Error bounds for any regression model using Gaussian processes with gradient information

Rafael Savvides
University of Helsinki

Hoang Phuc Hau Luu
University of Helsinki

Kai Puolamäki
University of Helsinki

Abstract

We provide an upper bound for the expected quadratic loss on new data for any regression model. We derive the bound by modelling the underlying function by a Gaussian process (GP). Instead of a single kernel or family of kernels of the same form, we consider all GPs with translation-invariant and continuously twice differentiable kernels having a bounded signal variance and prior covariance of the gradient. To obtain a bound for the expected posterior loss, we present bounds for the posterior variance and squared bias. The squared bias bound depends on the regression model used, which can be arbitrary and not based on GPs. The bounds scale well with data size, in contrast to computing the GP posterior by a Cholesky factorisation of a large matrix. More importantly, our bounds do not require strong prior knowledge as we do not specify the exact kernel form. We validate our theoretical findings by numerical experiments and show that the bounds have applications in uncertainty estimation and concept drift detection.

1 INTRODUCTION

In a typical regression problem, we fit a regression model $\hat{f}$ on training data and then evaluate its performance through its expected error on new data, called the generalisation loss. Generalisation loss is most commonly estimated using a labelled holdout set or cross-validation, assuming that the new data come from the training distribution $\mathcal{P}$.

Suppose that we want to estimate the (unknown) loss of the regressor at a new, unlabelled test point $\mathbf{x}^*$. One approach is to use the estimate of the generalisation loss, or an estimate using a different test distribution, $\mathcal{P}'$, from which we have holdout data. Under this approach, we estimate the loss at a data point using the expected loss over the distribution from which that data point was drawn.

Estimating the loss at a point this way is often not enough. First, in many real-world cases, such as under concept drift (Gama et al., 2014) or domain adaptation (Redko et al., 2022), we cannot reasonably assume $\mathbf{x}^*$ comes from the training distribution $\mathcal{P}$, and we may not have enough labelled data from the testing distribution $\mathcal{P}'$. Secondly, suppose that we estimate the unknown loss at a point using the expected loss over its distribution. In this case, we get the same result for two different unlabeled test points $\mathbf{x}_1^*$ and $\mathbf{x}_2^*$, even when they are very far apart. This means that we cannot compare data points in applications such as active learning (Settles, 2009) or concept drift detection without ground truth (Oikarinen et al., 2021). To say something about the loss at a particular test point, we have to change our viewpoint from the expectation $E_{(\mathbf{x},y)}$ over the data space to the expectation E_f over the space of functions f that generate observations.

In this general setting, we want to upper bound the expected loss of an arbitrary fixed regressor on an arbitrary, fixed, unlabelled testing data point. Since the underlying function f that generates data is generally unknown, the loss we calculate depends on our beliefs about f. Our beliefs are encoded through the distribution we specify for f, and the expectation over this distribution gives the expected loss for each test point.

A powerful framework for specifying distributions over functions is given by Gaussian processes (GP), where the distribution of interest is the *GP posterior*. Under the usual GP workflow, we can estimate the expected loss as follows: (i) specify a kernel $k(\mathbf{x}, \mathbf{x}')$ encoding one's prior knowledge about f, such as smoothness or periodicity, (ii) condition on training observations to get the posterior distribution, and (iii) use the posterior to evaluate the *(expected) posterior loss*.

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. Copyright 2024 by the author(s).

Using GPs to estimate the expected loss has two drawbacks. Firstly, specifying the exact form for the kernel turns out to be a strong prior assumption that heavily affects the posterior distribution. This can make the posterior loss less reliable since different kernels result in different losses and since setting a kernel can be hard to reason about. A domain expert may not be able to confidently decide between different kernels, and eliciting a prior from them is in general non-trivial (Mikkola et al., 2023). Secondly, even if we have sufficient prior knowledge to specify the kernel's analytical form, computing the posterior loss can be prohibitively expensive, as it requires performing a Cholesky factorisation for a potentially large matrix.

In this paper, we derive an upper bound for the posterior loss of an arbitrary regressor on an arbitrary test point using the GP framework using *milder*, yet *easier to specify* modelling assumptions. The bound is *computationally efficient*, as it does not require Cholesky factorisation for a large matrix. Using milder assumptions, our bound holds uniformly for a general class of kernels satisfying specific properties. Using assumptions that are easier to specify allows the user to obtain an uncertainty estimate without specifying a kernel. The prior knowledge we require involves upper bounds on the gradient and the signal variance, which can be easier to elicit from domain experts than an exact kernel.

Our primary assumption is *translation-invariance*: how much two points, $\mathbf{p}$ and $\mathbf{q}$, affect each other depends on their relative locations (the difference $\mathbf{p}$-$\mathbf{q}$), but not their absolute locations ($\mathbf{p}$ and $\mathbf{q}$ solely). This general class of kernels contains many commonly used kernels, such as the radial basis function (RBF), rational quadratic (RQ), and Matern kernels. Without any other information about the underlying function outside the training data, if we want to say anything about extrapolation behaviour, we need to assume some properties are shared between the training data and the testing data. Translation invariance is a sensible assumption since it allows us to extrapolate and inject prior *directional* information. The latter is not possible if we restrict ourselves further to the *isotropy* assumption in which how $\mathbf{p}$ and $\mathbf{q}$ affect each other is based solely on $\|\mathbf{p} - \mathbf{q}\|$.

The importance of directional information for the error is illustrated in Figure 1, in which the actual function $f(x_1, x_2)$ varies faster on average in the x_1 direction than in x_2. We use training data from f to train a regressor (not pictured) and want to extrapolate outside the training data. In that case, the extrapolation behaviour depends on the direction: the error rises on average faster when extrapolating in the x_1 direction than in the x_2 direction.

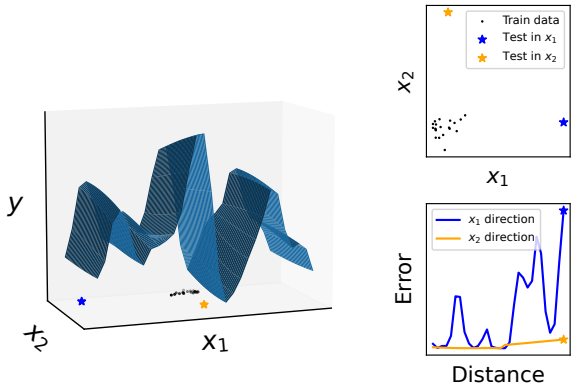

Figure 1: Left: A function $y = f(x_1, x_2)$, in which the average gradient in the x_1 direction is larger than in the x_2 direction. Right: The errors of a regressor $\hat{f}(x_1, x_2) \approx y$ can rise faster in the x_1 direction.

The structure and the contributions of this paper are as follows. In Sect. 2 we survey the related work, in Sect. 3, we define the problem and provide desirable upper bounds and theoretical proofs, and in Sect. 4 we demonstrate that our proposed method works correctly and efficiently to upper bound the regression loss. We discuss future work in Sect. 5 and conclude the paper in Sect. 6.

2 RELATED WORK

Estimating the generalisation error is an essential subject in statistical learning. In its classical setting, only one unique distribution generates the data. The generalisation error can then be estimated, with cross-validation (Stone, 1974; Nadeau and Bengio, 1999; Hastie et al., 2009) or bootstrapping (Efron, 1992).

When there is more than one data distribution, the related fields are out-of-distribution generalisation (Liu et al., 2023), domain generalisation (Wang et al., 2022), and domain adaptation (Redko et al., 2022), in which the goal is to find a model that can generalise to an unseen test distribution. The main difference with our work is that we want to upper bound the generalisation error for a fixed regressor and a fixed test point rather than learning the regressor that best generalises to a test distribution.

Our work is closely related to the literature on posterior variance analysis of GPs, where the goal is a tight bound for the *posterior variance* while remaining computationally affordable (Williams and Vivarelli, 2000; Lederer et al., 2019; Sollich, 1998; Sollich and Halees, 2002; Le Gratiet and Garnier, 2015). The essential distinction of our work is twofold. Firstly, we do not

follow the classical setting where the form of the GP kernel must be specified in advance. Our proposed bound applies to a vast class of translation-invariant and twice continuously differentiable kernels and only requires prior information on the signal variance and the gradient covariance. It is worth addressing that knowing the kernel implies knowing the gradient covariance and signal variance, but the converse does not hold. Secondly, we also provide a bound for the squared bias (the difference between regressor and posterior mean), which is needed to bound the posterior error. We are not aware of an existing bound for the squared bias in our general setting with an arbitrary regressor and an unknown kernel.

Although the posterior variance of GPs is sometimes used as an error bound, it only holds under the correct prior assumption, i.e., when the regressor is also a GP with the same prior as the true function (Sollich, 1998). Capone et al. (2022) propose a uniform error bound for GPs with translation-invariant kernels that holds under misspecification of the kernel length scale. However, the kernel form must be given in advance and the regressor must be a GP with that kernel form. We do not require the kernel form and consider a general regressor.

3 THEORY

3.1 Definitions and setting

Assume a function $f : \mathbb{R}^p \to \mathbb{R}$ that generates data as $y = f(\mathbf{x}) + \epsilon$, where the covariate $\mathbf{x} \in \mathbb{R}^p$ and ϵ is independent Gaussian noise with zero mean $E[\epsilon] = 0$ and variance $E\left[\epsilon^2\right] = \sigma^2$. Furthermore, assume that we have n *training data* points $\mathbf{x}_1, \ldots, \mathbf{x}_n$, where $\mathbf{x}_i \in \mathbb{R}^p$. We denote the training data set by $D = \{(\mathbf{x}_i, y_i)\}_{i\in[n]}$, where we use the shorthand notation $[n] = \{1, \ldots, n\}$, the set of training covariates by $\mathbf{X} = \{\mathbf{x}_i\}_{i\in[n]}$ and the vector of training labels by $\mathbf{y} = [y_1, y_2, \ldots, y_n]^\top$.

Assume that we have a regression model $\hat{f}$ and a *testing data* point $\mathbf{x}^* \in \mathbb{R}^p$. The main object of interest is the discrepancy between the prediction and its true value at the test point, $(y^* - \hat{f}(\mathbf{x}^*))^2 = (f(\mathbf{x}^*) + \epsilon^* - \hat{f}(\mathbf{x}^*))^2$.

We model f as a GP with a zero mean function and covariance between points of f characterised by a symmetric, positive definite kernel $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$. The GP specifies a prior distribution $\mathcal{F}$ over functions f. When conditioning on the observations D, we obtain the posterior distribution of f, denoted by $\mathcal{F}'$. The predictive distribution of $f(\mathbf{x}^*)$ is given by a normal distribution (Rasmussen and Williams, 2006, Page 16) with a (posterior) mean, denoted by $\bar{f}(\mathbf{x}^*)$,

$$E_{f\sim\mathcal{F}'}[f(\mathbf{x}^*)] = K(\mathbf{x}^*, \mathbf{X})[K(\mathbf{X}, \mathbf{X}) + \sigma^2\mathbf{I}]^{-1}\mathbf{y}, \quad (1)$$

and (posterior) variance

$$V_{f\sim\mathcal{F}'}[f(\mathbf{x}^*)] = k(\mathbf{x}^*, \mathbf{x}^*) - K(\mathbf{x}^*, \mathbf{X})[K(\mathbf{X}, \mathbf{X}) + \sigma^2\mathbf{I}]^{-1}K(\mathbf{X}, \mathbf{x}^*), \quad (2)$$

where $K(\mathbf{U}, \mathbf{V})$ denotes the matrix whose elements are $k(\mathbf{u}, \mathbf{v})$, $\mathbf{u} \in \mathbf{U}, \mathbf{v} \in \mathbf{V}$. To ensure that Equations (1) and (2) are well-defined even when the kernel matrix $K(\mathbf{X}, \mathbf{X})$ is singular, we assume $\sigma > 0$, but we can always take the limit of $\sigma \to 0^+$.

We consider kernels that are (i) continuously twice differentiable, (ii) translation invariant, (iii) have the prior signal variance upper bounded by $\tilde{\sigma}_0^2$, and (iv) the prior covariance of the gradient upper bounded by $\tilde{\mathbf{C}}$. Translation invariance means that $k(\mathbf{p}, \mathbf{q}) = k(\mathbf{p} + \mathbf{u}, \mathbf{q} + \mathbf{u})$ for any $\mathbf{p}, \mathbf{q}, \mathbf{u} \in \mathbb{R}^p$. Since k only depends on the difference of its inputs, it can be expressed as $k(\mathbf{p}, \mathbf{q}) = \kappa(\mathbf{p} - \mathbf{q})$ for some function κ. Examples are the radial basis function (RBF), rational quadratic (RQ), and Matern kernels. We assume the signal variance is bounded as:

$$\sigma_0^2 := E_{f\sim\mathcal{F}}\left[f(\mathbf{p})^2\right] \leq \tilde{\sigma}_0^2, \quad (3)$$

and the prior gradient covariance is bounded as:

$$\mathbf{C} := E_{f\sim\mathcal{F}}\left[\nabla f(\mathbf{p})(\nabla f(\mathbf{p}))^\top\right] \preceq \tilde{\mathbf{C}}, \quad (4)$$

which means $\tilde{\mathbf{C}} - \mathbf{C} \in \mathbb{R}^{p\times p}$ is positive semi-definite. Lemma 2 will show that $\mathbf{C}$ is not dependent on $\mathbf{p}$.

Since f is generally unknown, the best we can do in this paradigm is to upper-bound the expected posterior loss:

$$L(\mathbf{x}^*) = E_{f\sim\mathcal{F}',\epsilon^*}\left[\left(y^* - \hat{f}(\mathbf{x}^*)\right)^2\right]. \quad (5)$$

Our main problem is as follows.

Problem 1. *Given the definitions above, find an upper bound $U(\mathbf{x}^*)$ for the expected posterior loss $L(\mathbf{x}^*) \leq U(\mathbf{x}^*)$ when the kernel is translation-invariant, continuously twice differentiable, with a prior signal variance bounded by $\tilde{\sigma}_0^2$ as in Eq. (3) and a prior gradient covariance bounded by $\tilde{\mathbf{C}}$ as in Eq. (4).*

We intend to find an upper bound U that is as tight as possible in a computationally efficient manner. We will use the bias-variance decomposition of $L(\mathbf{x}^*)$:

$$L(\mathbf{x}^*) = \mathcal{I}(\mathbf{x}^*) + \mathcal{V}(\mathbf{x}^*) + \mathcal{B}(\mathbf{x}^*)^2, \quad (6)$$

where the *irreducible loss* is given by $\mathcal{I}(\mathbf{x}^*) = \sigma^2$, the *posterior variance* by

$$\mathcal{V}(\mathbf{x}^*) = E_{f\sim\mathcal{F}'}\left[\left(f(\mathbf{x}^*) - \bar{f}(\mathbf{x}^*)\right)^2\right], \quad (7)$$

where $\bar{f}(\mathbf{x}) = E_{f\sim\mathcal{F}'}[f(\mathbf{x})]$, and the *squared bias* by

$$\mathcal{B}(\mathbf{x}^*)^2 = \left(\hat{f}(\mathbf{x}^*) - \bar{f}(\mathbf{x}^*)\right)^2. \tag{8}$$

We specify a fixed value for the irreducible loss σ^2. The problem remains to give a tight upper bound for the posterior variance (Sect. 3.2) and the squared bias (Sect. 3.3).

3.2 Posterior variance bound

This section gives an upper bound for the posterior variance (Theorem 3). First, we present some lemmas that will be used in the proofs. The following celebrated theorem gives characterisation for a translation-invariant kernel.

Lemma 1 (Bochner's representation theorem (Rudin, 1960))**.** *Let $k(\mathbf{p},\mathbf{q}) = \kappa(\mathbf{p}-\mathbf{q})$ be a continuous, positive definite, translation-invariant kernel. Then κ is the Fourier transformation of a finite, non-negative measure $\hat{\mu}$:*

$$\kappa(\mathbf{p}-\mathbf{q}) = \frac{1}{(2\pi)^p}\int_{\mathbb{R}^p} \exp\left(i(\mathbf{p}-\mathbf{q})^\top\boldsymbol{\omega}\right) d\hat{\mu}(\boldsymbol{\omega}),$$

which can be simplified since we consider real kernels:

$$\kappa(\mathbf{p}-\mathbf{q}) = \frac{1}{(2\pi)^p}\int_{\mathbb{R}^p} \cos((\mathbf{p}-\mathbf{q})^\top\boldsymbol{\omega}) d\hat{\mu}(\boldsymbol{\omega}).$$

Lemma 2. *Let k be a twice continuously differentiable, positive definite, translation invariant kernel. Then:*

1. *$k(\mathbf{p},\mathbf{p}) = \sigma_0^2$ is constant for all $\mathbf{p}\in\mathbb{R}^p$.*
2. *$|k(\mathbf{p},\mathbf{q})| \le \sigma_0^2$ for any $\mathbf{p},\mathbf{q}\in\mathbb{R}^p$.*
3. *$\mathbf{C}_{ij} = \left.\dfrac{\partial^2 k(\mathbf{p},\mathbf{q})}{\partial\mathbf{p}_i\partial\mathbf{q}_j}\right|_{\mathbf{q}=\mathbf{p}}$, and it is constant with respect to $\mathbf{p}$.*

Proof. Item 1 is straightforward and Item 2 follows from the Cauchy-Schwartz inequality. Item 3 follows from the fact that differentiation is a linear operator, so for each i, $\partial f(\mathbf{p})/\partial\mathbf{p}_i$ is again a Gaussian process, and the covariance between these processes can be computed in terms of the second derivative of k (Rasmussen and Williams, 2006, Sect 9.4). In particular,

$$E_{f\sim\mathcal{F}}\left[\frac{\partial f(\mathbf{p})}{\partial\mathbf{p}_i}\frac{\partial f(\mathbf{q})}{\partial\mathbf{q}_j}\right] = \frac{\partial^2 k(\mathbf{p},\mathbf{q})}{\partial\mathbf{p}_i\partial\mathbf{q}_j}.$$

By translation invariance, it then further holds

$$\frac{\partial^2 k(\mathbf{p},\mathbf{q})}{\partial\mathbf{p}_i\partial\mathbf{q}_j} = -\frac{\partial^2\kappa(\mathbf{p}-\mathbf{q})}{\partial\mathbf{p}_i\partial\mathbf{q}_j}.$$

□

We denote the norm $\|\mathbf{p}\|_{\mathbf{C}}^2 := \mathbf{p}^\top\mathbf{C}\mathbf{p}$. Recall $\mathbf{C}$ is the matrix whose elements are $\mathbf{C}_{ij}$ from Lemma 2, or $\mathbf{C} = E_{f\sim\mathcal{F}}\left[(\nabla f)(\nabla f)^\top\right] = -\mathbf{Hess}_\kappa(\mathbf{0})$, where $\mathbf{Hess}_\kappa(\mathbf{0})$ denotes the Hessian matrix of κ at $\mathbf{0}$.

