Spectral Representation of Robustness Measures for Optimization Under Input Uncertainty

Jixiang Qing [1] **Tom Dhaene** [1] **Ivo Couckuyt** [1]

Abstract

We study the inference of mean-variance robustness measures to quantify input uncertainty under the Gaussian Process (GP) framework. These measures are widely used in applications where the robustness of the solution is of interest, for example, in engineering design. While the variance is commonly used to characterize the robustness, Bayesian inference of the variance using GPs is known to be challenging. In this paper, we propose a *Spectral Representation of Robustness Measures* based on the GP's spectral representation, i.e., an analytical approach to approximately infer both robustness measures for normal and uniform input uncertainty distributions. We present two approximations based on different Fourier features and compare their accuracy numerically. To demonstrate their utility and efficacy in robust Bayesian Optimization, we integrate the analytical robustness measures in three standard acquisition functions for various robust optimization formulations. We show their competitive performance on numerical benchmarks and real-life applications.

1. Introduction

Bayesian **O**ptimization (BO) is a well-established approach for solving expensive non-convex black-box optimization problems efficiently (Frazier, 2018; Shahriari et al., 2015). Its agnostic treatment of black-box functions has made it applicable in many real-life application domains. Starting with a limited number of training data, BO constructs a probabilistic *surrogate model*, with a Gaussian Process (GP) as the common choice (Rasmussen, 2003), to provide a computationally efficient approximation to the problem. By leveraging an *acquisition function* based on the model posterior, BO is able to efficiently quantify the potential of candidate solutions and will only query the most interesting candidate(s) using the expensive black-box function. After querying the new data, BO updates the posterior of the surrogate model and searches for the next most interesting candidate(s). Eventually, the deemed optimal solution is recommended to the practitioners.

Living in a chaotic world full of undetermined events, we strive for, yet are far from, obtaining an exact representation of the world. When ignored, uncertainties can degrade the performance of the chosen optimal solutions significantly. Among different types of uncertainties, **input uncertainties**, for instance, robotics position error (Nogueira et al., 2016), manufacturing tolerances (Cui et al., 2021), and imprecise measurements (Wu et al., 2021), are of practical interest in this work. To mitigate the side effects of uncertainties and aim for providing a robust solution, a sensible approach is *robust optimization*. In robust optimization, the goal is to maximize an objective function which is formulated upon robustness measure(s), and, thus, that is able to take uncertainty into account (Beyer & Sendhoff, 2007).

Two common robustness measures (Manfredi & Trinchero, 2021; Wauters et al., 2020; Hoque & Low, 2020; Markowitz & Todd, 2000) are the **mean** (a.k.a., Bayes risk, expected regret (Beland & Nair, 2017; Oliveira et al., 2019)) and the **variance**, representing the first raw moment and second central moment of the objective function distribution induced by input uncertainty, respectively. Without loss of generality, they can be defined as[1]:

$$\begin{aligned} \mathbb{J}_{\boldsymbol{\xi}}(f) &= \mathbb{E}_{\boldsymbol{\xi}}\left(f(\boldsymbol{x}+\boldsymbol{\xi})\right) \\ \mathbb{V}_{\boldsymbol{\xi}}(f) &= \mathbb{E}_{\boldsymbol{\xi}}\left(f^2(\boldsymbol{x}+\boldsymbol{\xi})\right) - \left[\mathbb{E}_{\boldsymbol{\xi}}\left(f(\boldsymbol{x}+\boldsymbol{\xi})\right)\right]^2 \end{aligned} \tag{1}$$

where the random variable $\boldsymbol{\xi}$ denotes a pre-defined additive

[1]Ghent University, – imec, IDLab, Department of Information Technology (INTEC), Tech Lane – Zwijnaarde 126, 9052 Ghent, Belgium. Correspondence to: Jixiang Qing <Jixiang.Qing@UGent.be>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

[1]Unless needed for clarity, we drop the subscript $\boldsymbol{\xi}$ for robustness measure notation hereafter.

input uncertainty[2] on candidate $\boldsymbol{x}$ and f represents the black-box objective function.

