Practical Global and Local Bounds in Gaussian Process Regression via Chaining

Junyi Liu[1], **Stanley Kok**[1]

[1]School of Computing, National University of Singapore, Singapore
liujunyi@u.nus.edu, skok@comp.nus.edu.sg

Abstract

Gaussian process regression (GPR) is a popular nonparametric Bayesian method that provides predictive uncertainty estimates and is widely used in safety-critical applications. While prior research has introduced various uncertainty bounds, most existing approaches require access to specific input features, and rely on posterior mean and variance estimates or the tuning of hyperparameters. These limitations hinder robustness and fail to capture the model's global behavior in expectation. To address these limitations, we propose a chaining-based framework for estimating upper and lower bounds on the expected extreme values over unseen data, without requiring access to specific input features. We provide kernel-specific refinements for commonly used kernels such as RBF and Matérn, in which our bounds are tighter than generic constructions. We further improve numerical tightness by avoiding analytical relaxations. In addition to global estimation, we also develop a novel method for local uncertainty quantification at specified inputs. This approach leverages chaining geometry through partition diameters, adapting to local structures without relying on posterior variance scaling. Our experimental results validate the theoretical findings and demonstrate that our method outperforms existing approaches on both synthetic and real-world datasets.

Code —
https://github.com/Liu-Jun-Yi/Chaining-Bounds-in-GPR
Extended version — http://arxiv.org/abs/2511.09144

Introduction

Gaussian process regression (GPR), a flexible nonparametric Bayesian method, has emerged as a powerful tool for learning-based control, offering predictive uncertainty estimates through its posterior distribution. Classical works such as Wu and Schaback (1993) and Schaback (1999) provide deterministic error bounds in noiseless settings using reproducing kernel Hilbert space (RKHS) theory and fill-distance metrics. In contrast, the frequentist literature focuses on uncertainty bounds under noise, including time-uniform bounds based on information gain and RKHS norms for sequential decision tasks (Srinivas et al. 2010, 2012; Chowdhury and Gopalan 2017; Whitehouse, Ramdas, and Wu 2023).

In safety-critical domains like autonomous driving and robotics (Berkenkamp 2019), rigorous uncertainty quantification is essential. Although GPR is often used in these settings, existing methods offer limited theoretical guarantees. One approach combines posterior variance with robust control to construct uncertainty intervals. To improve robustness under model mismatch, recent works introduce probabilistic bounds via Lipschitz assumptions (Lederer, Umlauft, and Hirche 2019), model-aware terms (Fiedler, Scherer, and Trimpe 2021), or hyperparameter variation (Capone, Lederer, and Hirche 2022). Others calibrate posterior errors or use conformal prediction (Capone, Hirche, and Pleiss 2023; Papadopoulos 2024).

Despite this progress, most existing approaches aim to control errors or construct uncertainty intervals at specified input locations. These bounds typically require access to specific test input features, and rely on posterior mean and variance or intensive hyperparameter tuning, limiting their adaptability and often resulting in conservative or unstable intervals. Moreover, they do not directly address global behavior, such as the expected maximum bound, which is critical for assessing model reliability in safety-critical settings. For instance, autonomous systems must ensure trajectory deviations remain within safety thresholds over time, while disaster and financial risk models focus on expected peaks, such as the highest flood level or the worst market loss.

To address these limitations, we propose a novel framework that directly controls the global behavior of Gaussian processes without requiring access to specific input features. Specifically, we derive an expected upper bound on the global maximum of the process, which, by symmetry, also provides a corresponding lower bound. To our knowledge, this is the first application of chaining-based bounds in the context of GPR.

In practical applications, loose bounds can lead to inefficient decisions and resource waste, motivating the need for tighter estimates. We tighten the chaining bounds by exploiting properties of kernels, such as RBF and Matérn, yielding provably tighter and more practical bounds. Our implementation avoids analytical relaxations, such as loose constant factors, that are typically introduced for mathematical convenience but often result in overly conservative estimates.

While our method captures global behavior, we also propose a complementary method to quantify uncertainty at

specific input features. Instead of relying on chaining accumulation, this method is inspired by chaining and leverages geometric partitioning to define locally adaptive reference sets. The uncertainty bounds are then computed by scaling the diameters of these partitions, aligning with the local geometry of the input space, and avoiding direct reliance on posterior mean and variance scaling.

Background

Gaussian Process Regression

Gaussian Process Regression (GPR) is a non-parametric Bayesian approach to regression (Williams and Rasmussen 2006). A Gaussian process (GP) is a collection of random variables $\{f(x)\}_{x\in T}$, where any finite subset follows a multivariate normal distribution. We write $f \sim \mathcal{GP}(m, K)$ to indicate that f is drawn from a GP with mean function $m(x)$ and covariance function $K(x, x')$. The covariance function defines dependencies between inputs and is typically chosen as the RBF or Matérn kernel. For simplicity, the mean function is often assumed zero.

Chaining

Chaining is a mathematical technique consisting of a succession of steps that provide successive approximations of an index space (T, d), where T is an index set or input set, and d is a metric on T. Its fundamental idea is to group variables X_t (or, equivalently, their corresponding indices) that are nearly identical and approximate them at successive levels of granularity (Talagrand 2014). By doing this, we achieve tighter bounds, especially in cases where many variables are similar (Asadi and Abbe 2020). This approach mitigates the risk of large errors that can arise from such correlations.

To illustrate, consider a stochastic process $(X_t)_{t\in T}$, in which the difference between X_n and X_0 is expressed as $X_n - X_0 = \sum_{t=1}^{n} (X_t - X_{t-1})$. When many variables X_t in T are nearly identical, strong correlations between them can obscure the true variation in the process. Grouping similar variables together helps reduce this redundancy by allowing us to approximate these highly correlated variables with a representative value, thereby simplifying the analysis and making the process easier to interpret and work with. A more detailed explanation can be found in Appendix A.1.