One of our main results is the following theorem: a local cosine-type lower bound for any translation-invariant and continuously twice differentiable kernel.

Theorem 1. *Given the definitions above, it holds*

$$k(\mathbf{p},\mathbf{q}) \ge \sigma_0^2\cos\left(\frac{\|\mathbf{q}-\mathbf{p}\|_{\mathbf{C}}}{\sigma_0}\right)$$

for any $\mathbf{p},\mathbf{q}\in\mathbb{R}^p$ such that $\|\mathbf{q}-\mathbf{p}\|_{\mathbf{C}} \le \pi\sigma_0$.

Proof. We prove here a special case that provides geometric intuition. The rigorous proof for the general claim is in Appendix A.1.

We can write $k(\mathbf{p},\mathbf{q}) = \langle\phi(\mathbf{p}),\phi(\mathbf{q})\rangle$ where ϕ is a feature map whose values are in some Hilbert space $\mathcal{H}$. Since $k(\mathbf{p},\mathbf{p}) = \sigma_0^2$ for all $\mathbf{p}$, it follows that $\|\phi(\mathbf{p})\| = \sigma_0$ for all $\mathbf{p}$, meaning that the range of the feature map ϕ is restricted to a sphere of radius σ_0.

Suppose the original feature space is 1D ($p = 1$), $\mathbf{C} = 1$, $\sigma_0 = 1$, and ϕ is differentiable and maps to a Euclidean space $\mathbb{R}^q$ with the canonical inner product. The range of ϕ is then restricted to the unit sphere of $\mathbb{R}^q$. Given $\mathbf{p},\mathbf{q}\in\mathbb{R}$, consider the following path on the sphere running from $\phi(\mathbf{p})$ to $\phi(\mathbf{q})$:

$$\begin{aligned}\alpha : [0,1] &\to \mathbb{R}^q,\\ t &\mapsto \phi((1-t)\mathbf{p} + t\mathbf{q}).\end{aligned}$$

Let R be the length of the shortest path in the sphere connecting $\phi(\mathbf{p})$ and $\phi(\mathbf{q})$. It follows that $R \le \text{length}(\alpha)$ (illustrated in Figure 2), and $\cos R = \langle\phi(\mathbf{p}),\phi(\mathbf{q})\rangle = k(\mathbf{p},\mathbf{q})$, since we consider a unit sphere.

Since $\mathbf{C} = 1$, it holds

$$1 = \left.\frac{\partial^2 k(\mathbf{p},\mathbf{q})}{\partial\mathbf{p}\partial\mathbf{q}}\right|_{\mathbf{q}=\mathbf{p}} = \langle\phi'(\mathbf{p}),\phi'(\mathbf{p})\rangle = \|\phi'(\mathbf{p})\|^2.$$

The length of α is then given by

$$\begin{aligned}\text{length}(\alpha) &= \int_0^1 \|\alpha'(t)\|dt\\ &= \int_0^1 \|\phi'((1-t)\mathbf{p}+t\mathbf{q})(\mathbf{q}-\mathbf{p})\|dt\\ &= |\mathbf{q}-\mathbf{p}|.\end{aligned}$$

It follows that $R \le |\mathbf{q}-\mathbf{p}|$. For $|\mathbf{q}-\mathbf{p}| \le \pi$, then $\cos R \ge \cos|\mathbf{q}-\mathbf{p}|$, or $k(\mathbf{p},\mathbf{q}) \ge \cos|\mathbf{q}-\mathbf{p}|$.

For the general claim, the main idea is that the triangle inequality for angles holds for a general Hilbert

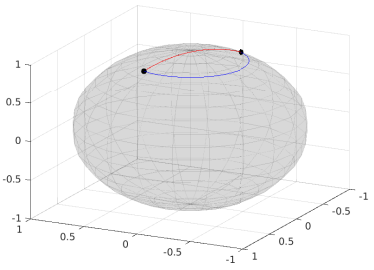

Figure 2: A feature map $\phi : \mathbb{R} \to \mathbb{R}^3$ given by $\phi(\mathbf{x}) = [\frac{1}{2}\sin(2\mathbf{x}), \frac{1}{2}\cos(2\mathbf{x}), \frac{\sqrt{3}}{2}]^\top$, for which $\|\phi(\mathbf{x})\| = \|\phi'(\mathbf{x})\| = 1$ and the corresponding kernel is translation-invariant. The black points are $\phi(\mathbf{p})$ and $\phi(\mathbf{q})$ with $\mathbf{p} = 5$, $\mathbf{q} = 6.5$; the red curve is the shortest path connecting these two points; the blue curve corresponds to the path $t \mapsto \phi((1-t)\mathbf{p} + t\mathbf{q})$.

space (Rao, 1976), and the shortest path idea can still be deployed by segmenting a path into small pieces. Everything is then translated back to the kernel, and final evaluations are done directly on the kernel itself using Bochner's representation theorem (Lemma 1).

□

Using Theorem 1, we derive a bound for the posterior variance when there is only one training data point:

Theorem 2. *When there is one training point,* $\mathbf{x}$*,*

$$\mathcal{V}(\mathbf{x}^*) \leq \frac{\sigma_0^4 \sin_*^2\left(\frac{\|\mathbf{x}-\mathbf{x}^*\|_\mathbf{C}}{\sigma_0}\right) + \sigma^2\sigma_0^2}{\sigma_0^2 + \sigma^2} \tag{9}$$

where $\sin_*$ *is defined as:*

$$\sin_*(t) = \begin{cases} \sin(t) & \text{if } 0 \leq t \leq \pi/2, \\ 1 & \text{otherwise.} \end{cases}$$

The bound increases faster when $\mathbf{x}^*$ moves in the direction of high variation because the matrix $\mathbf{C}$ contains directional information (illustrated in Figure 3). In the special case of an RBF, RQ, or Matern kernel, $\mathbf{C}$ recovers the notion of *length scale*, and the bound rises faster in directions of small length scales. For these kernels, $\mathbf{C}$ is diagonal with $\mathbf{C}_i := \mathbf{C}_{ii} \propto l_i^{-2}$ where l_i is the length scale with respect to the i-th coordinate (proved in Appendix A.10).

By observing that adding data points cannot increase the variance of a GP (Rasmussen and Williams, 2006, page 31) (shown in Appendix A.3 for completeness), we obtain the following bound for the posterior variance when there are n training data points.

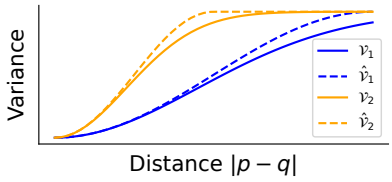

Figure 3: The posterior variance of Eq. (9) (solid lines) rises slower in the direction with a smaller $\mathbf{C}_i$ ($\mathbf{C}_1 < \mathbf{C}_2$). The bound (dashed lines) has the same behavior.

Theorem 3. *When there are* n *training data points* $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$*,*

$$\mathcal{V}(\mathbf{x}^*) \leq \min_{i \in [n]} \hat{\mathcal{V}}_i(\mathbf{x}^*) := \hat{\mathcal{V}}(\mathbf{x}^*), \tag{10}$$

where $\hat{\mathcal{V}}_i(\mathbf{x}^*)$ *is the RHS of Eq.* (9) *for* $\mathbf{x} = \mathbf{x}_i$*.*

When the test point is one of the training data points, the bound cannot exceed the irreducible loss σ^2.

Finally, in Theorem 4 we derive a uniform bound for the posterior variance (proof in Appendix A.4). We write $\mathcal{V}$ as $\mathcal{V}_\kappa$ and $\hat{\mathcal{V}}$ as $\hat{\mathcal{V}}_{\mathbf{C},\sigma_0}$ to show that the posterior variance depends on the kernel, while the bound only depends on $\mathbf{C}$ and σ_0. We also write $\mathcal{F}$ as $\mathcal{F}_\kappa$ to show that the prior distribution is kernel-dependent.

Theorem 4. *Theorems 2 and 3 still hold if we replace* $\mathbf{C}$ *and* σ_0 *with* $\tilde{\mathbf{C}}$ *and* $\tilde{\sigma}_0$*, as long as* $\mathbf{C} \preceq \tilde{\mathbf{C}}$ *and* $\sigma_0 \leq \tilde{\sigma}_0$*. As a consequence, we derive the uniform bound*

$$\sup_{\mathcal{F}_\kappa \in \mathbb{F}} \mathcal{V}_\kappa(\mathbf{x}^*) \leq \hat{\mathcal{V}}_{\tilde{\mathbf{C}},\tilde{\sigma}_0}(\mathbf{x}^*),$$

where $\mathbb{F}$ *denotes a class of translation-invariant and continuously twice differentiable priors satisfying* $E_{f\sim\mathcal{F}_\kappa}\left[f^2\right] \leq \tilde{\sigma}_0^2$ *and* $E_{f\sim\mathcal{F}_\kappa}\left[\nabla f(\nabla f)^\top\right] \preceq \tilde{\mathbf{C}}$*.*

Theorem 4 has the following implications:

- If we use our prior knowledge about $\tilde{\mathbf{C}}$ and $\tilde{\sigma}_0$, then the bound holds for all posterior variances whose prior kernel is compatible with our prior knowledge: $E\left[f^2\right] \leq \tilde{\sigma}_0^2$, and $E\left[\nabla f(\nabla f)^\top\right] \preceq \tilde{\mathbf{C}}$.
- If we do not have prior knowledge about $\tilde{\mathbf{C}}$ and $\tilde{\sigma}_0$, we can use empirical Bayes estimates: $\tilde{\sigma}_0$ can be estimated with the empirical signal variance and $\tilde{\mathbf{C}}$ can be estimated using the gradient of the underlying function with numerical methods.

Incorporating prior knowledge through $\tilde{\sigma}_0^2$ and $\tilde{\mathbf{C}}$ can be easier to understand and interpret than by eliciting a kernel function. We only require a small piece of prior information to get an uncertainty estimate: compare specifying a full kernel κ to specifying $\kappa(\mathbf{0}) = \sigma_0^2$

and $\mathbf{Hess}_\kappa(\mathbf{0}) = -\mathbf{C}$, or even just upper bounds for these quantities (by Theorem 4). One consequence of being able to overestimate σ_0^2 and $\mathbf{C}$ is that the user can vary these upper bounds to implicitly vary their level of prior confidence and obtain different posteriors; a non-trivial task in the classical setting.

3.3 Squared bias bound

Although the posterior variance is surprisingly stable (e.g., it cannot exceed σ_0^2) even at the limit of vanishing noise ($\sigma \to 0^+$), regardless of the structure of the training data, the squared bias can be arbitrarily large in the worst case. This makes it difficult to derive a sensible bound for the bias for all realistic scenarios.

We first demonstrate that the bias could be arbitrarily large via the following adversarial examples. The main factor making the bias explode is that the noise is at the limit of zero and the kernel matrix is at the limit of singularity (data points are too close in the following RBF example, and data points are well aligned with the kernel's periodicity in the cosine example).

As a first example, consider an RBF kernel $k(p,q) = \exp\left(-(p-q)^2/2\right)$ where $p, q \in \mathbb{R}$, two training data points at $x_1 = -\epsilon$ and $x_2 = \epsilon > 0$, with $y_1 = -1$ and $y_2 = 1$. Applying Eq. (1) for a test point at $x^* = 1$ we obtain $\bar{f}(x^*) = \frac{\exp(\frac{5\epsilon^2+2\epsilon+1}{2}) - \exp(\frac{5\epsilon^2-2\epsilon+1}{2})}{\exp(3\epsilon^2+1)\sigma^2 - \exp(\epsilon^2+1) + \exp(3\epsilon^2+1)}$. If we let $\sigma = \epsilon \to 0^+$, then $\bar{f}(x^*) \to +\infty$, and so does the bias.

As another example, consider the cosine kernel $k(p,q) = \cos(p-q)$ where $p, q \in \mathbb{R}$, two training data points $x_1 = 0$ and $x_2 = 2\pi + \epsilon$, with $y_1 = 0$ and $y_2 = 1$. Then the posterior mean at the test point $x^* = \pi/2$ is $\bar{f}(x^*) = \sin(\epsilon)(\sigma^2+1)/(\sin^2(\epsilon) + 2\sigma^2 + \sigma^4)$, which also tends to infinity if we let $\sigma = \sin(\epsilon)$ and let $\epsilon \to 0^+$.

We do not provide bounds for such adversarial cases; instead, we provide bounds that work if the noise term σ is not at the limit of zero (Theorem 5) or the training data points are well separated in the sense that the kernel matrix can be approximated by its diagonal (Theorem 6), thus avoiding the potentially large posterior gradients. Proofs of these theorems are given in Appendices A.5 and A.6.

Theorem 5. *If $\sigma > 0$, a bound for the bias is given by $|\mathcal{B}(\mathbf{x}^*)| \le \hat{\mathcal{B}}_1(\mathbf{x}^*)$, where:*

$$\hat{\mathcal{B}}_1(\mathbf{x}^*) := \frac{\|\mathbf{y}\|}{2} + \min_{i\in[n]} \left\{ |\hat{f}(\mathbf{x}^*) - y_i| + \frac{|y_i|}{2} + \frac{\|\mathbf{y}\|\sigma_0}{\sigma} \sin_* \left(\frac{\|\mathbf{x}_i - \mathbf{x}^*\|_{\mathbf{C}}}{2\sigma_0} \right) \right\}, \tag{11}$$

where $\sin_*$ is defined in Theorem 2.

Theorem 5 is still valid if we replace $\mathbf{C}, \sigma_0$ by $\tilde{\mathbf{C}}, \tilde{\sigma}_0$ where $\mathbf{C} \preceq \tilde{\mathbf{C}}$ and $\sigma_0 \le \tilde{\sigma}_0$, so a uniform bound can be derived for the bias, similar to Theorem 4 (shown in Appendix A.4).

As discussed in Remark 3.6 and the experimental results of Capone et al. (2022), the term $\|\mathbf{y}\|/\sigma$ can be overly conservative and in many applications it is commonly replaced with a fixed value, such as 2; see also Berkenkamp et al. (2017); Umlauft et al. (2017); Srinivas et al. (2012). In Section 5, we propose replacing it instead with $\beta\|\mathbf{y}\|/\sqrt{n}$ (for some $\beta > 0$) using a probabilistic argument to obtain a practical bias bound.

When the training data points are well separated, the kernel matrix can be effectively approximated by its diagonal, leading to the approximate bound in Theorem 6, which also works at the limit of vanishing noise.

Theorem 6. *Assuming the kernel matrix is diagonal, a bound for the bias is given by $|\mathcal{B}(\mathbf{x}^*)| \le \hat{\mathcal{B}}_2(\mathbf{x}^*)$, where $\hat{\mathcal{B}}_2(\mathbf{x}^*)$ is given by*

$$\min_{i\in[n]} \left\{ \left| \hat{f}(\mathbf{x}^*) - \frac{\sigma_0^2 y_i}{\sigma_0^2 + \sigma^2} \right| + \frac{2\sigma_0^2\|\mathbf{y}\|}{\sigma_0^2 + \sigma^2} \sin_* \left(\frac{\|\mathbf{x}_i - \mathbf{x}^*\|_{\mathbf{C}}}{2\sigma_0} \right) \right\},$$

where $\sin_*$ is defined as in Theorem 2. Consequently, at the limit of zero noise, $\sigma \to 0^+$, we obtain the following bound

$$\min_{i\in[n]} \left\{ |\hat{f}(\mathbf{x}^*) - y_i| + 2\|\mathbf{y}\| \sin_* \left(\frac{\|\mathbf{x}_i - \mathbf{x}^*\|_{\mathbf{C}}}{2\sigma_0} \right) \right\}.$$

4 EXPERIMENTS

The experiments show that the bounds of Theorems 3, 5 and 6 are valid, fast, and useful. The experiments were run in Python 3.10.8 on a high-performance computing cluster[1]. Appendix B contains experimental details and additional results.