Pursuing optimal solutions being robust against input uncertainties is of practical interest in Bayesian Optimization, which we term as **R**obust **B**ayesian **O**ptimization (RBO). In this context, a proper approach to infer these non-observable robustness measures is of immediate interest. Unfortunately, while it has been well-known that the posterior of the mean[3] is tractable under the GP framework (O'Hagan, 1991; Papoulis & Pillai, 2002), the variance is much more difficult to infer. Practitioners have commonly used a *degenerated variance approximation*, i.e., performing Monte Carlo sampling on the GP posterior mean (e.g., Rivier & Congedo (2018)) to approximate the variance as a point estimation. Nevertheless, as we will theoretically elaborate later, it is unfortunately a biased estimation of the variance distribution's first moment.

Contribution We provide an effective inference scheme for the mean and the variance distributions under the GP framework, supporting normal and uniform input uncertainty distribution forms. This has been achieved using an analytical calculation of the robustness measures through the spectral representation of the GP, which is based on the kernel approximation using Fourier features. We derive two different expressions for the robustness measures based on Fourier features (FF) methods, namely Random Fourier Features (RFF) (Rahimi et al., 2007) and Quadrature Fourier Features (QFF) (Mutnỳ & Krause, 2019), as shown in Fig. 1. We empirically compare their accuracy through synthetic uncertainty calibration experiments. As a direct application of this scheme, we elaborate on how to utilize these robustness measures in RBO. To that end, we extend three standard acquisition functions for handling robustness, corresponding to three common robust optimization formulations. We provide empirical and real-life benchmarking of these acquisition functions, demonstrating their competitive performance against relevant methods for recommending robust optimal solutions.

Related Work Bayesian inference of the mean $\mathbb{J}_{\boldsymbol{\xi}}(f)$ and variance $\mathbb{V}_{\boldsymbol{\xi}}(f)$ is of practical interest in the GP literature. Pioneer works (Girard et al., 2002; McHutchon & Rasmussen, 2011) focus on approximating the propagated objective function distribution. However, the objective function distribution generally does not disentangle the model and input uncertainties. Regarding variance inference, O'Hagan (2011) provides an approach to calculate its first two moments, though expressions still rely on approximating a set of integrals. Iwazaki et al. (2021) consider the task of performing robust pool-based active learning under environmental uncertainty, where confidence intervals for both mean and variance are derived and utilized in the formulation of acquisition functions. However, the confidence interval does not provide a principled way of inspecting the distribution form (e.g., through sampling the distribution). In addition, while an extension for continuous input spaces is proposed, the continuous input uncertainty scenario is not considered.

Fruitful RBO research has been conducted solely based on the mean distribution as it is preserved as a GP. Regarding the acquisition functions for RBO, the expected improvement (Beland & Nair, 2017), its modified formulation (Nogueira et al., 2016), and other common acquisition functions (Oliveira et al., 2019; Wang et al., 2020a) are considered. Fröhlich et al. (2020) also sample an expression for the mean using Fourier features in entropy search, but is restricted to independent normal uncertainty distributions, while we derive an extended expression for generic multivariate normal input uncertainties. Besides the common mean measure, input uncertainty has also been considered in more conservative approaches such as the adversarial form (Bogunovic et al., 2018; Sessa et al., 2020; Weichert & Kister, 2021), quantile form (Torossian et al., 2020) value-at-risk and conditional value-at-risk form (Cakmak et al., 2020; Nguyen et al., 2021a;b).

Other types of robustness are also extensively investigated in RBO such as the robustness to model mis-specification (Bogunovic & Krause, 2021; Neiswanger & Ramdas, 2021). Finally, more generic types of input uncertainties are also investigated in (Kirschner et al., 2020; Rahimian & Mehrotra, 2019), and heteroscedastic noise is considered in (Makarova et al., 2021).