For $n \geq 0$, we select a subset T_n, and for each $t \in T$, we choose an approximation $\pi_n(t)$ from T_n. Using these $\{\pi_n(t)\}$ points, we obtain the corresponding $\{X_{\pi_n(t)}\}$ variables, which serve as successive approximations of $\{X_t\}$. We start by assuming that T_0 contains only one element t_0, and thus $\pi_0(t) = t_0$ for all $t \in T$. The core relation is: $X_t - X_{t_0} = \sum_{n\geq 1} \left(X_{\pi_n(t)} - X_{\pi_{n-1}(t)}\right)$.

This equality holds because, for sufficiently large n, $\pi_n(t)$ equals t, meaning that beyond a certain point, the approximation stops, and the series becomes a finite sum. Specifically, as n increases, the sets T_n become progressively finer, eventually covering all points in T. Once T_n contains t, we have $\pi_n(t) = t$, so no new information is added by further terms in the series. As a result, the infinite series truncates to a finite sum. This ensures convergence in practical settings where the process X_t is fully captured after a finite number of terms. The efficacy of this approach is rooted in the fact that for each approximation $\pi_i(t)$, the variables $X_t - X_{\pi_i(t)}$ are smaller than $X_t - X_{t_0}$, making their supremum easier to handle. We present a simple example in Appendix A.2.

This stepwise refinement converts the intractable global bound estimation into manageable local problems, simplifying the overall calculation, thus avoiding the complexity and error accumulation associated with global estimation.

Related Work

The concept of bounds in GPR originates from the Bayesian confidence intervals, which assume the target function is drawn from a Gaussian process prior and are widely used in traditional GPR setups to reflect posterior uncertainty.

In contrast, the frequentist literature develops high-probability uncertainty bounds, which do not assume a prior but instead provide guarantees over repeated sampling. For example, Srinivas et al. (2010), Srinivas et al. (2012), and Chowdhury and Gopalan (2017) derive time-uniform bounds on the prediction error by leveraging information gain and RKHS norms. These bounds are often used in Bayesian optimization and bandit settings. Similarly, Fiedler, Scherer, and Trimpe (2021) refine these bounds by introducing a model-misspecification-aware error term, still within a frequentist framework. A different class of results involves deterministic error bounds from the field of scattered data approximation. Classical works such as Wu and Schaback (1993) and Schaback (1999), as summarized in Wendland (2004), provide interpolation-based error bounds using fill distance and smoothness assumptions. These results are fundamentally non-probabilistic, rely on properties of the function space (e.g., RKHS regularity), and assume noiseless observations and the absence of distributional randomness.

Recent approaches adopt a Bayesian-style probabilistic setup. Lederer, Umlauft, and Hirche (2019) introduce probabilistic bounds using Lipschitz continuity, which are derived on a finite grid and extended via covering arguments. Although their presentation is probabilistic, it deviates from the strict frequentist setting of time-uniform guarantees. Similarly, Capone, Lederer, and Hirche (2022) design probabilistic error bounds based on a given hyperparameter range, aiming to mitigate the risk from inaccurate kernel choices.

Later, Song et al. (2019) propose distribution calibration by adjusting predictive distributions post-hoc using Gaussian processes and Beta calibration. Capone, Hirche, and Pleiss (2023) introduce a posterior adjustment that sharpens GP intervals via variance calibration but lacks distribution-free coverage. More recently, Papadopoulos (2024) develops conformal prediction to construct distribution-free intervals based on nonconformity scores.

These methods bound uncertainty pointwise but not global deviations. Our work advances GPR bounds in two directions. First, we implement Talagrand's generic chaining bound to estimate the global expected supremum error, capturing expected global deviation rather than input-specific guarantees. While existing frequentist bounds upper-bound this supremum by maximizing over uncertainty intervals at given inputs, our method offers a theoretical alternative that

targets the expected supremum. Second, we propose a new method for uncertainty quantification at specified inputs; Inspired by chaining geometry, it uses partitions to construct locally adaptive bounds around each input, reducing reliance on global posterior variance. Our methods jointly extend GPR bounds from local to global uncertainty control.

Chaining-based methods have recently gained attention in machine learning through Chaining Mutual Information (CMI) (Asadi, Abbe, and Verdú 2018), which has been used to bound generalization error (Clerico et al. 2022) and establish risk bounds for neural networks via hierarchical coverings (Asadi and Abbe 2020). For GPR, we instead apply chaining directly to covariance functions, which naturally encode the process structure.

Upper and Lower Bounds

We now present our technical contributions. We avoid posterior variance scaling compared to existing methods and apply chaining directly to the distance metric induced by the kernel function. While GPR typically fits data and selects kernel hyperparameters, we use the kernel solely to define a distance metric, bypassing posterior inference. This eliminates dependence on posterior variance and calibration, yielding tighter, more reliable bounds. We assume a bounded RKHS function and i.i.d. sub-Gaussian noise.

To formalize our analysis, we consider a Gaussian process $(X_t)_{t\in T}$, where each X_t follows a normal distribution with mean zero and variance σ^2, and T is an index set (e.g., $T \subseteq \mathbb{R}^n$). For any two points $s, t \in T$, the process satisfies $\mathbb{E}[(X_s - X_t)^2] = d(s,t)^2$ and tail bound $\mathbb{P}(|X_s - X_t| \geq u) \leq 2\exp\left(-\frac{u^2}{2d(s,t)^2}\right)$, where $d(s,t)$ is the canonical pseudometric induced by the covariance function. Under this setting, we study two objectives: (i) estimating an expected upper bound , and (ii) providing pointwise uncertainty bounds. We first introduce the expected upper bound, starting from Talagrand's theorem (Talagrand 2014).