The bounds are evaluated using the difference between the value and its bound, denoted by $\hat{\mathcal{V}} - \mathcal{V}$, $\hat{\mathcal{B}} - \mathcal{B}$, and $U - L$, where $U = \hat{\mathcal{V}} + \hat{\mathcal{B}}^2$. We report the median and quantiles of these differences (divided by the median value) over test data points and repeated samplings.

4.1 Bound evaluation with synthetic data

We evaluate the bounds on synthetic data generated from a GP. We show that the bounds are valid for various translation-invariant kernels and compare their tightness to baseline bounds, described below.

[1] Our code is publicly available at `https://github.com/edahelsinki/gpbound`.

One training point. We first examine the variance bound in 1D for one training point at $p = 0$ and test points $q \in [0, \pi/2]$. In Figure 4 (left), the cosine bound of Theorem 1 lower bounds the RBF, RQ, and Matern kernels. For the same kernels, in Figure 4 (right) the variance bound $\hat{\mathcal{V}}$ of Equation (9) upper bounds the true posterior variance $\mathcal{V}$. The variance bound holds for all kernels and is tightest for the cosine kernel[2]. If $\mathbf{C}$ is overestimated (dashed lines), the bound is more conservative, but it is still valid, due to Theorem 4.

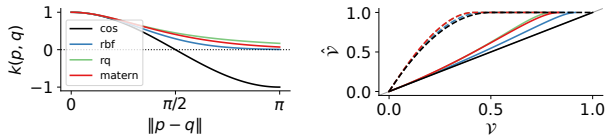

Figure 4: 1D with one training point. Left: Various kernels against the cosine kernel lower bound. Right: The variance bound $\hat{\mathcal{V}}$ upper bounds the true variance $\mathcal{V}$. The dashed lines denote an overestimated $\tilde{\mathbf{C}} = 4\mathbf{C}$.

Many training points. We examine the variance and bias bounds for 2D data with $N = 50$ train points. Higher dimensional cases are in Appendix B.3. Table 1 compares three variance bounds to the true posterior variance $\mathcal{V}$. The variance bound $\hat{\mathcal{V}}$ of Theorem 3 is always valid, as $\hat{\mathcal{V}} - \mathcal{V} > 0$, and tighter than the baseline variance bound of a constant $\sigma_0^2 = 1$, since $\hat{\mathcal{V}} - \mathcal{V} \approx 0.1 \cdot (\sigma_0^2 - \mathcal{V})$. The third entry $\mathcal{V}_{1/2}$ is the exact variance when using half as many points as the true variance $\mathcal{V}$. $\mathcal{V}_{1/2}$ is not a bound, but it is wider than $\mathcal{V}$, and it provides context for the tightness of the bound $\hat{\mathcal{V}}$, since $\hat{\mathcal{V}} - \mathcal{V} \approx 6 \cdot (\mathcal{V}_{1/2} - \mathcal{V})$.

Table 2 compares three bias bounds with the actual bias $\mathcal{B}$ of a random forest. The bias bound $\hat{\mathcal{B}}_{\text{Thm. 5}}$ is always valid, similarly to the variance bound $\hat{\mathcal{V}}$. The bias bound $\hat{\mathcal{B}}_{\text{Thm. 6}}$ is tighter, but may not be valid when the kernel matrix cannot be approximated by a diagonal matrix. Since, to the best of our knowledge, there is no prior work on the bias bound when the kernel form is unknown, for the third entry we include the following baseline bias bound (derived in Appendix A.9):

$$|\mathcal{B}| \leq \hat{\mathcal{B}}_{\hat{\mathcal{V}}} := \frac{|\hat{f}(\mathbf{x}^*)| + \sqrt{\|\mathbf{y}\|^2 + \hat{f}(\mathbf{x}^*)^2}}{2\sigma^2} \left(\hat{\mathcal{V}}(\mathbf{x}^*) + \sigma^2\right). \tag{12}$$

This baseline bound uses the variance bound $\hat{\mathcal{V}}$ and is looser than the proposed bias bounds $\hat{\mathcal{B}}_{\text{Thm. 5}}$ and $\hat{\mathcal{B}}_{\text{Thm. 6}}$.

[2]The cosine kernel $\cos(p-q)$ is a valid kernel for $p, q \in \mathbb{R}$.

Table 1: Evaluation of variance bound $\hat{\mathcal{V}}$. Each value is a median (and 5% quantile) over 100 test points sampled 200 times. See Appendix B for details. Parameters for $\hat{\mathcal{V}}$: $\mathbf{C} = 1$, $\sigma_0^2 = 1$, $\sigma^2 = 0.1$.

Kernel	$\frac{\hat{\mathcal{V}} - \mathcal{V}}{\text{median}(\mathcal{V})}$	$\frac{\sigma_0^2 - \mathcal{V}}{\text{median}(\mathcal{V})}$	$\frac{\mathcal{V}_{1/2} - \mathcal{V}}{\text{median}(\mathcal{V})}$
RBF	5.64 (4.06)	56.81 (52.15)	0.96 (0.183)
RQ	3.75 (2.34)	40.55 (36.47)	0.82 (0.044)
Matern	3.61 (2.35)	38.71 (34.84)	0.78 (0.077)

Table 2: Evaluation of bias bounds for the bias $\mathcal{B}$ of a random forest. The values and parameters are as in Table 1.

Kernel	$\frac{\hat{\mathcal{B}}_{\text{Thm. 5}} - \mathcal{B}}{\text{median}(\mathcal{B})}$	$\frac{\hat{\mathcal{B}}_{\text{Thm. 6}} - \mathcal{B}}{\text{median}(\mathcal{B})}$	$\frac{\hat{\mathcal{B}}_{\text{Eq. (12)}} - \mathcal{B}}{\text{median}(\mathcal{B})}$
RBF	40.49 (19.98)	7.44 (1.41)	59.54 (29.22)
RQ	50.22 (26.35)	9.74 (2.13)	74.12 (38.23)
Matern	42.83 (23.59)	8.32 (1.71)	62.01 (35.09)

4.2 Bound evaluation with real data

We next examine the total bound $U(x) = \hat{\mathcal{V}}(x) + \hat{\mathcal{B}}^2_{\text{Thm. 6}}(x)$ with real datasets and regressors. $U(x)$ upper bounds the expected loss $L(x) = E_y[(y - \hat{f}(x))^2]$ of a regression model $\hat{f}$ at a given point x, ignoring the irreducible error. When used in practice, $U(x)$ is compared to the observed error $e(x, y) = (y - \hat{f}(x))^2$ at a point x with an unknown y. Figure 5 shows the range of $U(x) - e(x, y)$ on test points (x, y) from several real datasets (described in Appendix B.1), using as regressors a random forest and an SVM. The bound is valid for all datasets, as $U(x) - e(x, y) > 0$ on average. Since $\mathbf{C}$ is not known for the real datasets, it is estimated using the regressor's gradient, approximated and averaged over training points (see Appendix B.3).

4.3 Scalability

Figure 6 shows the scaling of the variance bound $\hat{\mathcal{V}}$ of Equation (10) evaluated on N points in D dimensions. Table 6 contains additional results. The bound is compared to a GP trained and evaluated on the same points using the RBF, RQ, and Matern kernels. The variance bound is faster than a GP for all N and D. For example, for $N = 10000$ and $D = 25$, a GP requires ≈ 50 seconds while the bound requires ≈ 4 seconds. Besides being faster, recall that the bound is more general, as we don't need to specify a kernel. The computational complexity of the variance bound for N train points and M test points is $O(MND^2)$, while a GP is $O(N^3 + N^2(M + D) + NMD)$.

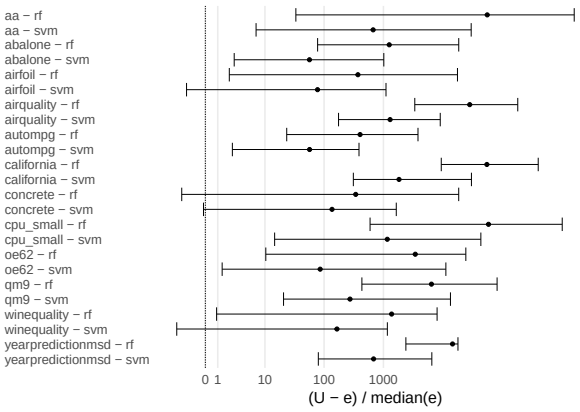

Figure 5: The total bound $U = \hat{\mathcal{V}} + \hat{\mathcal{B}}^2_{\text{Thm.6}}$ evaluated on real data sets to bound the regression errors $e = (y - \hat{f}(x))^2$, for $\hat{f} \in \{\text{rf}, \text{svm}\}$. Points are medians, and error bars are 1% and 99% quantiles over test data points (x, y) and 100 random splits.

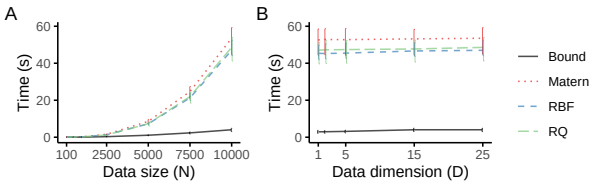

Figure 6: Scalability experiment. Error bars are 5% and 95% quantiles over 100 repetitions. For Figure A, $D = 25$. For Figure B, $N = 10000$.

4.4 Application: Drift detection

The potential application areas for the variance bound are the same as for the posterior variance of a GP. The total error bound U can be used when the squared error is needed, but ground truth labels are not available. One such application is drift detection.

In this experiment, we view drift detection as a binary classification: there is drift at a given point if the error exceeds a threshold (80% quantile of the validation set errors), and we detect drift if a drift indicator exceeds a threshold. We compare the AUC of two drift indicators: the bound U and the Euclidean distance from the training data centroid.

Figure 7 shows the squared error of a random forest regressor against the bound $U = \hat{\mathcal{V}} + \hat{\mathcal{B}}^2_{\text{Thm.6}}$ (left figure), the practical bound $U' = \hat{\mathcal{V}} + \hat{\mathcal{B}}^2_{\text{Thm.7}}$ (middle figure; discussed in Sect. 5), and the distance (right figure) evaluated at data points from AIRQUALITY; Appendix B.3 shows more datasets. Visually, both the bound and the distance follow the error reasonably well, suggesting that they can act as drift indicators.

Table 3 shows the AUC score of the total bound U and the distance for several real data sets and regressors. For most datasets, the bound can detect high errors fairly reliably. The lowest AUC scores can be attributed to the noisy estimation of $\mathbf{C}$, which is not known a priori here. Note also that while the distance is a baseline for drift detection, it is not an error bound.

Table 3: Drift detection results. AUC scores for the bound $U = \hat{\mathcal{V}} + \hat{\mathcal{B}}^2_{\text{Thm.6}}$ and the Euclidean distance as drift indicators.

Dataset	$\hat{f}$	Bound	Distance
AA	RF	**0.85**	0.52
	SVM	**0.92**	0.49
ABALONE	RF	**0.65**	0.59
	SVM	**0.65**	0.59
AIRFOIL	RF	**0.84**	0.66
	SVM	**0.81**	0.73
AIRQUALITY	RF	**0.90**	0.83
	SVM	**0.87**	0.83
AUTOMPG	RF	0.72	**0.74**
	SVM	0.88	**0.90**
CALIFORNIA	RF	**0.70**	0.59
	SVM	**0.66**	0.56
CONCRETE	RF	**0.60**	0.51
	SVM	**0.59**	0.53
CPU_SMALL	RF	**0.75**	0.63
	SVM	**0.77**	0.67
OE62	RF	**0.64**	0.62
	SVM	**0.64**	0.62
QM9	RF	**0.76**	0.71
	SVM	**0.77**	0.71
WINEQUALITY	RF	**0.67**	0.52
	SVM	**0.61**	0.54
YEARPREDICTIONMSD	RF	**0.58**	0.55
	SVM	**0.55**	0.53

5 DISCUSSION & FUTURE WORK

Non-negative kernels. We postulate that for non-negative kernels the kernel lower bound of Theorem 1 can be tightened to $k(\mathbf{p}, \mathbf{q}) \geq (1/2)\cos\left(\sqrt{2}\|\mathbf{q} - \mathbf{p}\|\right) + 1/2$ whenever $\|\mathbf{p} - \mathbf{q}\| \leq \pi/\sqrt{2}$ and $k(\mathbf{p}, \mathbf{q}) \geq 0$ otherwise (assuming unit signal variance and identity matrix $\mathbf{C}$). This lower bound allows us to tighten both the posterior variance and the squared bias bounds.

Practical bias bound. The term $\|\mathbf{y}\|$ in the bias bounds can be overly conservative, and it is commonly replaced with a smaller value when applying the bound in practice (Capone et al., 2022; Berkenkamp et al., 2017; Umlauft et al., 2017; Srinivas et al., 2012). We can replace $\|\mathbf{y}\|$ in a principled way to obtain a prac-

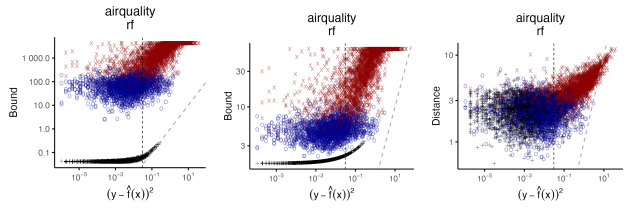

Figure 7: Drift detection on the AIRQUALITY dataset. Left: Using $U = \hat{\mathcal{V}} + \hat{\mathcal{B}}^2_{\text{Thm.6}}$. Middle: Using $U' = \hat{\mathcal{V}} + \hat{\mathcal{B}}^2_{\text{Thm.7}}$. Right: Using the distance from the training data centroid. Black crosses are train points, blue circles are validation points, red x's are test points, vertical lines denote error thresholds, tilted lines denote $x = y$.

tical bias bound as follows. Theorem 5 relies on the inequality $\langle \frac{\mathbf{y}}{\|\mathbf{y}\|}, \frac{\boldsymbol{\alpha}}{\|\boldsymbol{\alpha}\|} \rangle \leq 1$, where the equality holds iff $\mathbf{y}$ and $\boldsymbol{\alpha}$ point toward the same direction. The LHS is the cosine of two unit vectors, which is likely to be close to zero in large dimensions (and is non-negative since $\langle \mathbf{y}, \boldsymbol{\alpha} \rangle \geq 0$). Since the dimension of $\mathbf{y}$ and $\boldsymbol{\alpha}$ is n, the number of training points, which can be very large, the above inequality considers a worst-case scenario. In practice, the RHS can be replaced with a smaller value with high confidence by modeling $\boldsymbol{\alpha}$ as random ($\boldsymbol{\alpha}$ is unknown, since the kernel is unknown). Using a uniform distribution on $\frac{\boldsymbol{\alpha}}{\|\boldsymbol{\alpha}\|}$ leads to the practical bound of Theorem 7 (proof in Appendix A.7). Future work can consider similar arguments using more informative distributions for $\frac{\boldsymbol{\alpha}}{\|\boldsymbol{\alpha}\|}$, or even directly for the unknown kernel.

Theorem 7. *Assume $\frac{\boldsymbol{\alpha}}{\|\boldsymbol{\alpha}\|}$ is distributed uniformly on the half unit sphere with a pole $\frac{\mathbf{y}}{\|\mathbf{y}\|}$. Then, with probability of at least γ, it holds $|\mathcal{B}(\mathbf{x}^*)| \leq \hat{\mathcal{B}}_3(\mathbf{x}^*)$, where:*

$$\hat{\mathcal{B}}_3(\mathbf{x}^*) := \frac{\beta\|\mathbf{y}\|}{\sqrt{n}} + \min_{i\in[n]} \left\{ |\hat{f}(\mathbf{x}^*) - y_i| + \frac{\beta\|\mathbf{y}\|\sigma_0}{\sqrt{n}\sigma} \sin_* \left(\frac{\|\mathbf{x}_i - \mathbf{x}^*\|_{\mathbf{C}}}{2\sigma_0} \right) \right\} \tag{13}$$

where $\beta = \sqrt{2\log\left(\frac{2}{1-\gamma}\right)}$.

Bias bound from optimization. Theorem 8 presents an alternative direction for obtaining a bias bound in the case of a diagonal kernel matrix in Theorem 6. The idea is to obtain a bias bound by directly maximizing (or minimizing) the bias, under constraints implied by the kernel's properties. Future work can study this optimization problem to find efficient solutions. The optimization problem can be infeasible if the diagonality assumption is unmet, i.e., if data points are not well-separated. This assumption can be easily tested by checking that two arbitrary training data points $\mathbf{x}_i$ and $\mathbf{x}_j$ are distant in the sense that $\|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{C}} > \pi\sigma_0$. In addition, to meet the assumption in practice, the data can be modified by, for example, dropping or combining overly close data points.

