.

Reed, R.; Laurenti, L.; and Lahijanian, M. 2024. Error Bounds For Gaussian Process Regression Under Bounded Support Noise With Applications To Safety Certification. *arXiv preprint arXiv:2408.09033*.

Santoyo, C.; Dutreix, M.; and Coogan, S. 2021. A Barrier Function Approach to Finite-Time Stochastic System Verification and Control. *Automatica*, 125: 109439.

Snelson, E.; Ghahramani, Z.; and Rasmussen, C. 2003. Warped gaussian processes. *Advances in neural information processing systems*, 16.

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

Steinwart, I. 2001. On the influence of the kernel on the consistency of support vector machines. *Journal of machine learning research*, 2(Nov): 67–93.

Wajid, R.; Awan, A. U.; and Zamani, M. 2022. Formal synthesis of safety controllers for unknown stochastic control systems using Gaussian process learning. In *Learning for Dynamics and Control Conference*, 624–636. PMLR.