Theorem 1. *(Talagrand 2014) (Eq 2.33) Let T be an index set, $t_0 \in T$ be an initial index, $T_n \subseteq T$ for $n \geq 0$, and $T_0 = \{t_0\}$. For each $t \in T$, let $\pi_n(t) \in T_n$ for each $n \geq 0$, where each $\pi_n(t)$ represents a successive approximation of t, and let $\pi_n(t) = t$ for sufficiently large n. Then*

$$P\left(\sup_{t\in T} |X_t - X_{t_0}| > uS\right) \leq L\exp\left(-\frac{u^2}{2}\right),$$

where L is a universal constant (e.g., $L = \exp(9/2)$ satisfies the condition), $u \in \mathbb{R}\cup\{0\}$, $d : T\times T \to \mathbb{R}$ is a distance metric on T, and

$$S := \sup_{t\in T} \sum_{n\geq 1} 2^{n/2} d(\pi_n(t), \pi_{n-1}(t)).$$

Theorem 1 is purely theoretical, lacking practical implementation details. For instance, no method is provided for determining S, $\{T_n\}_{n\geq 0}$, $\{\pi_n(t)\}_{n\geq 0}$, and t_0. We address some of these deficiencies in subsequent sections. A generic bound that applies to all kernels is presented below.

Theorem 2. *(Generic Bound) Theorem 1, combined with the formula $\mathbb{E}[Y] = \int_0^\infty P(Y \geq u)\, du$, which gives the expectation of a non-negative random variable, leads to the following upper bound:*

$$\mathbb{E}\sup_{t\in T} X_t \leq X_{t_0} + \mathbb{E}\left[\sup_{t\in T} |X_t - X_{t_0}|\right]$$
$$\leq X_{t_0} + (1+\sqrt{2})L \sup_{t\in T} \sum_{n\geq 0} 2^{n/2} d(t, T_n),$$

where $d(t, T_n) = \inf_{s\in T_n} \sqrt{K(t,t) + K(s,s) - 2K(t,s)}$ and t_0 is chosen such that X_{t_0} is close to zero due to the zero-mean property and the symmetry of the kernel.

Proof sketch.. We combine theorem 1 with the integral formula for non-negative expectations to get the upper bound $\mathbb{E}[\sup_{t\in T} |X_t - X_{t_0}|] \leq LS\sqrt{\pi/2}$. The approximation $d(t, \pi_n(t)) = \inf_{s\in T_n} d(t,s)$ and the triangle inequality control chaining increments. Detailed proofs of Theorems 1 and 2 are in Appendices B.1 and B.2. □

Unlike Talagrand's purely theoretical formulation, we provide a concrete and fully implementable version of the chaining procedure, including explicit constructions, pseudocode, and runnable code. In addition, rather than relying on arbitrary constants sufficient for theoretical validity, we explicitly control quantities such as prefactors (e.g., L) to obtain tighter and more practically useful bounds, thereby enhancing both rigor and applicability.

We further apply the general bounds to compute tighter bounds for specific kernels by deriving more precise estimates of $\mathbb{E}\left[\sup_{t\in T} |X_t - X_{t_0}|\right]$. The following subsections will first introduce the RBF and Matérn kernels, and then provide detailed proofs for their respective tighter bounds.

Kernels

In GPR, the distance between two input points is measured using a kernel, or covariance function, which quantifies their similarity in the feature space and defines the structure of the Gaussian process by controlling its smoothness and generalization ability. A widely used example is the radial basis function (RBF) kernel. It produces smooth, continuous estimates and is often combined with a constant kernel to model signal variance. It is defined as:

$$K(s,t) = \sigma^2 \exp\left(-\frac{\|s-t\|^2}{2l^2}\right),$$

where $\|s-t\|$ is the Euclidean distance between the (multi-dimensional) input points s and t, the σ^2 term represents the constant kernel, and l is the length-scale parameter.

The Matérn kernel is also a widely used covariance function that controls the smoothness of sampled functions through its parameter $\nu > 0$. It is defined as $K(s,t) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}\|s-t\|}{l}\right)^\nu B_\nu\left(\frac{\sqrt{2\nu}\|s-t\|}{l}\right)$, where l is the length scale, and B_ν is the modified Bessel function of the second kind. Larger values of ν imply smoother sample paths.

When $\nu = p + 1/2$ with $p \in \mathbb{N}$, the kernel simplifies to a product of an exponential and a polynomial of order p (Seeger 2004). Commonly $\nu = 3/2$, giving

$$K(s,t) = \left(1 + \frac{\sqrt{3}\|s-t\|}{l}\right)\exp\left(-\frac{\sqrt{3}\|s-t\|}{l}\right).$$

The distance between two points s and t in the context of GPs is $d(s,t) = \sqrt{\mathbb{E}[(X_s - X_t)^2]}$, where X_s and X_t are the values at points s and t respectively. This distance metric is derived from the covariance function $K(s,t)$, which describes the covariance between the random variables X_s and X_t. Specifically, it can be expanded as:

$$d(s,t)^2 = \mathbb{E}[(X_s - X_t)^2] = K(s,s) + K(t,t) - 2K(s,t).$$

It is worth noting that s and t can each represent a vector describing a (multi-dimensional) input in a feature space, with X_s and X_t corresponding to the outputs evaluated at those input vectors. In this case, the covariance function $K(s,t)$ reflects how similar the outputs are given their respective input vectors s and t.

Tighter Bounds

We will now show in Theorem 5 how to refine the upper bound on $\mathbb{E}\left[\sup_{t\in T} |X_t - X_{t_0}|\right]$, yielding a tighter and more practical result for Gaussian processes with RBF kernels, compared to the bound in Theorem 2.