Theorem 8. *Given a test point $\mathbf{x}^*$, a bias bound is given by $|\mathcal{B}(\mathbf{x}^*)| \leq \hat{\mathcal{B}}_4(\mathbf{x}^*)$, defined as:*

$$\hat{\mathcal{B}}_4(\mathbf{x}^*) := \max\left\{ |\hat{f}(\mathbf{x}^*) - \bar{f}_{min}(\mathbf{x}^*)|, |\hat{f}(\mathbf{x}^*) - \bar{f}_{max}(\mathbf{x}^*)| \right\}$$

where $\bar{f}_{min}(\mathbf{x}^*)$ and $\bar{f}_{max}(\mathbf{x}^*)$ denote respectively the optimal values of the following optimization problems:

$$\min \ (\max) \quad \frac{1}{\sigma^2 + \sigma_0^2}\langle \mathbf{y}, \mathbf{v} \rangle$$
$$s.t. \begin{cases} \|\mathbf{v}\| \leq \sigma_0^2, \\ \mathbf{v}_i \geq \sigma_0^2 \cos\left(\frac{\|\mathbf{x}_i - \mathbf{x}^*\|_{\mathbf{C}}}{\sigma_0} \right) & \forall i \in I, \end{cases}$$

where $I = \{i \in [n] : \|\mathbf{x}_i - \mathbf{x}^*\|_{\mathbf{C}} \leq \pi\sigma_0\}$.

Joint bound on multiple points. A potentially tighter bound can be obtained by bounding the joint error of test points rather than the error of each test point separately. However, considering the interactions between data points makes this joint bound more challenging to derive.

Estimation of C. When a suitable value for $\mathbf{C}$ is not available, it can be estimated from the data. However, estimating $\mathbf{C}$ from data is non-trivial, especially in high dimensions. An estimator that improves the simple estimator used in the experiments can improve the bounds. An estimator for $\mathbf{C}$ can also replace the costly $O(n^3)$ maximization of the marginal likelihood required for learning the length scales l_i of the RBF kernel since $\mathbf{C}_{ii} \propto l_i^{-2}$ (see Appendix A.10).

6 CONCLUSION

We derived a tight bound for the posterior variance and a bound for the squared bias of a regressor when the true function follows a GP. The bounds apply to a large class of kernels and can be viewed as the worst-case loss over this class. Our results offer a trade-off: the user provides less information (upper bounds on gradients and signal variance, instead of a kernel function) to obtain a fast error bound over a general class of kernels (instead of an error estimate for a single kernel).

Acknowledgements

We thank the anonymous reviewers for their valuable and constructive suggestions. R. Savvides was

funded by the Doctoral School of Computer Science at the University of Helsinki, and the Research Council of Finland (346376). H.P.H. Luu was funded by the Research Council of Finland (345704). We thank the Finnish Computing Competence Infrastructure (FCCI) for supporting this project with computational and data storage resources.

References

Felix Berkenkamp, Matteo Turchetta, Angela Schoellig, and Andreas Krause. Safe Model-based Reinforcement Learning with Stability Guarantees. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Alexandre Capone, Armin Lederer, and Sandra Hirche. Gaussian Process Uniform Error Bounds with Unknown Hyperparameters for Safety-Critical Applications. In *Proceedings of the 39th International Conference on Machine Learning*, pages 2609–2624. PMLR, 2022.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

Bradley Efron. Bootstrap Methods: Another Look at the Jackknife. In Samuel Kotz and Norman L. Johnson, editors, *Breakthroughs in Statistics*, pages 569–593. Springer New York, 1992. ISBN 978-0-387-94039-7 978-1-4612-4380-9.

João Gama, Indrė Žliobaitė, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):1–37, 2014.

Trevor Hastie, Robert Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, 2nd ed edition, 2009. ISBN 978-0-387-84857-0 978-0-387-84858-7.

I-Cheng Yeh. Concrete Compressive Strength, 1998.

Loic Le Gratiet and Josselin Garnier. Asymptotic analysis of the learning curve for Gaussian process regression. *Machine Learning*, 98(3):407–433, 2015.

Armin Lederer, Jonas Umlauft, and Sandra Hirche. Posterior Variance Analysis of Gaussian Processes with Application to Average Learning Curves, 2019.

S. Li. Concise Formulas for the Area and Volume of a Hyperspherical Cap. *Asian Journal of Mathematics & Statistics*, 4(1):66–70, 2010.

Jiashuo Liu, Zheyan Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards Out-Of-Distribution Generalization: A Survey, 2023.

Petrus Mikkola, Osvaldo A. Martin, Suyog Chandramouli, Marcelo Hartmann, Oriol Abril Pla, Owen Thomas, Henri Pesonen, Jukka Corander, Aki Vehtari, Samuel Kaski, Paul-Christian Bürkner, and Arto Klami. Prior knowledge elicitation: The past, present, and future, 2023.

Claude Nadeau and Yoshua Bengio. Inference for the generalization error. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999.

Emilia Oikarinen, Henri Tiittanen, Andreas Henelius, and Kai Puolamäki. Detecting virtual concept drift of regressors without ground truth values. *Data Mining and Knowledge Discovery*, 35(3):726–747, 2021.

A. Cerdeira Paulo Cortez. Wine Quality, 2009.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

R. Quinlan. Auto MPG, 1993.

Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1):140022, 2014.

D. K. Rao. A triangle inequality for angles in a Hilbert space. *Revista Colombiana de Matemáticas*, 10(3): 95–97, 1976.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. Adaptive Computation and Machine Learning. MIT Press, 2006. ISBN 978-0-262-18253-9.

Ievgen Redko, Emilie Morvant, Amaury Habrard, Marc Sebban, and Younès Bennani. A survey on domain adaptation theory: Learning bounds and theoretical guarantees, 2022.

Matti Ropo, Markus Schneider, Carsten Baldauf, and Volker Blum. First-principles data set of 45,892 isolated and cation-coordinated conformers of 20 proteinogenic amino acids. *Scientific Data*, 3(1):160009, 2016.

W. Rudin. *Fourier Analysis on Groups*. Number 12 in Interscience Tracts in Pure and Applied Mathematics. Interscience Publisher, 1960. ISBN 978-0-470-74481-9.

Burr Settles. Active Learning Literature Survey. Technical Report, University of Wisconsin-Madison Department of Computer Sciences, 2009.

Maciej Skorski. Bernstein-type bounds for beta distribution. *Modern Stochastics: Theory and Applications*, 10(2):211–228, 2023.

Peter Sollich. Learning Curves for Gaussian Processes. In *Advances in Neural Information Processing Systems*, volume 11. MIT Press, 1998.

Peter Sollich and Anason Halees. Learning Curves for Gaussian Process Regression: Approximations and Bounds. *Neural Computation*, 14(6):1393–1428, 2002.

Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias W. Seeger. Information-Theoretic Regret Bounds for Gaussian Process Optimization in the Bandit Setting. *IEEE Transactions on Information Theory*, 58(5):3250–3265, 2012.

M. Stone. Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):111–133, 1974.

Annika Stuke, Milica Todorović, Matthias Rupp, Christian Kunkel, Kunal Ghosh, Lauri Himanen, and Patrick Rinke. Chemical diversity in molecular orbital energy predictions with kernel ridge regression. *The Journal of Chemical Physics*, 150(20): 204121, 2019.

Annika Stuke, Christian Kunkel, Dorothea Golze, Milica Todorović, Johannes T. Margraf, Karsten Reuter, Patrick Rinke, and Harald Oberhofer. Atomic structures and orbital energies of 61,489 crystal-forming organic molecules. *Scientific Data*, 7(1):58, 2020.

T. Bertin-Mahieux. YearPredictionMSD, 2011.

D. Pope Thomas Brooks. Airfoil Self-Noise, 1989.

Jonas Umlauft, Thomas Beckers, Melanie Kimmel, and Sandra Hirche. Feedback linearization using Gaussian processes. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*, pages 5249–5255, 2017.

Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. OpenML: Networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60, 2014.

Saverio Vito. Air Quality, 2008.

Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to Unseen Domains: A Survey on Domain Generalization. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2022.

Tracy Sellers Warwick Nash. Abalone, 1994.

Christopher K.I. Williams and Francesco Vivarelli. Upper and Lower Bounds on the Learning Curve for Gaussian Processes. *Machine Learning*, 40(1): 77–102, 2000.

A PROOFS

This Appendix contains technical proofs for the theorems.

A.1 Proof of Theorem 1

We first state some auxiliary lemmas.

Lemma 3 (Triagle inequality for angles in Hilbert space (Rao, 1976))**.** *Let $\mathbf{x}, \mathbf{y}, \mathbf{z}$ be unit vectors in a Hilbert space, and define the angle $\theta_{\mathbf{xy}}$ by $\cos\theta_{\mathbf{xy}} = \langle \mathbf{x}, \mathbf{y}\rangle$, $0 \le \theta_{\mathbf{xy}} \le \pi$. The following triangle inequality for angles holds:*

$$\theta_{\mathbf{xz}} \le \theta_{\mathbf{xy}} + \theta_{\mathbf{yz}}.$$

As a consequence, for three vectors $\mathbf{x}, \mathbf{y}, \mathbf{z}$ with the same length σ_0, it holds:

$$\arccos \frac{\langle \mathbf{x}, \mathbf{z}\rangle}{\sigma_0^2} \le \arccos \frac{\langle \mathbf{x}, \mathbf{y}\rangle}{\sigma_0^2} + \arccos \frac{\langle \mathbf{y}, \mathbf{z}\rangle}{\sigma_0^2}.$$

Lemma 4.

$$\lim_{\delta \to 0^+} \frac{\arccos\left(1 - \frac{\delta^2}{2}\right)}{\delta} = 1. \tag{14}$$

Proof. From l'Hôpital's one-sided rule:

$$\begin{aligned}
\lim_{\delta \to 0^+} \frac{\arccos\left(1 - \frac{\delta^2}{2}\right)}{\delta} &= \lim_{\delta \to 0^+} \left[\arccos\left(1 - \frac{\delta^2}{2}\right)\right]' \\
&= \lim_{\delta \to 0^+} -\frac{1}{\sqrt{1 - \left(1 - \frac{\delta^2}{2}\right)^2}}(-\delta) \\
&= \lim_{\delta \to 0^+} \frac{\delta}{\sqrt{\delta^2 - \frac{\delta^4}{4}}} = 1.
\end{aligned}$$

□

The proof of Theorem 1 is given below.

Proof. Given two points $\mathbf{p}, \mathbf{q}$, by chopping the segment from $\mathbf{p}$ to $\mathbf{q}$ into m equal pieces and applying the triangle inequality in Lemma 3:

$$\arccos \frac{\langle \phi(\mathbf{p}), \phi(\mathbf{q})\rangle}{\sigma_0^2} \le \sum_{r=1}^{m} \arccos \frac{\langle \phi(\mathbf{p} + \frac{r-1}{m}(\mathbf{q} - \mathbf{p})), \phi(\mathbf{p} + \frac{r}{m}(\mathbf{q} - \mathbf{p}))\rangle}{\sigma_0^2}$$

or

$$\arccos \frac{k(\mathbf{p}, \mathbf{q})}{\sigma_0^2} \le \sum_{r=1}^{m} \arccos \frac{k(\mathbf{p} + \frac{r-1}{m}(\mathbf{q} - \mathbf{p}), \mathbf{p} + \frac{r}{m}(\mathbf{q} - \mathbf{p}))}{\sigma_0^2}. \tag{15}$$

By translation-invariance

$$k\left(\mathbf{p} + \frac{r-1}{m}(\mathbf{q} - \mathbf{p}), \mathbf{p} + \frac{r}{m}(\mathbf{q} - \mathbf{p})\right) = k\left(0, \frac{1}{m}(\mathbf{q} - \mathbf{p})\right),$$

the inequality (15) becomes:

$$\arccos \frac{k(\mathbf{p}, \mathbf{q})}{\sigma_0^2} \le m \arccos \frac{k(0, \frac{1}{m}(\mathbf{q} - \mathbf{p}))}{\sigma_0^2}, \quad \forall m \in \mathbb{N},$$

therefore,

$$\arccos \frac{k(\mathbf{p},\mathbf{q})}{\sigma_0^2} \leq \lim_{\epsilon \to 0^+} \frac{\arccos \frac{k(0,\epsilon(\mathbf{q}-\mathbf{p}))}{\sigma_0^2}}{\epsilon}. \tag{16}$$

By Bochner's representation theorem, the RHS of (16) can be written as:

$$\lim_{\epsilon \to 0^+} \frac{\arccos \frac{k(0,\epsilon(\mathbf{q}-\mathbf{p}))}{\sigma_0^2}}{\epsilon} = \lim_{\epsilon \to 0^+} \frac{\arccos \frac{\int_{\mathbb{R}^p} \cos(\epsilon(\mathbf{q}-\mathbf{p})^\top \boldsymbol{\omega}) d\hat{\mu}(\boldsymbol{\omega})}{(2\pi)^p \sigma_0^2}}{\epsilon}.$$

By applying the inequality:

$$\cos(\epsilon(\mathbf{q}-\mathbf{p})^\top \boldsymbol{\omega}) \geq 1 - \frac{\epsilon^2((\mathbf{q}-\mathbf{p})^\top \boldsymbol{\omega})^2}{2},$$

we obtain (notice $\hat{\mu}(\boldsymbol{\omega})$ is a non-negative measure)

$$\frac{\int_{\mathbb{R}^p} \cos(\epsilon(\mathbf{q}-\mathbf{p})^\top \boldsymbol{\omega}) d\hat{\mu}(\boldsymbol{\omega})}{(2\pi)^p \sigma_0^2} \geq 1 - \frac{\epsilon^2}{2\sigma_0^2(2\pi)^p} \int_{\mathbb{R}^p} ((\mathbf{q}-\mathbf{p})^\top \boldsymbol{\omega})^2 d\hat{\mu}(\boldsymbol{\omega}).$$

Note that $\mathbf{q}, \mathbf{p}$ are fixed from the beginning, and here we are at the limit of ϵ tending to 0, by choosing ϵ small enough, we can always make sure $1 - \frac{\epsilon^2}{2\sigma_0^2(2\pi)^p} \int_{\mathbb{R}^p} ((\mathbf{q}-\mathbf{p})^\top \boldsymbol{\omega})^2 d\hat{\mu}(\boldsymbol{\omega})$ is close to 1. Furthermore, arccos is a decreasing function in its domain $[-1, 1]$, therefore:

$$\arccos \frac{\int_{\mathbb{R}^p} \cos(\epsilon(\mathbf{q}-\mathbf{p})^\top \boldsymbol{\omega}) d\hat{\mu}(\boldsymbol{\omega})}{(2\pi)^p \sigma_0^2} \leq \arccos \left[1 - \frac{\epsilon^2}{2\sigma_0^2(2\pi)^p} \int_{\mathbb{R}^p} ((\mathbf{q}-\mathbf{p})^\top \boldsymbol{\omega})^2 d\hat{\mu}(\boldsymbol{\omega}) \right].$$

Combining this inequality with (16), we obtain:

$$\arccos \frac{k(\mathbf{p},\mathbf{q})}{\sigma_0^2} \leq \lim_{\epsilon \to 0^+} \frac{\arccos \left[1 - \frac{\epsilon^2}{2\sigma_0^2(2\pi)^p} \int_{\mathbb{R}^p} ((\mathbf{q}-\mathbf{p})^\top \boldsymbol{\omega})^2 d\hat{\mu}(\boldsymbol{\omega}) \right]}{\epsilon}.$$

By applying Lemma 4, we get

$$\lim_{\epsilon \to 0^+} \frac{\arccos \left[1 - \frac{\epsilon^2}{2\sigma_0^2(2\pi)^p} \int_{\mathbb{R}^p} ((\mathbf{q}-\mathbf{p})^\top \boldsymbol{\omega})^2 d\hat{\mu}(\boldsymbol{\omega}) \right]}{\epsilon} = \sqrt{\frac{\int_{\mathbb{R}^p} ((\mathbf{q}-\mathbf{p})^\top \boldsymbol{\omega})^2 d\hat{\mu}(\boldsymbol{\omega})}{\sigma_0^2(2\pi)^p}}.$$

On the other hand,

$$\frac{1}{(2\pi)^p} \int_{\mathbb{R}^p} \boldsymbol{\omega}_i \boldsymbol{\omega}_j d\hat{\mu}(\boldsymbol{\omega}) = E_{f \sim \mathcal{F}} \left(\frac{\partial f}{\partial \mathbf{q}_i} \frac{\partial f}{\partial \mathbf{q}_j} \right).$$

It follows that:

$$\begin{aligned}
&\frac{1}{(2\pi)^p} \int_{\mathbb{R}^p} ((\mathbf{q}-\mathbf{p})^\top \boldsymbol{\omega})^2 d\hat{\mu}(\boldsymbol{\omega}) \\
&= \frac{1}{(2\pi)^p} \sum_{i,j} \int_{\mathbb{R}^p} \boldsymbol{\omega}_i \boldsymbol{\omega}_j d\hat{\mu}(\boldsymbol{\omega}) (\mathbf{q}_i - \mathbf{p}_i)(\mathbf{q}_j - \mathbf{p}_j) \\
&= \sum_{i,j} E_{f \sim \mathcal{F}} \left(\frac{\partial f}{\partial \mathbf{q}_i} \frac{\partial f}{\partial \mathbf{q}_j} \right) (\mathbf{q}_i - \mathbf{p}_i)(\mathbf{q}_j - \mathbf{p}_j) \\
&= (\mathbf{q}-\mathbf{p})^\top \mathbf{C} (\mathbf{q}-\mathbf{p}).
\end{aligned}$$

Putting these together, we derive:

$$\arccos \frac{k(\mathbf{p},\mathbf{q})}{\sigma_0^2} \leq \frac{\|\mathbf{q}-\mathbf{p}\|_{\mathbf{C}}}{\sigma_0}.$$

Now, we want to apply cos to two sides of the above inequality. Notice that as cos is a decreasing function on $[0,\pi]$, we will require $\|\mathbf{q}-\mathbf{p}\|_{\mathbf{C}} \leq \pi\sigma_0$. □

A.2 Proof of Theorem 2

Proof.

$$\begin{aligned}
\mathcal{V}(\mathbf{x}^*) &= k(\mathbf{x}^*,\mathbf{x}^*) - \frac{k(\mathbf{x}^*,\mathbf{x})^2}{k(\mathbf{x},\mathbf{x})+\sigma^2} \\
&= \sigma_0^2 - \frac{k(\mathbf{x}^*,\mathbf{x})^2}{\sigma_0^2+\sigma^2} \\
&\leq \sigma_0^2 - \frac{\sigma_0^4 \cos^2\left(\dfrac{\|\mathbf{x}-\mathbf{x}^*\|_{\mathbf{C}}}{\sigma_0}\right)}{\sigma_0^2+\sigma^2} \\
&= \frac{\sigma_0^4 \sin^2\left(\dfrac{\|\mathbf{x}-\mathbf{x}^*\|_{\mathbf{C}}}{\sigma_0}\right) + \sigma^2\sigma_0^2}{\sigma_0^2+\sigma^2}.
\end{aligned} \tag{17}$$

□

A.3 Proof of Theorem 3

We prove that adding points does not increase the variance of a GP.