2. Preliminaries

2.1. Weight-Space Approximations of Gaussian Process

GP is well-known to be a generalization of a parametric *Bayesian linear model* in *weight-space* (Rasmussen, 2003; Wilson et al., 2020). Starting with a linear model $f(\boldsymbol{x}) = \boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{\theta}$, where $\boldsymbol{\phi}(\boldsymbol{x})$ represents the feature mapping function with weights $\boldsymbol{\theta}$, the kernel trick (Schölkopf et al., 2002) can be utilized to replace a feature mapping function with infinite features by a kernel. This transformation has endowed advantages, including the freedom of parametric feature functions and flexible expressivity scaling with training data. Nevertheless, mainly encouraged by avoiding the expensive kernel inversion in big data scenarios, one may still want to look back to its weight-space formulation (Lázaro-Gredilla et al., 2010),

[2] It is straightforward to extend this formulation, as well as our approaches proposed subsequently, to more generic problems containing *environmental variables*, see discussion in appendix E.

[3] For clarity, we use *GP posterior mean* and *GP posterior variance* to represent the mean and the variance of the conditional distribution of the GP posterior.

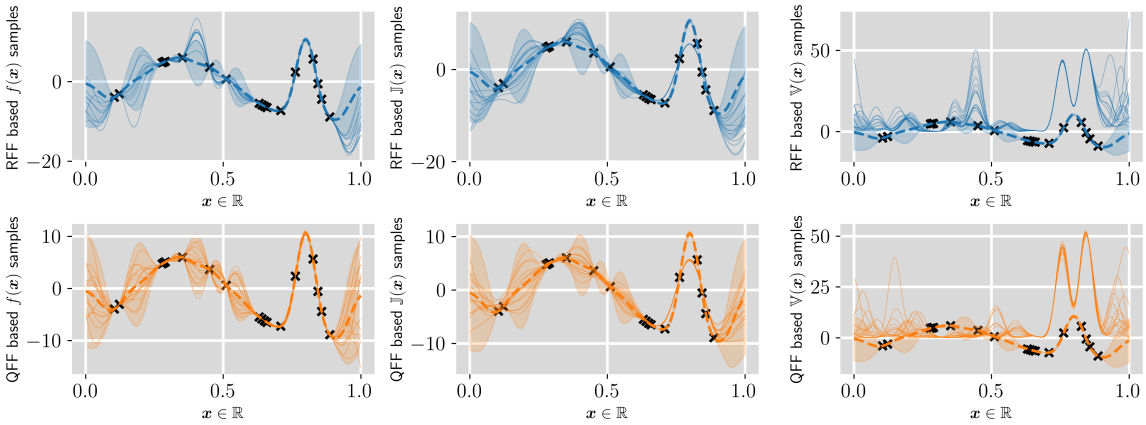

Figure 1. Comparison of different sample trajectories based on Fourier features (RFF and QFF) on a synthetic function: $f = -20\left[0.5 - 0.8\exp\left(-(\frac{x-0.35}{0.25})^2\right) - \exp\left(-(\frac{x-0.8}{0.05})^2\right)\right]$. With 256 Fourier features and uniform input uncertainty $\xi \sim U(-0.05, 0.05)$. Black crosses are training data, dashed line represents GP posterior mean, shaded areas are 95 confidence intervals of GP posterior, and the plotted trajectories represent posterior samples based on Fourier features.