Theorem 3. *(Tighter RBF Bound) Consider a Gaussian process $(X_t)_{t\in T}$ with a radial basis function (RBF) kernel $K(s,t) = \sigma^2 \exp\left(-\frac{\|s-t\|^2}{2l^2}\right)$, where T is an input/index set, $\|s-t\|$ is the Euclidean distance between input points $s \in T$ and $t \in T$, the term σ^2 represents the constant kernel, and l represents the length-scale parameter. Let $t_0 \in T$ be an initial point, and $(T_n)_{n\geq 0}$ be a sequence such that $T_n \subseteq T$. In addition, for each $t \in T$, let $\{\pi_n(t) \in T_n\}_{n\geq 0}$ represent a chain of successive approximations of t such that $X_t - X_{t_0} = \sum_{n\geq 1}\left(X_{\pi_n(t)} - X_{\pi_{n-1}(t)}\right)$ with the condition that $\pi_n(t) = t$ for sufficiently large n and $\pi_0(t) = t_0$. Then*

$$\mathbb{E}\sup_{t\in T} |X_t - X_{t_0}| \leq (1+\sqrt{2})L \sup_{t\in T}\sum_{n\geq 0} 2^{n/2} d'(t, T_n),$$

where $d'(t,T_n) = \inf_{s\in T_n}\sqrt{K(t,t) + K(s,s) - 2\sigma K^{\frac{1}{2}}(t,s)}$.

Proof sketch.. We use the kernel-induced distance $d(s,t)^2 = 2\sigma^2(1 - \exp(-\|s-t\|^2/2l^2))$, and apply the inequality $\|s-t\|^2 + \|t-u\|^2 \geq \|s-u\|^2/2$ with the convexity of the exponential to get $d(s,u)^2 \leq d'(s,t)^2 + d'(t,u)^2$, where $d'(s,t)^2 = 2\sigma^2 - 2\sigma K^{1/2}(s,t)$. We use the approximation and triangle inequality. A detailed proof is given in Appendix B.3. □

While the RBF kernel is widely used, other kernels, such as the Matérn kernel, are better suited for specific applications. In the following, we prove the upper and lower bounds for the Matérn kernel with $\nu = 3/2$.

Theorem 4. *(Tighter Matérn Bound) Consider a Gaussian process $(X_t)_{t\in T}$ with a Matérn kernel $K(s,t) = \left(1 + \frac{\sqrt{3}\|s-t\|}{l}\right)\exp\left(-\frac{\sqrt{3}\|s-t\|}{l}\right)$, where T is an input/index set, $\|s-t\|$ is the Euclidean distance between input points $s \in T$ and $t \in T$, and l is the length-scale parameter. Then*

$$\mathbb{E}\sup_{t\in T}|X_t - X_{t_0}| \leq (1+\sqrt{2})L\sup_{t\in T}\sum_{n\geq 0} 2^{\frac{n}{2}}[d''(t,T_n) + \sqrt{2} - 2],$$

where $d''(t,T_n)) = \inf_{s\in T_n}\sqrt{K(t,t) + K(s,s) - 2K'(t,s)}$, *and* $K'(s,t) = \left(1 + \frac{\sqrt{3}\|s-t\|}{l}\right)\left[\exp\left(-\frac{\sqrt{3}\|s-t\|}{l}\right) - \frac{1}{2}\right]$.

Proof sketch.. By applying Chebyshev's sum inequality to the sequences $(1+x_i)$ and $\exp(-x_i)$ with $x_i = \frac{\sqrt{3}\|\cdot\|}{l}$, and leveraging the monotonicity of the function $f(x) = (1+x)\exp(-x)$, we establish that the kernel satisfies a relaxed subadditivity condition $d(s,u)^2 \leq d'(s,t)^2 + d'(t,u)^2 - 2$, where $d'(s,t)^2 = 2 - 2K'(s,t)$. We then follow the approximation and triangle inequality. A detailed proof is provided in Appendix B.4. □

By leveraging the kernel-dependent bounds in Theorems 3 and 4, we provide both broadly applicable results and tighter, more practical bounds for Gaussian processes with commonly used kernels such as Matérn and RBF.

However, beyond kernel-specific flexibility, we also address a more fundamental limitation in classical chaining theory. Generic chaining is not optimized for sharp constants, resulting in loose conservative numerical bounds. In particular, classical results (e.g., the proof in Appendix B.1) often include a large prefactor $L = \exp(9/2)$.

To obtain tighter estimates, we replace the constant L by directly integrating the tail bound derived from chaining. Since the failure probability $p(u)$ is defined via a union bound over rare events and satisfies $p(u) \leq 1$, we truncate the integrand by $\min(p(u), 1)$, yielding a sharper estimate.

Theorem 5. *The expected supremum satisfies:*

$$\mathbb{E}\sup_{t\in T} X_t \leq X_{t_0} + \mathbb{E}\left[\sup_{t\in T}|X_t - X_{t_0}|\right]$$
$$\leq X_{t_0} + S\int_0^\infty \min\left(\sum_{n\geq 1} 2^{2^{n+1}+1}\exp(-v^2 2^{n-1}),\, 1\right) dv,$$

where $S \leq (1+\sqrt{2})\sup_{t\in T}\sum_{n\geq 0} 2^{n/2} d'(t,s)$ *for the RBF kernel in Theorem 3;* $S \leq (1+\sqrt{2})\sup_{t\in T}\sum_{n\geq 0} 2^{n/2} d''(t,s)$ *for the Matérn kernel in Theorem 4.*

Proof sketch.. The Gaussian tail bound is applied to the increment $X_{\pi_n(t)} - X_{\pi_{n-1}(t)}$, bounding its tail probability by $\exp(-u^2 2^{n-1})$, where the threshold is $u\, 2^{n/2} d(\pi_n(t), \pi_{n-1}(t))$. A union bound $p(u)$ is used for the failure probability of the event Ω_u^c. To avoid introducing large analytic prefactors L. We note that $p(u) \leq 1$ to get $\mathbb{E}[\sup_{t\in T}|X_t - X_{t_0}|] \leq \int_0^\infty \min(p(u/S), 1)\, du$,, where S is bounded via theorems 3 and 4. A detailed proof is provided in Appendix B.5. □

Uncertainty Bounds

In the application of chaining methods for uncertainty quantification in Gaussian process regression, relying solely on the expected supremum upper bound is often insufficient. It is also necessary to evaluate the probability of the supremum *exceeding a given threshold*. To address this, we propose a method for deriving uncertainty bounds.