Proof. Denote by $\mathcal{D}$ the set of n observations $(\mathbf{x}_1,y_1),(\mathbf{x}_2,y_2),\ldots,(\mathbf{x}_n,y_n)$ and by $\mathcal{D}'$ the set of $\mathcal{D}$ and one extra observation $(\mathbf{x}_{n+1},y_{n+1})$. We also denote by X and X' the covariate variables extracted from $\mathcal{D}$ and $\mathcal{D}'$, respectively. We find the relation between the posterior variance given $\mathcal{D}$ and the posterior variance given $\mathcal{D}'$. Conditioned on $\mathcal{D}$, the posterior variance of the prediction at $\mathbf{x}^*$ is expressed as

$$\mathcal{V}_n(f(\mathbf{x}^*)) = k(\mathbf{x}^*,\mathbf{x}^*) - \mathbf{k}_*^\top \mathbf{G}_n^{-1}\mathbf{k}_*$$

where $\mathbf{G}_n = K(X,X)+\sigma^2\mathbf{I}$ and $\mathbf{k}_* = K(X,\mathbf{x}^*)$. On the other hand, conditioned on $\mathcal{D}'$, the variance is given by

$$\mathcal{V}_{n+1}(f(\mathbf{x}^*)) = k(\mathbf{x}^*,\mathbf{x}^*) - \begin{pmatrix}\mathbf{k}_* \\ a\end{pmatrix}^\top \mathbf{G}_{n+1}^{-1}\begin{pmatrix}\mathbf{k}_* \\ a\end{pmatrix}, \tag{18}$$

where $a = k(\mathbf{x}^*,\mathbf{x}_{n+1})$ and

$$\mathbf{G}_{n+1} = \begin{pmatrix}\mathbf{G}_n & \boldsymbol{\nu}_{n+1} \\ \boldsymbol{\nu}_{n+1}^T & c_{n+1}\end{pmatrix},$$

whereas $\boldsymbol{\nu}_{n+1} = K(X,\mathbf{x}_{n+1}), c_{n+1} = k(\mathbf{x}_{n+1},\mathbf{x}_{n+1})+\sigma^2$.

The inverse of $\mathbf{G}_{n+1}$ is given as (see, e.g., Rasmussen and Williams (2006, Appendix A.3))

$$\mathbf{G}_{n+1}^{-1} = \begin{pmatrix}\tilde{\mathbf{G}}_n & \tilde{\boldsymbol{\nu}}_{n+1} \\ \tilde{\boldsymbol{\nu}}_{n+1}^\top & \tilde{c}_{n+1}\end{pmatrix}$$

where

$$\begin{cases}
\tilde{\mathbf{G}}_n = \mathbf{G}_n^{-1} + \mathbf{G}_n^{-1}\boldsymbol{\nu}_{n+1}r^{-1}\boldsymbol{\nu}_{n+1}^\top\mathbf{G}_n^{-1} \\
r = c_{n+1} - \boldsymbol{\nu}_{n+1}^\top\mathbf{G}_n^{-1}\boldsymbol{\nu}_{n+1} \\
\tilde{\boldsymbol{\nu}}_{n+1} = -\mathbf{G}_n^{-1}\boldsymbol{\nu}_{n+1}r^{-1} \\
\tilde{c}_{n+1} = r^{-1}.
\end{cases}$$

We observe that r is the posterior variance of the prediction with noise at $\mathbf{x}_{n+1}$ conditioned on $\mathcal{D}$, i.e.,

$$r = \mathcal{V}_n(f(\mathbf{x}_{n+1})) + \sigma^2 > 0. \tag{19}$$

By extending (18), we get

$$\begin{aligned}
\mathcal{V}_{n+1}(f(\mathbf{x}^*)) &= k(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{k}_*^\top \tilde{\mathbf{G}}_n \mathbf{k}_* - a\tilde{\boldsymbol{\nu}}_{n+1}^\top \mathbf{k}_* - a\mathbf{k}_*^\top \tilde{\boldsymbol{\nu}}_{n+1}^\top - a^2 \tilde{c}_{n+1} \\
&= k(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{k}_*^\top \left[\mathbf{G}_n^{-1} + \mathbf{G}_n^{-1}\boldsymbol{\nu}_{n+1} r^{-1} \boldsymbol{\nu}_{n+1}^\top \mathbf{G}_n^{-1}\right] \mathbf{k}_* - 2a\mathbf{k}_*^\top \tilde{\boldsymbol{\nu}}_{n+1}^\top - a^2 \tilde{c}_{n+1} \\
&= \mathcal{V}_n(f(\mathbf{x}^*)) - r^{-1}\left[\mathbf{k}_*^\top \mathbf{G}_n^{-1} \boldsymbol{\nu}_{n+1} \boldsymbol{\nu}_{n+1}^\top \mathbf{G}_n^{-1} \mathbf{k}_* - 2a\boldsymbol{\nu}_{n+1}^\top \mathbf{G}_n^{-1} \mathbf{k}_* + a^2\right] \\
&= \mathcal{V}_n(f(\mathbf{x}^*)) - r^{-1}\left(\mathbf{k}_*^\top \mathbf{G}_n^{-1} \boldsymbol{\nu}_{n+1} - a\right)^2 \\
&\leq \mathcal{V}_n(f(\mathbf{x}^*)).
\end{aligned}$$

□

A.4 Proof of Theorem 4

We prove that the posterior variance bounds (Theorems 2 and 3) still hold if we overestimate $\mathbf{C}$ and σ_0 (Theorem 4). We also prove the bias bound of Theorem 5 still holds for overestimated $\mathbf{C}$ and σ_0.

Proof. We consider two cases.

Case 1: σ_0 is fixed, $\mathbf{C}$ is overestimated

In Theorem 2, when σ_0 is fixed, if we overestimate $\mathbf{C}$ by $\tilde{\mathbf{C}}$ in the sense that $\mathbf{C} \preceq \tilde{\mathbf{C}}$, the informative region, $\|\mathbf{x} - \mathbf{x}^*\|_{\mathbf{C}} \leq \pi\sigma_0/2$, becomes smaller, while the bound rises higher in that region since $\sin^2(t)$ is increasing in $[0, \pi/2]$. For Theorem 3, taking the min retains the monotonicity of $\mathbf{C}$. For Theorem 5, similar arguments apply: the informative region is shrunk, while the bias bound rises higher in that region.

Case 2: $\mathbf{C}$ is fixed, σ_0 is overestimated

Since taking the min does not change the monotonicity, it suffices to show the monotonicity with respect to σ_0 of the core function under the min. Let us fix $A := \|\mathbf{x} - \mathbf{x}^*\|_{\mathbf{C}}$.

For the posterior variance, we prove that the following function is increasing with respect to $t > 0$

$$\varphi(t) := \begin{cases} \dfrac{t^4 \sin^2\left(\dfrac{A}{t}\right) + \sigma^2 t^2}{t^2 + \sigma^2} & \text{if } t \geq \dfrac{2A}{\pi} \\ t^2 & \text{if } 0 < t < \dfrac{2A}{\pi}. \end{cases}$$

The derivative of φ is given by

$$\varphi'(t) = \begin{cases} \dfrac{4t^3 \sin^2\left(\dfrac{A}{t}\right) + 2\sigma^2 t - 2At^2 \cos\left(\dfrac{A}{t}\right) \sin\left(\dfrac{A}{t}\right)}{\sigma^2 + t^2} - \dfrac{2\sigma^2 t^3 + 2t^5 \sin^2\left(\dfrac{A}{t}\right)}{(\sigma^2 + t^2)^2} & \text{if } t > \dfrac{2A}{\pi} \\ 2t & \text{if } 0 < t < \dfrac{2A}{\pi}. \end{cases}$$

It then holds

$$\lim_{t \to \left(\frac{2A}{\pi}\right)^+} \varphi'(t) = \lim_{t \to \left(\frac{2A}{\pi}\right)^-} \varphi'(t) = \frac{4A}{\pi},$$

so φ is differentiable in $(0, \infty)$. We prove $\varphi'(t) \geq 0$ for all $t \in (0, \infty)$. It is trivially the case for $t \in (0, 2A/\pi]$. Let $t > (2A)/\pi$, $\varphi'(t) \geq 0$ is equivalent to

$$2t^2\sigma^2 \sin^2\left(\frac{A}{t}\right) + \sigma^4 + t^4 \sin^2\left(\frac{A}{t}\right) - At\sigma^2 \cos\left(\frac{A}{t}\right) \sin\left(\frac{A}{t}\right) - At^3 \cos\left(\frac{A}{t}\right) \sin\left(\frac{A}{t}\right) \geq 0,$$

which can be rewritten as

$$t^2\sigma^2\cos\left(\frac{A}{t}\right)\sin\left(\frac{A}{t}\right)\left[2\tan\left(\frac{A}{t}\right)-\frac{A}{t}\right]+t^4\cos\left(\frac{A}{t}\right)\sin\left(\frac{A}{t}\right)\left[\tan\left(\frac{A}{t}\right)-\frac{A}{t}\right]+\sigma^4\geq 0.$$

The last inequality holds by noting that $A/t<\pi/2$ so $\tan(A/t)\geq A/t$.

For the bias, we prove that the following function is increasing with respect to $t>0$

$$\vartheta(t):=t\sin_*\left(\frac{A}{2t}\right)=\begin{cases}t\sin\left(\frac{A}{2t}\right) & \text{if } t\geq\frac{A}{\pi}\\ t & \text{if } 0<t<\frac{A}{\pi}.\end{cases}$$

The derivative of ϑ is given by

$$\vartheta'(t)=\begin{cases}\sin\left(\frac{A}{2t}\right)-\frac{A}{2t}\cos\left(\frac{A}{2t}\right) & \text{if } t>\frac{A}{\pi}\\ 1 & \text{if } 0<t<\frac{A}{\pi}.\end{cases}$$

It holds

$$\lim_{t\to\left(\frac{A}{\pi}\right)^+}\vartheta'(t)=1,$$

so $\vartheta(t)$ is differentiable in $(0,\infty)$. Moreover, we shall show $\vartheta'(t)>0$ for all $t>0$. We only need to verify that for $t>A/\pi$. Let $s:=A/(2t)\in(0,\pi/2)$, using the inequality $\tan(s)\geq s$, we derive the conclusion. □

A.5 Proof of Theorem 5

Proof. The posterior mean is given by:

$$\bar{f}(\mathbf{x}^*)=\sum_{i=1}^n\boldsymbol{\alpha}_i k(\mathbf{x}_i,\mathbf{x}^*)=\left\langle\sum_{i=1}^n\boldsymbol{\alpha}_i\phi(\mathbf{x}_i),\phi(\mathbf{x}^*)\right\rangle$$

where $\boldsymbol{\alpha}=(\mathbf{K}+\sigma^2\mathbf{I})^{-1}\mathbf{y}$. Note that $\boldsymbol{\alpha}$ implies $\forall j\in[n]$:

$$y_j=\sum_{i=1}^n k(\mathbf{x}_j,\mathbf{x}_i)\boldsymbol{\alpha}_i+\sigma^2\boldsymbol{\alpha}_j=\left\langle\phi(\mathbf{x}_j),\sum_{i=1}^n\boldsymbol{\alpha}_i\phi(\mathbf{x}_i)\right\rangle+\sigma^2\boldsymbol{\alpha}_j,$$

and:

$$\sum_{i=1}^n\boldsymbol{\alpha}_i y_i=\left\|\sum_{i=1}^n\boldsymbol{\alpha}_i\phi(\mathbf{x}_i)\right\|^2+\sigma^2\sum_{i=1}^n\boldsymbol{\alpha}_i^2. \tag{20}$$

The posterior mean then is written $\forall j\in[n]$ as:

$$\begin{aligned}\bar{f}(\mathbf{x}^*)&=\left\langle\sum_{i=1}^n\boldsymbol{\alpha}_i\phi(\mathbf{x}_i),\phi(\mathbf{x}^*)-\phi(\mathbf{x}_j)\right\rangle+\left\langle\sum_{i=1}^n\boldsymbol{\alpha}_i\phi(\mathbf{x}_i),\phi(\mathbf{x}_j)\right\rangle\\&=\left\langle\sum_{i=1}^n\boldsymbol{\alpha}_i\phi(\mathbf{x}_i),\phi(\mathbf{x}^*)-\phi(\mathbf{x}_j)\right\rangle+y_j-\sigma^2\boldsymbol{\alpha}_j,\end{aligned}$$

and the bias is written $\forall j \in [n]$ as:

$$
\begin{aligned}
|\hat{f}(\mathbf{x}^*) - \bar{f}(\mathbf{x}^*)| &= \left| \hat{f}(\mathbf{x}^*) - \left\langle \sum_{i=1}^{n} \boldsymbol{\alpha}_i \phi(\mathbf{x}_i), \phi(\mathbf{x}^*) - \phi(\mathbf{x}_j) \right\rangle - y_j + \sigma^2 \boldsymbol{\alpha}_j \right| \\
&\leq |\hat{f}(\mathbf{x}^*) - y_j| + \sigma^2 |\boldsymbol{\alpha}_j| + \left\langle \sum_{i=1}^{n} \boldsymbol{\alpha}_i \phi(\mathbf{x}_i), \phi(\mathbf{x}^*) - \phi(\mathbf{x}_j) \right\rangle \\
&\leq |\hat{f}(\mathbf{x}^*) - y_j| + \sigma^2 |\boldsymbol{\alpha}_j| + \left\| \sum_{i=1}^{n} \boldsymbol{\alpha}_i \phi(\mathbf{x}_i) \right\| \|\phi(\mathbf{x}^*) - \phi(\mathbf{x}_j)\| \\
&= |\hat{f}(\mathbf{x}^*) - y_j| + \sigma^2 |\boldsymbol{\alpha}_j| + \left(\sum_{i=1}^{n} \boldsymbol{\alpha}_i y_i - \sigma^2 \|\boldsymbol{\alpha}\|^2 \right)^{1/2} \|\phi(\mathbf{x}^*) - \phi(\mathbf{x}_j)\| \qquad (21) \\
&\leq |\hat{f}(\mathbf{x}^*) - y_j| + \sigma^2 |\boldsymbol{\alpha}_j| + \left(\|\boldsymbol{\alpha}\| \|\mathbf{y}\| - \sigma^2 \|\boldsymbol{\alpha}\|^2 \right)^{1/2} \|\phi(\mathbf{x}^*) - \phi(\mathbf{x}_j)\| \\
&\leq |\hat{f}(\mathbf{x}^*) - y_j| + \sigma^2 |\boldsymbol{\alpha}_j| + \frac{\|\mathbf{y}\|}{2\sigma} \|\phi(\mathbf{x}^*) - \phi(\mathbf{x}_j)\| \qquad (22) \\
&\leq |\hat{f}(\mathbf{x}^*) - y_j| + \frac{|y_j| + \|\mathbf{y}\|}{2} + \frac{\|\mathbf{y}\|}{2\sigma} \|\phi(\mathbf{x}^*) - \phi(\mathbf{x}_j)\|, \qquad (23) \\
&\leq |\hat{f}(\mathbf{x}^*) - y_j| + \frac{|y_j| + \|\mathbf{y}\|}{2} + \frac{\|\mathbf{y}\| \sigma_0}{\sigma} \sin_* \left(\frac{\|\mathbf{x}^* - \mathbf{x}_i\|_{\mathbf{C}}}{2\sigma_0} \right), \qquad (24)
\end{aligned}
$$

where we used Equation (20) at (21), $a^2 + b^2 \geq 2ab$ at (22), Lemma 5 at (23), and Lemma 6 at (24).
By taking the min over $j \in [n]$, we obtain the desired bound. □

Lemma 5.

$$
|\boldsymbol{\alpha}_i| \leq \frac{|y_i| + \|\mathbf{y}\|}{2\sigma^2}, \quad \forall i \in [n]. \qquad (25)
$$

Proof. Derived for $i = 1$, from Equation (20):

$$
\begin{aligned}
\sigma^2 \sum_{i=1}^{n} \boldsymbol{\alpha}_i^2 &\leq \sum_{i=1}^{n} \boldsymbol{\alpha}_i y_i, \\
\sigma^2 \boldsymbol{\alpha}_1^2 + \sigma^2 \sum_{i \neq 1} \boldsymbol{\alpha}_i^2 - \boldsymbol{\alpha}_1 y_1 &\leq \sum_{i \neq 1} \boldsymbol{\alpha}_i y_i, \\
\sigma^2 \left(\boldsymbol{\alpha}_1 - \frac{y_1}{2\sigma^2} \right)^2 &\leq \sum_{i \neq 1} \boldsymbol{\alpha}_i y_i - \sigma^2 \sum_{i \neq 1} \boldsymbol{\alpha}_i^2 + \frac{y_1^2}{4\sigma^2} \leq \frac{1}{4\sigma^2} \|\mathbf{y}_{-1}\|^2 + \frac{y_1^2}{4\sigma^2} = \frac{\|\mathbf{y}\|^2}{4\sigma^2} \\
|\boldsymbol{\alpha}_1| &\leq \frac{|y_1| + \|\mathbf{y}\|}{2\sigma^2},
\end{aligned}
$$

where $\mathbf{y}_{-j}$ denotes the vector $\mathbf{y}$ excluded the j-th element y_j. □

Lemma 6. *Let k be a translation invariant kernel with signal variance σ_0^2, prior gradient covariance $\mathbf{C}$, and a feature space denoted by $\phi(\cdot)$. Then:*

$$
\|\phi(\mathbf{x}_1) - \phi(\mathbf{x}_2)\| \leq 2\sigma_0 \sin_* \left(\frac{\|\mathbf{x}_1 - \mathbf{x}_2\|_{\mathbf{C}}}{2\sigma_0} \right),
$$

where $\sin_*$ is defined in Theorem 2.

Proof. For $\|\mathbf{x}_1 - \mathbf{x}_2\|_{\mathbf{C}} \leq \pi\sigma_0$, from the kernel lower bound of Theorem 1 we get:

$$\begin{aligned}\|\phi(\mathbf{x}_1) - \phi(\mathbf{x}_2)\|^2 &= k(\mathbf{x}_1, \mathbf{x}_1) + k(\mathbf{x}_2, \mathbf{x}_2) - 2k(\mathbf{x}_1, \mathbf{x}_2) \\ &= 2\sigma_0^2 - 2k(\mathbf{x}_1, \mathbf{x}_2) \\ &\leq 2\sigma_0^2 - 2\sigma_0^2 \cos\left(\frac{\|\mathbf{x}_1 - \mathbf{x}_2\|_{\mathbf{C}}}{\sigma_0}\right) \\ &= 4\sigma_0^2 \sin^2\left(\frac{\|\mathbf{x}_1 - \mathbf{x}_2\|_{\mathbf{C}}}{2\sigma_0}\right)\end{aligned}$$

For $\|\mathbf{x}_1 - \mathbf{x}_2\|_{\mathbf{C}} > \pi\sigma_0$, from Bochner's theorem we get:

$$\begin{aligned}\|\phi(\mathbf{x}_1) - \phi(\mathbf{x}_2)\|^2 &= k(\mathbf{x}_1, \mathbf{x}_1) + k(\mathbf{x}_2, \mathbf{x}_2) - 2k(\mathbf{x}_1, \mathbf{x}_2) \\ &= \frac{2}{(2\pi)^p} \int_{\mathbb{R}^p} (1 - \cos((\mathbf{x}_2 - \mathbf{x}_1)^\top \boldsymbol{\omega})) d\hat{\mu}(\boldsymbol{\omega}) \\ &\leq \frac{4}{(2\pi)^p} \int_{\mathbb{R}^p} d\hat{\mu}(\boldsymbol{\omega}) = 4\sigma_0^2.\end{aligned}$$

□

A.6 Proof of Theorem 6

Proof. Since $\mathbf{K}$ is diagonal, the posterior mean can be expressed as

$$\begin{aligned}\bar{f}(\mathbf{x}^*) &= \frac{1}{\sigma_0^2 + \sigma^2} \sum_{i=1}^n y_i k(\mathbf{x}_i, \mathbf{x}^*) \\ &= \frac{1}{\sigma^2 + \sigma_0^2} \left\langle \phi(\mathbf{x}^*), \sum_{i=1}^n y_i \phi(\mathbf{x}_i) \right\rangle \\ &= \frac{1}{\sigma^2 + \sigma_0^2} \left\langle \phi(\mathbf{x}^*) - \phi(\mathbf{x}_j), \sum_{i=1}^n y_i \phi(\mathbf{x}_i) \right\rangle + \frac{1}{\sigma^2 + \sigma_0^2} \left\langle \phi(\mathbf{x}_j), \sum_{i=1}^n y_i \phi(\mathbf{x}_i) \right\rangle \\ &= \frac{1}{\sigma^2 + \sigma_0^2} \left\langle \phi(\mathbf{x}^*) - \phi(\mathbf{x}_j), \sum_{i=1}^n y_i \phi(\mathbf{x}_i) \right\rangle + \frac{\sigma_0^2 y_j}{\sigma_0^2 + \sigma^2}\end{aligned}$$

for all $j \in [n]$. Therefore,

$$\begin{aligned}\left|\bar{f}(\mathbf{x}^*) - \frac{\sigma_0^2 y_j}{\sigma_0^2 + \sigma^2}\right| &\leq \frac{1}{\sigma^2 + \sigma_0^2} \left\|\sum_{i=1}^n y_i \phi(\mathbf{x}_i)\right\| \cdot \|\phi(\mathbf{x}^*) - \phi(\mathbf{x}_j)\| \\ &\leq \frac{2\sigma_0^2 \|\mathbf{y}\|}{\sigma^2 + \sigma_0^2} \sin_*\left(\frac{\|\mathbf{x}^* - \mathbf{x}_j\|_{\mathbf{C}}}{2\sigma_0}\right).\end{aligned}$$

It follows that for all $j \in [n]$, it holds

$$|\hat{f}(\mathbf{x}^*) - \bar{f}(\mathbf{x}^*)| \leq \left|\hat{f}(\mathbf{x}^*) - \frac{\sigma_0^2 y_j}{\sigma_0^2 + \sigma^2}\right| + \frac{2\sigma_0^2 \|\mathbf{y}\|}{\sigma^2 + \sigma_0^2} \sin_*\left(\frac{\|\mathbf{x}^* - \mathbf{x}_j\|_{\mathbf{C}}}{2\sigma_0}\right)$$

By taking the min over $j \in [n]$, we derive the result. □

A.7 Proof of Theorem 7

Proof. Since $\boldsymbol{\alpha}/\|\boldsymbol{\alpha}\|$ is distributed uniformly on the half unit sphere with the pole $\mathbf{y}/\|\mathbf{y}\|$, it follows from Lemma 8 that $\frac{1}{2}\left(\langle \frac{\boldsymbol{\alpha}}{\|\boldsymbol{\alpha}\|}, \frac{\mathbf{y}}{\|\mathbf{y}\|} \rangle + 1\right) \sim Beta_{[1/2,1]}((n-1)/2, (n-1)/2)$. By using Lemma 10, for any $C > 0$,

$$P\left(\langle \boldsymbol{\alpha}, y \rangle \leq C\|\boldsymbol{\alpha}\|\|\mathbf{y}\|\right) \geq 1 - 2\exp\left(-\frac{nC^2}{2}\right). \tag{26}$$

Equivalently, with the probability of at least γ, it holds

$$\langle \boldsymbol{\alpha}, \mathbf{y} \rangle \leq \sqrt{\frac{2}{n} \log\left(\frac{2}{1-\gamma}\right)} \|\boldsymbol{\alpha}\| \|\mathbf{y}\| := \frac{\beta}{\sqrt{n}} \|\boldsymbol{\alpha}\| \|\mathbf{y}\|. \tag{27}$$

From (21), by applying the above probabilistic evaluation, we get

$$\begin{aligned} |\hat{f}(\mathbf{x}^*) - \bar{f}(\mathbf{x}^*)| &\leq |\hat{f}(\mathbf{x}^*) - y_j| + \sigma^2 |\boldsymbol{\alpha}_j| + \left(\frac{\beta}{\sqrt{n}} \|\boldsymbol{\alpha}\| \|\mathbf{y}\| - \sigma^2 \|\boldsymbol{\alpha}\|^2\right)^{1/2} \|\phi(\mathbf{x}^*) - \phi(\mathbf{x}_j)\| \\ &\leq |\hat{f}(\mathbf{x}^*) - y_j| + \sigma^2 |\boldsymbol{\alpha}_j| + \frac{\beta \|\mathbf{y}\|}{2\sigma\sqrt{n}} \|\phi(\mathbf{x}^*) - \phi(\mathbf{x}_j)\|. \end{aligned} \tag{28}$$

Similarly,

$$\sigma^2 \sum_{i=1}^{n} \boldsymbol{\alpha}_i^2 \leq \sum_{i=1}^{n} \boldsymbol{\alpha}_i y_i \leq \frac{\beta}{\sqrt{n}} \|\boldsymbol{\alpha}\| \|\mathbf{y}\|$$

so $\sigma^2 \|\boldsymbol{\alpha}\| \leq (\beta/\sqrt{n}) \|\mathbf{y}\|$, which further implies $\sigma^2 |\boldsymbol{\alpha}_j| \leq (\beta/\sqrt{n}) \|\mathbf{y}\|$ for all $j \in [n]$.

Therefore,

$$|\mathcal{B}(\mathbf{x}^*)| \leq \min_{i \in [n]} \left\{ |\hat{f}(\mathbf{x}^*) - y_i| + \frac{\beta \sigma_0 \|\mathbf{y}\|}{\sigma\sqrt{n}} \sin_* \left(\frac{\|\mathbf{x}_i - \mathbf{x}^*\|_{\mathbf{C}}}{2\sigma_0}\right) \right\} + \frac{\beta \|\mathbf{y}\|}{\sqrt{n}}. \tag{29}$$

□

Lemma 7 (Surface area of a hyperspherical cap (Li, 2010))**.** *Let $\mathbb{S}^{D-1}$ be the unit hypersphere in $\mathbb{R}^D$, and let it be cut into two parts (called caps) by a hyperplane above the equator in $\mathbb{R}^D$, where $\mathbf{e}_D = [0, 0, \ldots, 1]^\top$ is the northern direction. The upper cap's surface area is given by*

$$A_D^{cap} = \frac{1}{2} A_D I_{\sin^2(\phi)} \left(\frac{D-1}{2}, \frac{1}{2}\right)$$

where A_D is the surface area of the entire sphere, $A_D = 2\pi^{D/2}/\Gamma(D/2)$ (here Γ is the gamma function), $I_x(a, b)$ is the regularized incomplete beta function $I_x(a, b) = \int_0^x s^{a-1}(1-s)^{b-1} ds / B(a, b)$, $B(a, b)$ is the beta function $B(a, b) = \int_0^1 s^{a-1}(1-s)^{b-1} ds$, and $\phi \in [0, \pi]$ is the colatitude angle of a point in the intersection of the hyperplane and the hypersphere, i.e., the angle (in radians) down from the north pole to the point.

Lemma 8. *Let $\mathbf{u}$ be a random unit vector distributed uniformly on the upper half part of the unit sphere $\mathbb{S}^{D-1}$ (with the convention that $\mathbf{e}_D = [0, 0, \ldots, 0, 1]^\top$ is the north pole). Then $(\mathbf{u}_D + 1)/2$ follows a truncated Beta distribution $Beta_{[1/2,1]}((D-1)/2, (D-1)/2)$ where $\mathbf{u}_D = \langle \mathbf{u}, \mathbf{e}_D \rangle$ is the last coordinate of $\mathbf{u}$, which is also the height of the vector $\mathbf{u}$ in this case.*

Proof. For any $t \in [0, 1]$, the cumulative density function (CDF) of $\mathbf{u}_D$ at t is given by

$$F(t) = P(\mathbf{u}_D \leq t) = P(\mathbf{u}_D < t) = 1 - P(\mathbf{u}_D \geq t) = 1 - 2\frac{A_D^{cap}(t)}{A_D}$$

where $A_D^{cap}(t)$ is the surface area of the upper cap of the $\mathbb{S}^{D-1}$ cut by a horizontal hyperplane passing through $[0, 0, \ldots, 0, t]^\top$. According to Lemma 7 with the conversion $t = \cos(\phi)$, we have

$$A_D^{cap}(t) = \frac{1}{2} A_D I_{1-t^2} \left(\frac{D-1}{2}, \frac{1}{2}\right) \quad \text{if } 0 \leq t \leq 1.$$

We now differentiate F to get the probability density function f of $\mathbf{u}_D$. Let's say with $0 \leq t \leq 1$, we have

$$\begin{aligned}
f(t) &= F'(t) \\
&= -\left(I_{1-t^2}\left(\frac{D-1}{2}, \frac{1}{2}\right)\right)' \\
&= -\frac{1}{B\left(\frac{D-1}{2}, \frac{1}{2}\right)}\left(\int_0^{1-t^2} s^{\frac{D-3}{2}}(1-s)^{-\frac{1}{2}} ds\right)' \\
&= -\frac{1}{B\left(\frac{D-1}{2}, \frac{1}{2}\right)}(1-t^2)^{\frac{D-3}{2}} t^{-1}(-2t) \\
&= \frac{2}{B\left(\frac{D-1}{2}, \frac{1}{2}\right)}(1-t^2)^{\frac{D-3}{2}}.
\end{aligned}$$

Summarising, the probability density distribution of $\mathbf{u}_D$ is given by

$$f(t) \propto (1-t^2)^{\frac{D-3}{2}} \quad \text{for } t \in [0,1].$$

Let g be the probability density function of $(\mathbf{u}_D + 1)/2$, it follows that

$$g(s) \propto f(2s-1) \propto (1-s)^{\frac{D-3}{2}} s^{\frac{D-3}{2}} \quad \text{for } s \in [1/2, 1].$$

Therefore, $(\mathbf{u}_D + 1)/2 \sim Beta_{[1/2,1]}((D-1)/2, (D-1)/2)$, the truncated beta distribution with support $[1/2, 1]$. □

Lemma 9 (Bernstein inequality (Skorski, 2023))**.** *Let $X \sim Beta(\alpha, \alpha)$ then it holds*

$$P(X > 1/2 + \epsilon) \leq \exp\left(-(4\alpha + 2)\epsilon^2\right).$$

Lemma 10 (Bernstein inequality for truncated beta distribution)**.** *Let $Z \sim Beta_{[1/2,1]}(\alpha, \alpha)$ be the truncated beta distribution whose support is $[1/2, 1]$, then*

$$P(Z > 1/2 + \epsilon) \leq 2\exp\left(-(4\alpha + 2)\epsilon^2\right).$$

Proof. Let $X \sim Beta(\alpha, \alpha)$ and let f_X, f_Z be the probability distribution functions of X and Z, respectively. It follows that

$$f_Z(x) = \begin{cases} c f_X(x) & \text{if } x \in [1/2, 1] \\ 0 & \text{otherwise} \end{cases}$$

for some normalisation constant $c > 0$. Since $\int_{1/2}^1 f_Z(x)dx = 1$ and since $Beta(\alpha, \alpha)$ is symmetric, it follows that $c = \left(\int_{1/2}^1 f_X(x)dx\right)^{-1} = 2$. For $\epsilon > 0$, we have

$$\begin{aligned}
P(Z > 1/2 + \epsilon) &= \int_{\epsilon+1/2}^1 f_Z(x)dx \\
&= 2\int_{1/2+\epsilon}^1 f_X(x)dx \\
&= 2P(X > 1/2 + \epsilon) \\
&\leq 2\exp\left(-(4\alpha + 2)\epsilon^2\right)
\end{aligned}$$

where the last inequality uses Lemma 9. □

A.8 Proof of Theorem 8

Proof. Since $\mathbf{K}$ is assumed to be diagonal, $\{\phi(\mathbf{x}_i)\}_{i=1}^n$ is an orthogonal sequence of the Hilbert space $\mathcal{H}$. The coordinates of the projection of $\phi(\mathbf{x}^*)$ onto this basis are given by $\langle \phi(\mathbf{x}^*), \phi(\mathbf{x}_i)\rangle = k(\mathbf{x}^*, \mathbf{x}_i)$ for $i \in [n]$. By using Bessel's inequality, it holds

$$\sum_{i=1}^{n} k(\mathbf{x}_i, \mathbf{x}^*)^2 \leq \sigma_0^4. \tag{30}$$

On the other hand, from Theorem 1,

$$k(\mathbf{x}_i, \mathbf{x}^*) \geq \sigma_0^2 \cos\left(\frac{\|\mathbf{x}^* - \mathbf{x}_i\|_{\mathbf{C}}}{\sigma_0}\right) \tag{31}$$

whenever $\|\mathbf{x}_i - \mathbf{x}^*\| \leq \pi\sigma_0$.