When incorporating test set features to predict pointwise bounds, the objective shifts from inferring the global supremum (and infimum) to estimating bounds for a specific test point, t'. Symmetric intervals around a fixed reference (e.g., $X_{t_0} \approx 0$) may lack tightness, as the uncertainty should instead be symmetric around $X_{t'}$. To address this, we focus on $X_t - X_{\pi_n(t)}$, where $X_{\pi_n(t)}$ is a reference point of X_t in T_n, yielding tighter and more precise bounds. Such bounds are useful when we wish to bound the uncertainty around the predicted value of an *individual* test point.

Theorem 6. *Let T be an index set, $t_0 \in T$ be an initial index, $T_n \subseteq T$ for $n \geq 0$, and $T_0 = \{t_0\}$. For each $t \in T$, let $\pi_n(t) \in T_n$ for each $n \geq 0$, where each $\pi_n(t)$ represents a successive approximation of t, and let $\pi_n(t) = t$ for sufficiently large n. Then*

$$P\left(|X_t - X_{\pi_n(t)}| < \sqrt{2(\ln 2 + \ln \frac{1}{\delta})}d(t, T_n)\right) > 1 - \delta,$$

where $d(t, T_n) = \inf_{s \in T_n} \sqrt{K(t,t) + K(s,s) - 2K(t,s)}$.

Proof sketch.. By introducing the confidence parameter $\delta = 2\exp(-u^2 2^{n-1})$ and taking the natural logarithm of both sides, we have $\ln(\delta) = \ln(2) - u^2 2^{n-1}$. Rearranging for u^2, we get $u = \sqrt{\frac{\ln(2)+\ln(1/\delta)}{2^{n-1}}}$. and substituting this expression for u into the inequality for $|X_t - X_{\pi_n(t)}|$ gives the result. A detailed proof is in Appendix B.6. □

In the chaining method, it is challenging to construct the sets T_n. Talagrand (2014) reformulates this as a geometric problem, replacing $d(t, T_n)$ with partition diameters. This approach replaces the inherently combinatorial nature of supremum computations with geometric quantities that are easier to control and analyze. Partition diameters, as global quantities, reduce computational complexity to a manageable recursive process and provide better theoretical control.

To formalize this approach, an admissible sequence $(A_n)_{n \geq 0}$ is defined as an increasing sequence of partitions of T, where $A_0 = \{T\}$, and $|(A_n)| \leq 2^{2^n}$ for $n \geq 1$. Each $A_n(t)$ denotes the unique set in A_n containing $t \in T$.

From each partition A_n, a representative point is selected from every set $A \in A_n$ to construct the subset $T_n \subseteq T$. By construction, for any $t \in T$ and $n \geq 0$, the distance between t and T_n satisfies $d(t, T_n) \leq \Delta(A_n(t))$, where $\Delta(A_n(t))$ represents the diameter of $A_n(t)$ under the metric d. Thus Theorem 6 can be expressed as:

$$P\left(|X_t - X_{\pi_n(t)}| < \sqrt{2(\ln 2 + \ln \frac{1}{\delta})}\Delta(A_n(t))\right) > 1 - \delta.$$

This method retains the advantages of chaining without relying on posterior inference, providing tighter uncertainty pointwise bounds than global expected bounds.

Algorithm 1: Chaining Bounds Method

Input: Kernel function $K(s,t)$ and dataset $D \coloneqq \{(t, X_t)\}$, where $t \in \mathbb{R}^d$ is a d-dimensional input/index vector, and $X_t \in \mathbb{R}$ is its associated output value.

Output: B, a set containing the upper and lower bounds for each test example.

Split D into a training set D_{train} and a test set D_{test}.
Fit a Gaussian process using the kernel function $K(s,t)$ to the training data D_{train}.
Initialize $T_0 \leftarrow \{t_0\}$, $T \leftarrow \{t : (t, \cdot) \in D_{\text{train}}\}$, and $n_{\max} \leftarrow \lfloor \log_2(\log_2(|T|)) \rfloor$.
for $n = 1$ to $n_{\max}$ **do**
  $T_n \leftarrow T_{n-1}$
  while $|T_n| < N_n = \min(2^{2^n}, |T|)$ **do**
    $T_n \leftarrow T_n \cup \{\arg\max_{t \in D_{\text{train}}} \min_{t^* \in T_n} d(t, t^*)\}$
  end while
  $A_{n,k} = \{t_i \in D_{\text{train}} \mid k = \arg\min_j d(t_i, t_j), t_j \in T_n\}, \quad k = 1, \ldots, N_n$.
end for
Initialize bounds $B \leftarrow \emptyset$.
for $t \in D_{\text{test}}$ **do**
  Compute $\mathbb{E}\sup_t |X_t - X_{t_0}|$ using Theorem 3 (RBF kernel) or Theorem 4 (Matérn kernel).
  Find $k^* = \arg\min_k \min_{t_i \in A_{n_{\max},k}} d(t, t_i)$.
  $\mu_{A_{n_{\max},k^*}} = \frac{1}{|A_{n_{\max},k^*}|} \sum_{t_i \in A_{L,k^*}} X_{t_i}$
  $\Delta(A_{n_{\max},k^*}) = \max_{t_i, t_j \in A_{n_{\max},k^*}} d(t_i, t_j)$
  Compute uncertainty bound $u(t) = \Delta(A_{n_{\max},k^*}) \cdot \sqrt{2 \cdot (\log(1/\delta) + \log(2))}$.
  $B \leftarrow B \cup \{(\mu_{A_{n_{\max},k^*}} + u(t), \mu_{A_{n_{\max},k^*}} - u(t))\}$ for uncertainty bounds (Theorem 6) or $B \leftarrow B \cup \{(X_{t_0} + S\int_0^\infty \min(\sum_{n \geq 1} 2^{2^{n+1}+1} \exp(-v^2 2^{n-1}), 1)\, dv, X_{t_0} - S\int_0^\infty \min(\sum_{n \geq 1} 2^{2^{n+1}+1} \exp(-v^2 2^{n-1}), 1)\, dv)\}$ for unseen test points' bound (Theorem 5).
end for
Return B.