Again, by diagonality of $\mathbf{K}$, the posterior mean is given by

$$\bar{f}(\mathbf{x}^*) = \frac{1}{\sigma_0^2 + \sigma^2} \sum_{i=1}^{n} y_i k(\mathbf{x}_i, \mathbf{x}^*).$$

By denoting $\mathbf{v}_i = k(\mathbf{x}_i, \mathbf{x}^*)$, $\bar{f}(\mathbf{x}^*)$ has to be within the range the minimal value and maximal value of $\frac{1}{\sigma_0^2 + \sigma^2} \sum_{i=1}^n y_i \mathbf{v}_i$ given the constraints (30) and (31). □

A.9 Proof of the baseline bias bound in Equation (12)

Proof. We use the setting and notation from Appendix A.3, and denote by $\mathbf{y}_{[n]}$ the vector of the n first elements of $\mathbf{y}$. We first connect the posterior mean given $\mathcal{D}$ and the posterior mean given $\mathcal{D}'$.

The posterior mean at $\mathbf{x}^*$ conditioned on $\mathcal{D}$ is

$$\begin{aligned}
\bar{f}_{n+1}(\mathbf{x}^*) &= \begin{pmatrix} \mathbf{k}_* \\ a \end{pmatrix}^\top \mathbf{G}_{n+1}^{-1} \begin{pmatrix} \mathbf{y}_{[n]} \\ y_{n+1} \end{pmatrix} \\
&= \mathbf{k}_*^\top \tilde{\mathbf{G}}_n \mathbf{y}_{[n]} + a\tilde{\boldsymbol{\nu}}_{n+1}^\top \mathbf{y}_{[n]} + \mathbf{k}_*^\top \tilde{\boldsymbol{\nu}}_{n+1} y_{n+1} + a\tilde{c}_{n+1} y_{n+1} \\
&= \bar{f}_n(\mathbf{x}^*) + r^{-1}(y_{n+1} - \boldsymbol{\nu}_{n+1}^\top \mathbf{G}_n^{-1} \mathbf{y}_{[n]})(a - \mathbf{k}_*^\top \mathbf{G}_n^{-1} \boldsymbol{\nu}_{n+1}),
\end{aligned}$$

By noting that $r = \sigma^2 + \mathcal{V}_n(f(\mathbf{x}_{n+1}))$, we derive

$$\bar{f}_{n+1}(\mathbf{x}^*) = \bar{f}_n(\mathbf{x}^*) + (y_{n+1} - \bar{f}_n(\mathbf{x}_{n+1})) \frac{a - \mathbf{k}_*^\top \mathbf{G}_n^{-1} \boldsymbol{\nu}_{n+1}}{\mathcal{V}_n(f(\mathbf{x}_{n+1})) + \sigma^2}. \tag{32}$$

If we choose the new point $(\mathbf{x}_{n+1}, y_{n+1}) := (\mathbf{x}^*, \hat{f}(\mathbf{x}^*))$ where $\hat{f}$ is the given regressor, $\bar{f}_{n+1}(\mathbf{x}^*)$ can be written as (see Rasmussen and Williams (2006, Page 17))

$$\bar{f}_{n+1}(\mathbf{x}^*) = \boldsymbol{\alpha}_1 k(\mathbf{x}^*, \mathbf{x}_1) + \boldsymbol{\alpha}_2 k(\mathbf{x}^*, \mathbf{x}_2) + \ldots + \boldsymbol{\alpha}_n k(\mathbf{x}^*, \mathbf{x}_n) + \boldsymbol{\alpha}_{n+1} k(\mathbf{x}^*, \mathbf{x}^*), \tag{33}$$

where

$$\sum_{i=1}^{n} \boldsymbol{\alpha}_i k(\mathbf{x}_j, \mathbf{x}_i) + \boldsymbol{\alpha}_{n+1} k(\mathbf{x}_j, \mathbf{x}^*) + \sigma^2 \boldsymbol{\alpha}_j = y_j$$

for every $j \in [n]$, and

$$\sum_{i=1}^{n} \boldsymbol{\alpha}_i k(\mathbf{x}^*, \mathbf{x}_i) + k(\mathbf{x}^*, \mathbf{x}^*) \boldsymbol{\alpha}_{n+1} + \sigma^2 \boldsymbol{\alpha}_{n+1} = \hat{f}(\mathbf{x}^*). \tag{34}$$

From (33) and (34) we deduce $\bar{f}_{n+1}(\mathbf{x}^*) = \hat{f}(\mathbf{x}^*) - \sigma^2\boldsymbol{\alpha}_{n+1}$. By using this relation and setting $(\mathbf{x}_{n+1}, y_{n+1}) := (\mathbf{x}^*, \hat{f}(\mathbf{x}^*))$ in (32) we derive

$$\hat{f}(\mathbf{x}^*) - \sigma^2\boldsymbol{\alpha}_{n+1} = \bar{f}_n(\mathbf{x}^*) + (\hat{f}(\mathbf{x}^*) - \bar{f}_n(\mathbf{x}^*))\frac{\mathcal{V}_n(f(\mathbf{x}^*))}{\mathcal{V}_n(f(\mathbf{x}^*)) + \sigma^2}.$$

We then obtain the following

$$|\hat{f}(\mathbf{x}^*) - \bar{f}_n(\mathbf{x}^*)| = |\boldsymbol{\alpha}_{n+1}|\left(\mathcal{V}_n(f(\mathbf{x}^*)) + \sigma^2\right).$$

We can bound $|\boldsymbol{\alpha}_{n+1}|$ in the same way as in Lemma 5, leading to the following bound for the bias

$$|\hat{f}(\mathbf{x}^*) - \bar{f}_n(\mathbf{x}^*)| \leq \frac{|\hat{f}(\mathbf{x}^*)| + \sqrt{\|\mathbf{y}\|^2 + \hat{f}(\mathbf{x}^*)^2}}{2\sigma^2}\left(\mathcal{V}_n(f(\mathbf{x}^*)) + \sigma^2\right). \tag{35}$$

□

A.10 Derivation of C

This section derives the matrix $\mathbf{C}$ for the RBF, rational quadratic, and Matern kernels used in the experiments.

Remark 1. *For the RBF kernel* $k(\mathbf{p}, \mathbf{q}) = \sigma_0^2 \exp\left(-\frac{1}{2}\sum_{j=1}^p (\mathbf{p}_j - \mathbf{q}_j)^2/l_j^2\right)$*, we have* $\mathbf{C}_{ij} = \sigma_0^2/l_i^2$ *for* $i = j$ *and* $\mathbf{C}_{ij} = 0$ *for* $i \neq j$*.*

Proof. Let $\kappa(\mathbf{d}) = \sigma_0^2 \exp\left(-\sum_{j=1}^p \frac{\mathbf{d}_j^2}{2l_j^2}\right)$. Then:

$$\frac{\partial\kappa}{\partial d_i}(\mathbf{d}) = -\frac{\mathbf{d}_i}{l_i^2}\kappa(\mathbf{d})$$

$$\frac{\partial^2\kappa}{\partial d_i \partial d_j}(\mathbf{d}) = \frac{\mathbf{d}_i}{l_i^2}\frac{\mathbf{d}_j}{l_j^2}\kappa(\mathbf{d}) \tag{36}$$

$$\frac{\partial^2\kappa}{\partial d_i^2}(\mathbf{d}) = \frac{\mathbf{d}_i^2}{l_i^4}\kappa(\mathbf{d}) - \frac{1}{l_i^2}\kappa(\mathbf{d}) \tag{37}$$

From Lemma 2 we have: $\mathbf{C}_{ij} = -\frac{\partial^2\kappa}{\partial d_i \partial d_j}(\mathbf{0})$. The claim follows by substituting $\mathbf{d} = \mathbf{0}$ to Eq. (36) and (37). □

Remark 2. *For the rational quadratic kernel* $k(\mathbf{p}, \mathbf{q}) = \sigma_0^2\left(1 + \|\mathbf{p} - \mathbf{q}\|^2/(2\alpha l^2)\right)^{-\alpha}$ *with* $\alpha > 0$*, we have* $\mathbf{C}_{ij} = \sigma_0^2/l^2$ *for* $i = j$ *and* $\mathbf{C}_{ij} = 0$ *for* $i \neq j$*.*

Proof. Denote $\kappa(\mathbf{d}) = \sigma_0^2\left(1 + \frac{\|\mathbf{d}\|^2}{2\alpha l^2}\right)^{-\alpha}$. Then:

$$\frac{\partial\kappa}{\partial d_i}(\mathbf{d}) = -\frac{\mathbf{d}_i}{l^2}\sigma_0^2\left(1 + \frac{\|\mathbf{d}\|^2}{2\alpha l^2}\right)^{-\alpha-1}$$

$$\frac{\partial^2\kappa}{\partial d_i \partial d_j}(\mathbf{d}) = \frac{\mathbf{d}_i}{l^2}\frac{\mathbf{d}_j}{l^2}\sigma_0^2\left(1 + \frac{\|\mathbf{d}\|^2}{2\alpha l^2}\right)^{-\alpha-2} \tag{38}$$

$$\frac{\partial^2\kappa}{\partial d_i^2}(\mathbf{d}) = \frac{\mathbf{d}_i}{l^2}\frac{\mathbf{d}_j}{l^2}\sigma_0^2\left(1 + \frac{\|\mathbf{d}\|^2}{2\alpha l^2}\right)^{-\alpha-2} - \frac{\sigma_0^2}{l^2}\left(1 + \frac{\|\mathbf{d}\|^2}{2\alpha l^2}\right)^{-\alpha-1} \tag{39}$$

From Lemma 2 we have: $\mathbf{C}_{ij} = -\frac{\partial^2\kappa}{\partial d_i \partial d_j}(\mathbf{0})$. The claim follows by substituting $\mathbf{d} = \mathbf{0}$ to Eq. (38) and (39). □

Remark 3. *For the Matern kernel* $k(\mathbf{p}, \mathbf{q}) = \sigma_0^2 \frac{1}{\Gamma(\nu)2^{\nu-1}}\left(\frac{\sqrt{2\nu}}{l}\|\mathbf{p} - \mathbf{q}\|\right)^\nu K_\nu\left(\frac{\sqrt{2\nu}}{l}\|\mathbf{p} - \mathbf{q}\|\right)$ *with* $\nu = 5/2$*, we have* $\mathbf{C}_{ij} = \frac{5}{3}\sigma_0^2/l^2$ *for* $i = j$ *and* $\mathbf{C}_{ij} = 0$ *for* $i \neq j$

Proof. For $\nu = 5/2$, the kernel can be simplified to: $\kappa(\mathbf{d}) = \sigma_0^2 \left(1 + \frac{\sqrt{5}\|\mathbf{d}\|}{l} + \frac{5\|\mathbf{d}\|^2}{3l^2}\right) \exp\left(-\frac{\sqrt{5}\|\mathbf{d}\|}{l}\right)$. Then,

$$\frac{\partial \kappa}{\partial d_i}(\mathbf{d}) = \sigma_0^2 \exp\left(-\frac{\sqrt{5}\|\mathbf{d}\|}{l}\right)\left(-\frac{5}{3l^2}\mathbf{d}_i - \frac{5\sqrt{5}}{3l^3}\|\mathbf{d}\|\mathbf{d}_i\right).$$

For $i = j$,

$$\frac{\partial^2 \kappa}{\partial d_i^2}(\mathbf{d}) = \sigma_0^2 \exp\left(-\frac{\sqrt{5}\|\mathbf{d}\|}{l}\right)\left(\frac{25}{3l^4}\mathbf{d}_i^2 - \frac{5}{3l^2} - \frac{5\sqrt{5}}{3l^3}\|\mathbf{d}\|\right),$$

so $\mathbf{C}_{ii} = (5\sigma_0^2)/(3l^2)$. For $j \neq i$,

$$\frac{\partial^2 \kappa}{\partial d_i \partial d_j}(\mathbf{d}) = \frac{25\sigma_0^2}{3l^4} \exp\left(-\frac{\sqrt{5}\|\mathbf{d}\|}{l}\right)\mathbf{d}_i\mathbf{d}_j,$$

it follows that $\mathbf{C}_{ij} = 0$ if $i \neq j$.

□

B EXPERIMENTAL RESULTS

B.1 Datasets

The real datasets used in the experiments are shown in Table 4. The datasets are available in the UCI repository (Dua and Graff, 2017) or OpenML (Vanschoren et al., 2014). The molecular datasets QM9, AA, OE62 are available in Zenodo by Stuke et al. (2019).

Table 4: Datasets used in the experiments. N: data size, N_{tr}: train data size, p: dimensions.

Dataset	N	N_{tr}	p
AA	10000	3000	30
ABALONE	4177	1253	10
AIRFOIL	1503	450	5
AIRQUALITY	7355	2206	11
AUTOMPG	392	117	9
CALIFORNIA	20640	6192	8
CONCRETE	1030	309	8
CPU_SMALL	8192	2457	12
OE62	10000	3000	30
QM9	10000	3000	30
WINEQUALITY	6497	1949	13
YEARPREDICTIONMSD	10000	3000	90

We briefly describe the datasets and how they were preprocessed. For all datasets, the covariates and the target are scaled to zero mean and unit variance. The OpenML ID, when available, is in parentheses after the name.

- AA (Ropo et al., 2016) contains 64,710 amino acids and dipeptides described by 27,200-dimensional MBTR features, as computed in Stuke et al. (2019); the target is the highest occupied molecular orbital (HOMO) energy. We use a subset containing the 10,000 heaviest molecules and the first 30 PCA components of the MBTR features.

- ABALONE (Warwick Nash, 1994) (183) contains physical measurements of abalone snails; the target is their age.

- AIRFOIL (Thomas Brooks, 1989) (43919) contains NACA airfoils in different experimental conditions, such as wind tunnel speed and angle of attack; the target is the scaled sound pressure level.

- AIRQUALITY (Vito, 2008) contains hourly measurements of various gas concentrations; the target is the CO concentration. It is preprocessed as in Oikarinen et al. (2021).

- AUTOMPG (R. Quinlan, 1993) (196) contains cars described by attributes such as horsepower, origin and model year; the target is fuel consumption. We removed missing values.

- CALIFORNIA (Pedregosa et al., 2011) contains housing districts in California, described by attributes such as median house age and population; the target is the median house value.

- CONCRETE (I-Cheng Yeh, 1998) (4353) contains measurements on different types of concrete; the target is the compressive strength.

- CPU_SMALL (562) contains computer systems activity measures; the target is the portion of time that CPUs run in user mode.

- OE62 (Stuke et al., 2020) contains 44,004 opto-electronically active molecules described by 13,200-dimensional MBTR features; the target is the HOMO energy. We use a subset containing the 10,000 heaviest molecules and the first 30 PCA components of the MBTR features.

- QM9 (Ramakrishnan et al., 2014) contains 133,814 small organic molecules described by 3000-dimensional MBTR features; the target is the HOMO energy. We use a subset containing the 10,000 heaviest molecules and the first 30 PCA components of the MBTR features.

- YEARPREDICTIONMSD (T. Bertin-Mahieux, 2011) (44027) contains songs described by audio features; the target is the release year of the song. We use the first 10,000 rows.

- WINEQUALITY (Paulo Cortez, 2009) (287) contains red and white wines described by attributes such as pH and alcohol level; the target is wine quality.

B.2 Bound evaluation with synthetic data

Setup. We use $N_{tr} = 50$ training data points to fit the regressor $\hat{f}$ and the bounds $\hat{\mathcal{V}}$ and $\hat{\mathcal{B}}$. The regressor and the bounds are evaluated on $N_{te} = 100$ testing data points, for which we know the true targets y, posterior variance $\mathcal{V}$ and bias $\mathcal{B}$. The target y is sampled from a GP with a known kernel (RBF, RQ, Matern) that has $\mathbf{C} = 1$. The training data covariates are p-dimensional and sampled i.i.d. from $\mathrm{N}(0, 1/\sqrt{p})$. The test data lie along the axis of the first dimension from zero to π, to cover the domain of the $\sin_*$ function.

Results. Table 5 shows the variance and the bias bound of Theorem 5 under various dimensions p, kernels, noise levels σ^2, and regressors (RF: random forest regressor, SVM: support vector regressor), extending Tables 1 and 2 (without their baselines). Figure 8 shows the same bounds along with their baseline bounds, described in Tables 1 and 2. Notice that Table 5 and Figure 8 show absolute differences rather than relative ones, as in Tables 1 and 2, to highlight the maximum difference for the variance, which is equal to $\sigma_0^2 = 1$, and to avoid division by zero bias when using a GP regressor ($\mathcal{B}_{\mathrm{GP}} = 0$).

Table 5: Bound evaluation results, extending Tables 1 and 2 (without baseline bounds). p: dimension, σ^2: noise variance. In parentheses are 5% quantiles. Using $\mathbf{C} = 1$, $\sigma_0^2 = 1$. This table's entries are also in Figure 8 as black dots, along with baseline bounds.