Algorithm of Our Chaining Method

In this work, we convert theoretical constructs into a practical chaining method for calculating the uncertainty bounds of GPR. The full procedure is detailed in Algorithm 1.

First, we preprocess the data by randomly dividing it into training and test sets. Then, we calculate the average of the output values (labels), and center the training set by subtracting the average from the output values of each example (now their mean is 0). Similarly, we subtract this training average value from the test set. Next, we fit a Gaussian process (GP) to the training data via maximum likelihood estimation to learn the parameters of the GP's kernel function and ensure that the kernel effectively models the data distribution.

Next, we construct the sequence of partitions A_n to progressively refine the training set. At each level n, a representative set T_n is constructed by iteratively selecting points that maximize their minimum distance to the current representatives, ensuring that $\sup_{t \in D_{\text{train}}} d(t, T_n)$ is minimized. Each training point is then assigned to the closest represen-

tative in T_n, forming the partitions $A_n = \{A_{n,k}\}$, where $A_{n,k}$ contains all points nearest to the k-th representative.

We control the size of the set T_n by using the condition $|T_n| \leq N_n$, where $N_0 = 1$ and $N_n = 2^{2^n}$ for $n \geq 1$. This assumption leverages the approximation $\sqrt{\log N_n} \approx 2^{n/2}$, which is a critical component in our analysis, and is related to the term $\exp(-x^2)$, which governs the tails of a Gaussian distribution. Furthermore, the inequality $N_n^2 \leq N_{n+1}$ demonstrates the effectiveness of this sequence in controlling the size of the sets T_n (Talagrand 2014).

Finally, we compute the distances between the test points and the representatives in T_n to determine their closest subsets in A_n. For a test point t, the mean of the training targets in its subset is used to compute a central prediction. The uncertainty bounds are derived by scaling the diameter of the subset with a factor proportional to $\mathbb{E}[\sup_{t \in T} X_t]$ using Theorem 5 (for the upper and lower bounds over all unseen points) or $\sqrt{2(\log(1/\delta) + \log(2))}\Delta(A_n(t))$ using Theorem 6 (for uncertainty quantification at specified inputs). Computational complexity and time costs are provided in Appendix C.1.

Experiment

Datasets

The performance is evaluated on a synthetic dataset and five benchmark datasets that are widely used in prior studies:

- **Synthetic Data** consists of 50 random functions sampled from a RKHS over $D = [-1, 1]$ and evaluated at 1000 points. Each function combines kernels centered at random points, with 50 noisy samples drawn (Gaussian noise, standard deviation 0.5).
- **Boston Housing** (Cournapeau et al. 2007) contains 506 samples, each with 13 features (e.g., crime rates and pollution) predicting the median house price.
- **Sarcos** (Schaal 2009) consists of 44,484 training and 4,449 test samples, with each sample containing 21 input features from a seven-degree-of-freedom robotic arm. The task is to predict torque at seven joints.
- **USGS Earthquake** (Survey 2024) contains thousands of observations on earthquake occurrences, detailing the time, location, and magnitude of significant seismic events recorded by the U.S. Geological Survey.
- **Loa_CO2** (Laboratories 2024) contains CO2 concentration measurements from the Mauna Loa Observatory in Hawaii. Its inputs include date and CO2 concentration.
- **Auto-mpg** (Dua and Graff 2017) is a dataset focused on fuel consumption measured in miles per gallon (MPG). The original dataset consists of 398 observations, with 6 missing values discarded. It includes 7 input features such as engine capacity and number of cylinders.

Evaluation Metrics

The performance of our proposed approach is evaluated using standard metrics for prediction intervals, as described by Khosravi et al. (2010).

- **Prediction Interval Coverage Probability (PICP)**. This metric evaluates the percentage of test observations that lie within the bounds of the prediction intervals (PIs) at a given confidence level $(1 - \alpha)\%$. It is calculated as $\text{PICP} = \frac{1}{n}\sum_{i=1}^{n} c_i$, where $c_i = 1$ if the output at point i lies within the bounds $[L(X_i), U(X_i)]$, and $c_i = 0$ otherwise. Here, $L(X_i)$ and $U(X_i)$ denote the lower and upper bounds of the i^{th} PI.
- **Normalized Mean Prediction Interval Width (NMPIW)**. PIs that are too wide provide little useful information, so the NMPIW metric quantifies the width of the PIs as $\text{NMPIW} = \frac{\frac{1}{n}\sum_{i=1}^{n}(U(X_i) - L(X_i))}{R}$, where R is the range of the target variable. NMPIW expresses the average PI width as a percentage of the target range.
- **Coverage Width-Based Criterion (CWC)**. This is the *primary* evaluation metric because it balances the conflicting goals of achieving narrow PIs (low NMPIW) and high coverage (high PICP). (Note that a good PICP score can be trivially achieved at the expense of NMPIW (by using overly wide PIs) and vice versa (by using overly narrow PIs). Hence, either PICP or NMPIW alone is insufficient to completely reflect the goodness of bounds.) CWC is defined as $\text{CWC} = \text{NMPIW}\left(1 + \gamma(\text{PICP})e^{-\eta(PICP-\mu)}\right)$, where γ is a hyperparameter and $\eta = 50$ to penalize narrow intervals; and μ represents the nominal confidence level ($\mu = 1$ for extremal bounds). When PICP $\geq \mu$, $\gamma = 0$; otherwise, $\gamma = 1$.

Baseline Settings

We compare our chaining method to the following three state-of-the-art baselines, previously introduced in the related-work section: (i) **Lederer19** (Lederer, Umlauft, and Hirche 2019), which introduces probabilistic Lipschitz constants to reduce the reliance on prior knowledge, estimates errors on a finite grid, and extends them to the input space; (ii) **Fiedler21** (Fiedler, Scherer, and Trimpe 2021), which modifies its objective bound function by introducing an error term based on the work of (Chowdhury and Gopalan 2017); (iii) **Capone22** (Capone, Lederer, and Hirche 2022), which tackles hyperparameter misspecification by proposing a method to calculate error bounds across a given range of hyperparameters; and (iv) **Bayesian CI**, which uses the standard GP posterior mean and standard deviation to form Bayesian credible intervals $\mu(x) \pm z \cdot \sigma(x)$.

The baselines are evaluated on two tasks: (1) pointwise uncertainty estimation using per-point confidence intervals, and (2) estimation of upper and lower bounds for the expected maximum and minimum values on unseen data. We set δ to the confidence level in task (1), and to 0.0001 in task (2) to approximate the ideal case $\delta \to 0$ required for uniform bounds. Bayesian CI is only evaluated on task (1), as it requires a fixed test point x to provide $P(f(x) \in [\mu(x) \pm z \cdot \sigma(x)]) \geq 1 - \delta$, while task (2) requires a uniform guarantee for $\forall x \in \mathcal{X}$, which the others explicitly provide.

For Lederer19, we use the default implementation with 10000 grid points over a fixed domain and resolution $\tau = 10^{-8}$. For Fiedler21, we compare the default RKHS norm

$B = 2$ with a data-driven estimate $y^\top (K + \lambda I)^{-1} y$, using the better one, and fix the noise level to 0.0001. For Capone22, we use the provided hyperparameter grids when available, or adopt ranges from similar datasets otherwise. All methods use the RBF kernel as in their original versions.

Results

We perform experiments on the two tasks defined above. Table 1 compares the performance of our method with baseline methods in terms of CWC values at the 99%, 95%, and 90% confidence levels for conventional prediction on each test point. Our methods consistently achieve the lowest CWC values, demonstrating superior coverage while providing compact intervals. Complete results, including PICP, NMPIW, and CWC values, are provided in Appendix C.2. For the second experiment, Table 2 presents the results at the expected upper and lower bounds. Complete results, including PICP, NMPIW, and CWC values, are provided in Appendix C.3. From Table 2, it can be observed that our method produces the tightest bounds and has the best (lowest) CWC values in 5 of the 6 datasets, and is a close second on the last one.

Method	**Synthetic**			**Boston**		
	99%	95%	90%	99%	95%	90%
RBF (Ours)	1.80	1.50	1.35	1.01	0.84	0.76
Matérn (Ours)	**1.80**	1.50	**1.35**	**0.76**	**0.64**	**0.57**
Capone22	5.30	7.02	2.63	1.30	1.28	1.18
Fiedler21	3.95	**1.49**	1.49	3.46	3.46	3.46
Lederer19	3.63	3.56	3.53	1.47	1.35	1.41
Bayesian CI	37.19	69.02	113.08	5.95	4.94	2.66
Method	**Sarcos**			**USGS_EQ**		
	99%	95%	90%	99%	95%	90%
RBF (Ours)	0.48	0.40	0.36	**2.69**	2.25	2.03
Matérn (Ours)	**0.40**	**0.33**	**0.30**	2.76	**2.25**	**2.03**
Capone22	0.53	0.51	0.50	8.41	8.41	8.41
Fiedler21	1.42	1.42	1.42	3.26	3.26	3.26
Lederer19	0.56	0.50	0.49	4.23	4.13	4.07
Bayesian CI	2.03	3.44	3.22	6.13	3.00	2.82
Method	**Loa_CO2**			**Auto-mpg**		
	99%	95%	90%	99%	95%	90%
RBF (Ours)	0.81	0.68	0.61	1.09	0.91	0.82
Matérn (Ours)	**0.24**	**0.20**	**0.18**	**0.84**	**0.70**	**0.63**
Capone22	1761.85	239.41	20.68	6.85	3.23	1.95
Fiedler21	3.71	3.71	3.71	1.39	1.39	1.39
Lederer19	0.55	0.53	0.52	50.03	48.42	47.70
Bayesian CI	7.31	3.30	1.54	3.51	2.34	0.93

Table 1: Comparison of CWC at 99%, 95%, and 90% confidence levels across six datasets.

We illustrate the 99% uncertainty bounds in Figure 1 (large images are provided in Appendix C.4), which corresponds to the results shown in Table 1. In all figures, our method achieves over 99% coverage (with all black test points within bounds) and consistently produces tighter intervals, indicating its superior performance compared to all baselines. Computational cost results are reported in Appendix C.1. Our method achieves competitive computational efficiency and strong scalability across datasets. Statistical significance results can be found in Appendix C.5. In the vast majority of cases, our method demonstrates statistically significant improvements.