#	p	Kernel	σ^2	$\hat{\mathcal{V}} - \mathcal{V}$	$\hat{\mathcal{B}}_{\text{Thm. 5}} - \mathcal{B}_{\mathrm{GP}}$	$\hat{\mathcal{B}}_{\text{Thm. 5}} - \mathcal{B}_{\mathrm{RF}}$	$\hat{\mathcal{B}}_{\text{Thm. 5}} - \mathcal{B}_{\mathrm{SVM}}$
1	1	RBF	0.1	0.1 (0.1)	4.3 (2.2)	4.0 (2.0)	3.8 (2.1)
2	1	RQ	0.1	0.1 (0.1)	4.0 (2.2)	3.7 (1.9)	4.0 (2.1)
3	1	Matern	0.1	0.1 (0.1)	4.0 (2.0)	3.8 (1.9)	3.8 (2.1)
4	1	RBF	1.0	0.4 (0.4)	5.6 (4.2)	5.0 (3.6)	5.5 (3.9)
5	1	RQ	1.0	0.4 (0.4)	5.4 (4.0)	5.0 (3.6)	5.3 (3.9)
6	1	Matern	1.0	0.4 (0.4)	5.4 (3.8)	4.9 (3.6)	5.4 (3.8)
7	2	RBF	0.1	0.1 (0.1)	5.9 (3.3)	5.7 (2.8)	5.4 (3.0)
8	2	RQ	0.1	0.1 (0.1)	5.3 (3.0)	5.6 (2.9)	5.4 (2.9)
9	2	Matern	0.1	0.1 (0.1)	5.7 (3.0)	5.6 (3.1)	5.8 (3.1)
10	2	RBF	1.0	0.4 (0.3)	6.7 (4.6)	6.0 (4.2)	6.1 (4.4)
11	2	RQ	1.0	0.4 (0.3)	6.6 (4.7)	6.2 (4.4)	6.2 (4.4)
12	2	Matern	1.0	0.4 (0.3)	6.6 (4.7)	6.1 (4.3)	6.1 (4.4)
13	10	RBF	0.1	0.3 (0.2)	11.5 (7.3)	11.5 (7.2)	11.5 (7.3)
14	10	RQ	0.1	0.3 (0.2)	11.3 (7.1)	11.0 (6.9)	11.7 (7.2)
15	10	Matern	0.1	0.3 (0.2)	11.2 (7.2)	11.4 (6.7)	10.8 (6.7)
16	10	RBF	1.0	0.4 (0.2)	9.1 (6.9)	8.8 (6.6)	8.9 (6.8)
17	10	RQ	1.0	0.4 (0.3)	9.0 (6.5)	8.5 (6.2)	8.6 (6.3)
18	10	Matern	1.0	0.4 (0.3)	8.9 (6.6)	8.7 (6.4)	8.9 (6.3)
19	20	RBF	0.1	0.4 (0.2)	13.3 (8.8)	13.1 (8.8)	13.1 (8.7)
20	20	RQ	0.1	0.4 (0.3)	13.1 (8.3)	12.9 (8.2)	12.8 (8.4)
21	20	Matern	0.1	0.4 (0.3)	13.6 (8.6)	13.1 (8.5)	13.1 (8.1)
22	20	RBF	1.0	0.4 (0.2)	9.7 (7.3)	9.4 (7.2)	9.8 (7.3)
23	20	RQ	1.0	0.4 (0.3)	9.7 (7.1)	9.5 (6.9)	9.7 (7.1)
24	20	Matern	1.0	0.5 (0.3)	9.7 (7.2)	9.3 (7.1)	9.5 (7.1)

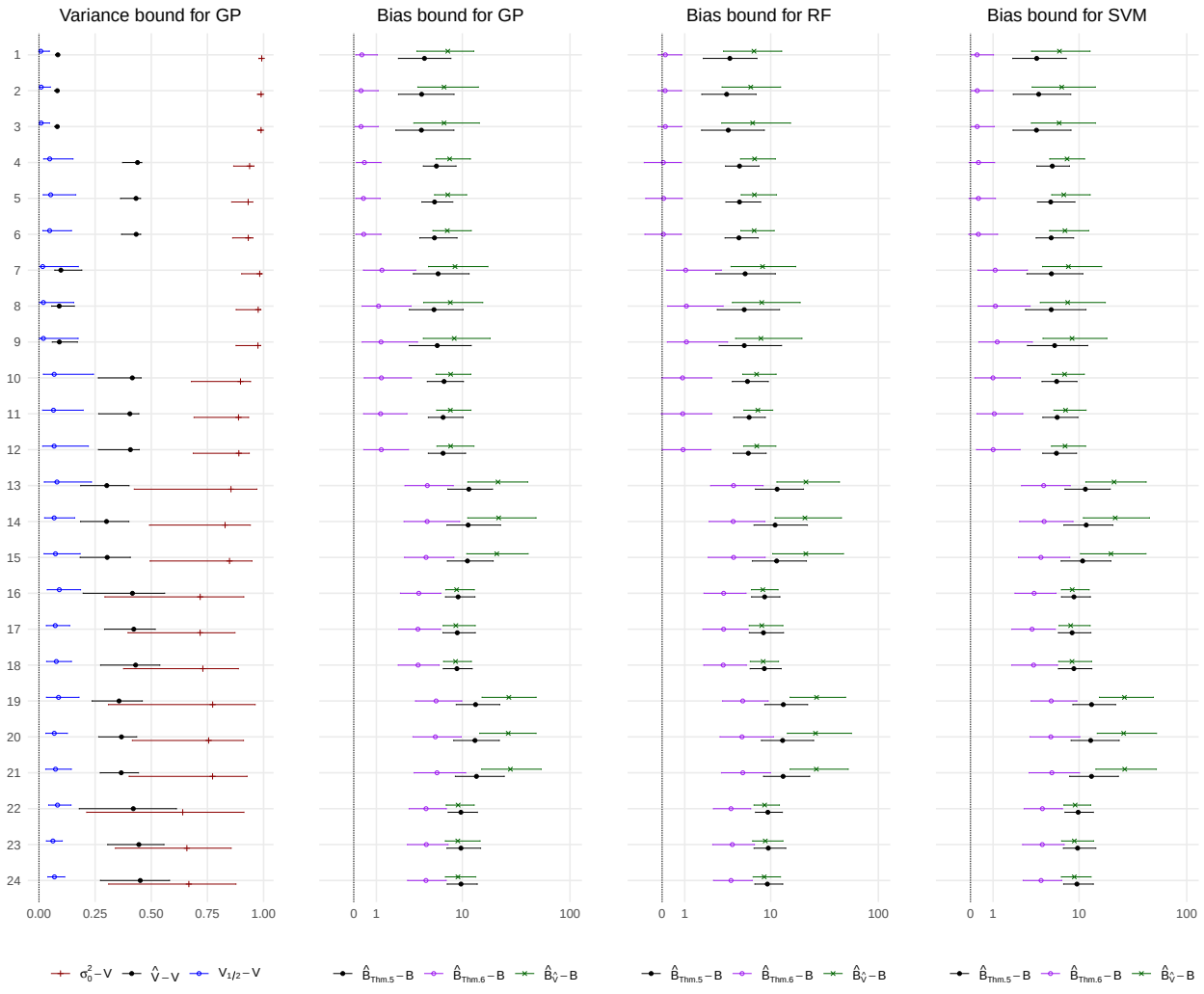

Figure 8: Bound evaluation results, extending Tables 1 and 2. Error bars span from 5% to 95% quantiles. The vertical ordering corresponds to Table 5. The values of Table 5 are shown as black dots. Note: bias bound x-axis is in (pseudo) log-scale.

B.3 Real data

Setup for bound evaluation with real data. The data are split randomly 30–70 into train-test 100 times. The reported values are the median over test points and repeated splits. The regressor hyperparameters are the defaults in `sklearn` 1.2.1. The bound parameters are $\sigma_0^2 = 1$ and $\sigma^2 = \text{MSE}_{test}$.

Setup for drift experiment. The data are split 30-30-40 into train-validation-test sets. The split is made to induce drift, meaning high errors on the test set, as follows. A random forest is trained on the whole data, which is then split using the variable with the highest impurity-based feature importance, i.e., the variable with the largest mean impurity decrease (which is variance reduction for regression) over all splits involving that variable. Figure 9 plots, for this split, the squared error against the bound $U = \hat{\mathcal{V}} + \hat{\mathcal{B}}^2_{\text{Thm.6}}$. The error threshold for drift is shown with a vertical line and is equal to the 80% quantile of the errors on the validation set. The regressor hyperparameters are selected by grid search with 5-fold cross-validation RMSE as a score. The bound parameters are $\sigma_0^2 = 1$ and $\sigma^2 = \text{MSE}_{CV}$ of the regressor.

Estimating C. For the real data experiments, we don't have prior knowledge about $\mathbf{C}$, so it is estimated from the training data as a diagonal matrix using the regressor $\hat{f}$'s gradient, approximated with finite-differences and averaged over N_{tr} training data points:

$$\mathbf{C}_{jj} \approx \frac{1}{N_{tr}} \sum_{i=1}^{N_{tr}} \left(\frac{\hat{f}(\mathbf{x}_i + \mathbf{e}_j h) - \hat{f}(\mathbf{x}_i)}{h} \right)^2 \tag{40}$$

where $\mathbf{e}_j$ denotes the unit vector of the j:th axis, $h = 10^{-6}$ for $\hat{f} = \text{SVM}$ and $h = 10^{-2}$ for $\hat{f} = \text{random forest}$. This estimate assumes: (i) the regressor gradient is close to the true function's gradient around the training points, (ii) the regressor gradient approximation has low error, (iii) the data are translation invariant, so that averaging over data points is equivalent to averaging over prior samples, (iv) the data contain enough gradient information in all directions, i.e., there are enough pairwise distances that align with all axes.

B.4 Scaling

Table 6 shows the numerical values of the scaling experiment of Figure 6.

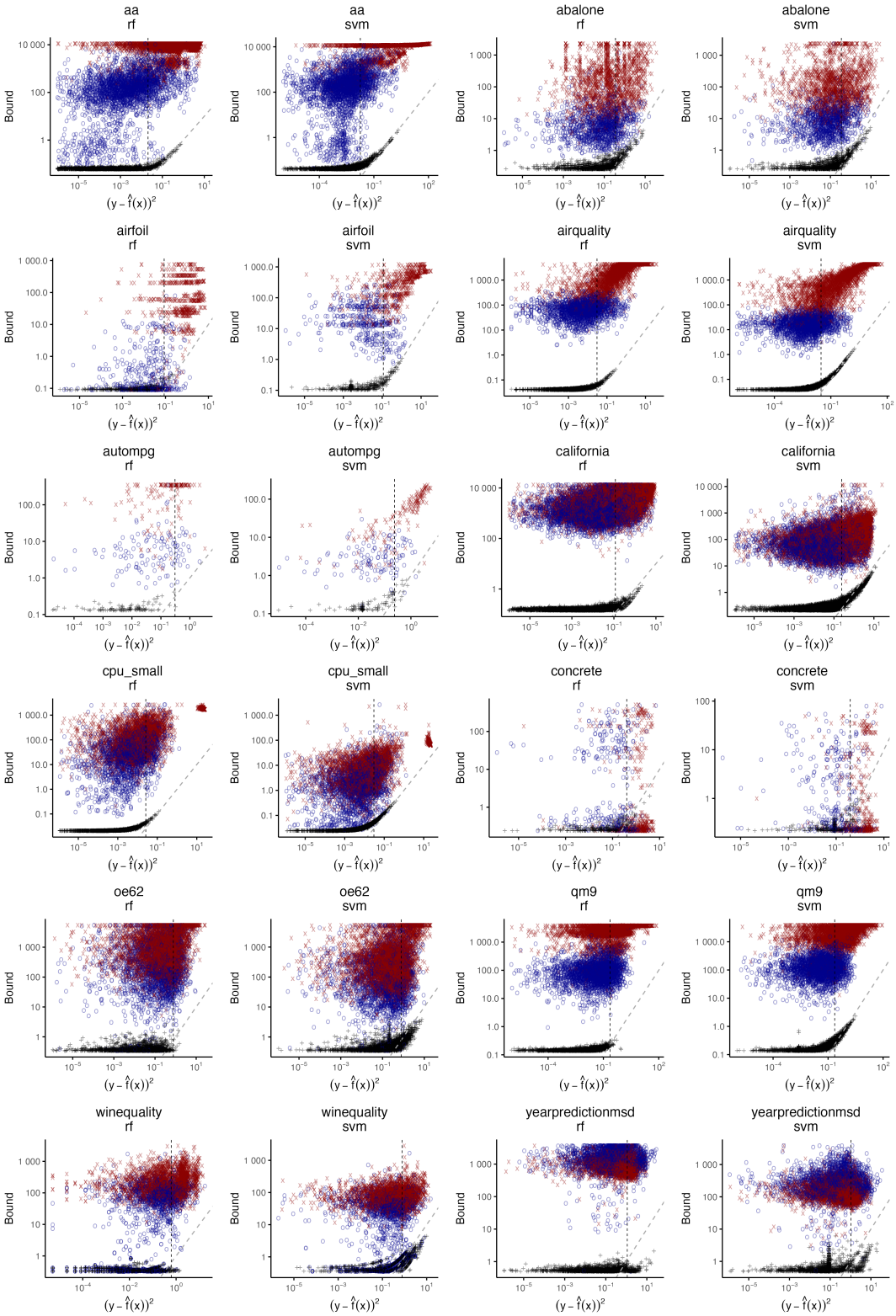

Figure 9: Squared error vs. the bound U for real datasets. Black crosses are train points, blue circles are validation points, red x's are test points. The vertical line denotes the error threshold for drift (80% quantile of validation errors).

Table 6: Numerical values of scaling results in seconds (average ± standard deviation over 100 runs) from Figure 6.

N	D	Bound	Matern	RBF	RQ
100	1	0.0084 ± 0.023	0.017 ± 0.017	0.031 ± 0.048	0.021 ± 0.021
100	2	0.014 ± 0.014	0.0028 ± 0.00047	0.0022 ± 0.00031	0.0021 ± 0.00034
100	5	0.0014 ± 0.0015	0.0027 ± 0.00054	0.0021 ± 0.00036	0.002 ± 0.00037
100	15	0.00076 ± 0.00026	0.0028 ± 0.00046	0.0023 ± 0.00048	0.002 ± 0.00036
100	25	0.00099 ± 0.00073	0.0028 ± 0.00045	0.0024 ± 0.00045	0.0021 ± 0.00037
500	1	0.0082 ± 0.0012	0.037 ± 0.0058	0.031 ± 0.011	0.023 ± 0.0055
500	2	0.0065 ± 0.00069	0.037 ± 0.0046	0.026 ± 0.0025	0.026 ± 0.005
500	5	0.0075 ± 0.00085	0.036 ± 0.0051	0.025 ± 0.0022	0.024 ± 0.0045
500	15	0.0096 ± 0.00083	0.038 ± 0.0043	0.03 ± 0.0024	0.027 ± 0.0039
500	25	0.01 ± 0.0012	0.042 ± 0.0048	0.032 ± 0.0026	0.029 ± 0.0044
1000	1	0.032 ± 0.0081	0.17 ± 0.02	0.12 ± 0.012	0.11 ± 0.014
1000	2	0.034 ± 0.0083	0.16 ± 0.02	0.12 ± 0.011	0.11 ± 0.014
1000	5	0.039 ± 0.0078	0.17 ± 0.021	0.12 ± 0.011	0.11 ± 0.014
1000	15	0.046 ± 0.0072	0.18 ± 0.023	0.14 ± 0.012	0.12 ± 0.015
1000	25	0.048 ± 0.0075	0.19 ± 0.024	0.15 ± 0.014	0.13 ± 0.016
2500	1	0.21 ± 0.041	1.4 ± 0.17	1.2 ± 0.1	1.1 ± 0.13
2500	2	0.22 ± 0.04	1.5 ± 0.18	1.2 ± 0.1	1.1 ± 0.11
2500	5	0.23 ± 0.035	1.5 ± 0.18	1.2 ± 0.1	1.1 ± 0.13
2500	15	0.28 ± 0.034	1.5 ± 0.18	1.3 ± 0.086	1.2 ± 0.13
2500	25	0.3 ± 0.049	1.6 ± 0.18	1.3 ± 0.078	1.2 ± 0.14
5000	1	0.78 ± 0.17	8.2 ± 1	6.8 ± 0.49	6.8 ± 0.58
5000	2	0.74 ± 0.14	8.3 ± 0.99	6.8 ± 0.53	6.9 ± 0.6
5000	5	0.83 ± 0.15	8.2 ± 0.95	6.8 ± 0.48	6.9 ± 0.62
5000	15	1 ± 0.13	8.4 ± 0.95	7.1 ± 0.53	7.1 ± 0.6
5000	25	1.1 ± 0.16	8.6 ± 0.96	7.2 ± 0.54	7.3 ± 0.63
7500	1	1.7 ± 0.31	24 ± 2.6	20 ± 1.5	21 ± 1.8
7500	2	1.7 ± 0.29	24 ± 2.6	20 ± 1.4	21 ± 1.7
7500	5	1.8 ± 0.27	24 ± 2.6	20 ± 1.4	21 ± 1.7
7500	15	2.3 ± 0.31	24 ± 2.6	21 ± 1.5	22 ± 1.8
7500	25	2.3 ± 0.36	25 ± 2.6	21 ± 1.4	22 ± 1.8
10000	1	2.9 ± 0.5	53 ± 5.6	45 ± 3.1	47 ± 3.6
10000	2	3 ± 0.47	53 ± 5.6	45 ± 3.1	47 ± 3.6
10000	5	3.2 ± 0.46	53 ± 5.7	45 ± 3.1	47 ± 3.6
10000	15	4 ± 0.54	53 ± 5.5	47 ± 3	48 ± 3.6
10000	25	4 ± 0.58	54 ± 5.3	47 ± 3.1	49 ± 3.7