	Synthetic	**Boston**	**Sarcos**
RBF(Ours)	1.68	1.75	1.03
Matérn(Ours)	**1.67**	**1.64**	**0.78**
Capone22	4.89	1.77	1.18
Fiedler21	2.20	5.04	2.31
Lederer19	2.02	1.78	1.34
	USGS EQ	**Loa_CO2**	**Auto-mpg**
RBF(Ours)	2.59	**1.70**	3.06
Matérn(Ours)	**2.56**	2.08	3.24
Capone22	4.62	2.07	**2.81**
Fiedler21	2.57	16.01	7.16
Lederer19	3.07	1.73	57.22

Table 2: Comparison of CWC across synthetic and real-world datasets of the expected upper and lower bounds over unseen points only using the training set.

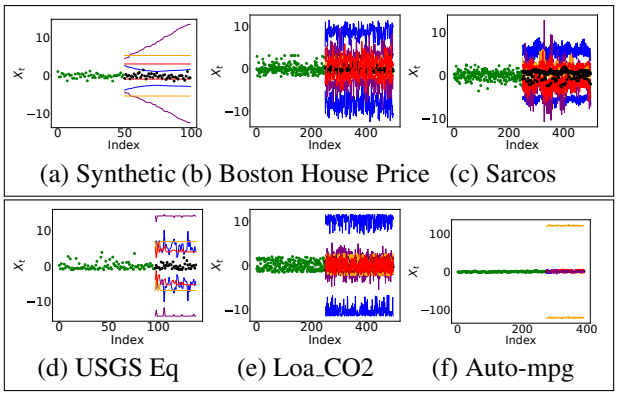

Figure 1: Comparison of our method with baselines for the test-point-specific bounds. The training set is in green, the test set in black, Lederer19 in orange, Fiedler21 in blue, Capone22 in purple, and our method in red.

Conclusion

Our work addresses the limitations of existing methods by introducing a novel chaining-based approach that improves error control and robustness. Leveraging Talagrand's techniques (Talagrand 2014), we derive more flexible and accurate prediction bounds without relying on posterior means or variances. Our method not only yields conventional uncertainty bounds but also estimates the expected value and variance of the supremum using only the training set. Furthermore, it provides tighter bounds for commonly used kernels such as RBF and Matérn. Future work includes case analyses, point-selection strategies, and broader related work.

Acknowledgments

We thank all anonymous reviewers for their valuable feedback and constructive suggestions.

References

Asadi, A.; Abbe, E.; and Verdú, S. 2018. Chaining mutual information and tightening generalization bounds. *Advances in Neural Information Processing Systems*, 31.

Asadi, A. R.; and Abbe, E. 2020. Chaining meets chain rule: Multilevel entropic regularization and training of neural networks. *Journal of Machine Learning Research*, 21(139): 1–32.

Berkenkamp, F. 2019. *Safe exploration in reinforcement learning: Theory and applications in robotics*. Ph.D. thesis, ETH Zurich.

Capone, A.; Hirche, S.; and Pleiss, G. 2023. Sharp calibrated gaussian processes. *Advances in Neural Information Processing Systems*, 36: 36579–36590.

Capone, A.; Lederer, A.; and Hirche, S. 2022. Gaussian process uniform error bounds with unknown hyperparameters for safety-critical applications. In *International Conference on Machine Learning*, 2609–2624. PMLR.

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

Clerico, E.; Shidani, A.; Deligiannidis, G.; and Doucet, A. 2022. Chained generalisation bounds. In *Conference on Learning Theory*, 4212–4257. PMLR.

Cournapeau, D.; Brucher, M.; Perrot, M.; and Duchesnay, E. 2007. Scikit-learn: Machine learning in Python. *JMLR*, 12: 2825–2830.

Dua, D.; and Graff, C. 2017. UCI Machine Learning Repository: Auto MPG Data Set.

Fiedler, C.; Scherer, C. W.; and Trimpe, S. 2021. Practical and rigorous uncertainty bounds for Gaussian process regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 7439–7447.

Khosravi, A.; Nahavandi, S.; Creighton, D.; and Atiya, A. F. 2010. Lower upper bound estimation method for construction of neural network-based prediction intervals. *IEEE transactions on neural networks*, 22(3): 337–346.

Laboratories, N. E. S. R. 2024. Mauna Loa CO2 Record. https://gml.noaa.gov/ccgg/trends/. Accessed: 2025-02-22.

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.

Papadopoulos, H. 2024. Guaranteed coverage prediction intervals with Gaussian process regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Schaal, S. 2009. The SL simulation and real-time control software package. Technical report, Citeseer.

Schaback, R. 1999. Improved error bounds for scattered data interpolation by radial basis functions. *Mathematics of Computation*, 201–216.

Seeger, M. 2004. Gaussian processes for machine learning. *International journal of neural systems*, 14(02): 69–106.

Song, H.; Diethe, T.; Kull, M.; and Flach, P. 2019. Distribution calibration for regression. In *International Conference on Machine Learning*, 5897–5906. PMLR.

Srinivas, N.; Krause, A.; Kakade, S.; and Seeger, M. 2010. Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. In *Proceedings of the 27th International Conference on Machine Learning*, 1015–1022. Omnipress.

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.

Survey, U. G. 2024. USGS Earthquake Data. https://earthquake.usgs.gov/earthquakes/map. Accessed: 2025-02-22.

Talagrand, M. 2014. *Upper and lower bounds for stochastic processes*, volume 60. Springer. Pp. 28–32.

Wendland, H. 2004. *Scattered data approximation*, volume 17. Cambridge university press.

Whitehouse, J.; Ramdas, A.; and Wu, S. Z. 2023. On the sublinear regret of GP-UCB. *Advances in Neural Information Processing Systems*, 36: 35266–35276.

Williams, C. K.; and Rasmussen, C. E. 2006. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA.

Wu, Z.-m.; and Schaback, R. 1993. Local error estimates for radial basis function interpolation of scattered data. *IMA journal of Numerical Analysis*, 13(1): 13–27.