Fourier Feature based Kernel Approximation The common way of looking back to weight-space is through degenerate kernels in a parametric form through approximation. We remark that Bochner's theorem (e.g., Stein (1999)) implies that any stationary kernel k and its spectral density $S(\omega)$ are Fourier dual as shown in the r.h.s. of Eq. 2. After scaling $S(\omega)$ to a probability distribution (Lázaro-Gredilla et al., 2010), the expectation integral can be approximated by a finite number of *Fourier features* and subsequently be written as an inner product (Rahimi et al., 2007):

$$\begin{aligned} k(\boldsymbol{x}, \boldsymbol{x}') &= \int S(\boldsymbol{\omega}) e^{j\boldsymbol{\omega}^T L(\boldsymbol{x}-\boldsymbol{x}')} d\boldsymbol{\omega} \\ &= \sigma^2 \int p(\boldsymbol{\omega}) \cos\left(\boldsymbol{\omega}^T L(\boldsymbol{x}-\boldsymbol{x}')\right) d\boldsymbol{\omega} \\ &\approx \boldsymbol{\phi}^T(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{x}') \end{aligned} \tag{2}$$

where $\boldsymbol{\phi}(\boldsymbol{x}) := [\phi_1(\boldsymbol{x}), \phi_2(\boldsymbol{x}), ..., \phi_{N_f}(\boldsymbol{x})]^T$ represents the stacked N_f parametric feature functions approximating the integration, $L : \text{diag}(\boldsymbol{\ell})^{-1} \in \mathbb{R}^{d\times d}$ built upon the kernel lengthscales $\boldsymbol{\ell}$ represents the Automatic Relevance Detection (ARD) (Rasmussen, 2003).

Spectral Representation of a Gaussian Process With this formulation, one can subsequently obtain an asymptotic approximation of the GP posterior in weight-space as a Bayesian linear model (Hernández-Lobato et al., 2014; Wilson et al., 2020), also known as the *spectral representation* of GPs (Hensman et al., 2017):

$$\mathcal{GP}(f) \dot{\sim} \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{\theta} \tag{3}$$

where $\dot{\sim}$ represents an approximation of distributions. The posterior of the weights $\boldsymbol{\theta} \in \mathbb{R}^{N_f}$, given prior $\mathcal{N}(\mathbf{0}, I)$ and N_D observations $D = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i\leq N_D}$, where $\boldsymbol{y}_i \sim \mathcal{N}(f(\boldsymbol{x}_i), \sigma_n^2)$, is multivariate normal[4] :

$$\boldsymbol{\theta} \sim \mathcal{N}(\Phi^T\Phi + \sigma_n^2 I)^{-1}\Phi^T\boldsymbol{y}, (\Phi^T\Phi + \sigma_n^2 I)^{-1}\sigma_n^2) \tag{4}$$

where $\Phi := [\boldsymbol{\phi}(\boldsymbol{x}_1), \boldsymbol{\phi}(\boldsymbol{x}_2), ..., \boldsymbol{\phi}(\boldsymbol{x}_{N_D})]$ represent stacked feature function values for all training inputs, and σ_n^2 represents the noise likelihood variance. The spectral representation of a GP has been extensively investigated (Lázaro-Gredilla et al., 2010; Hensman et al., 2017) as it has mitigated the $\mathcal{O}(N_D^3)$ complexity. Moreover, there is another advantage which enjoys growing interest in BO (Hernández-Lobato et al., 2014; Mutnỳ & Krause, 2019; Suzuki et al., 2020; Fröhlich et al., 2020). It provides a deterministic approach to sample GP posterior trajectories, which is useful for inference of interesting quantities.

2.2. Robust Bayesian Optimization

Problem Formulation Let $f : \mathcal{X} \to \mathbb{R}$ be a time-consuming black-box function, we can observe $y = f(\boldsymbol{x}) + \eta$, where η is independent Gaussian observation noise with variance σ_n^2. Assuming $\boldsymbol{\xi}$ is additive noise generated from a known distribution $p(\boldsymbol{\xi})$ specified by the practitioner, the goal of RBO is to maximize $\mathbb{J}_{\boldsymbol{\xi}}(f)$ and minimize $\mathbb{V}_{\boldsymbol{\xi}}(f)$ simultaneously in a bounded space $\mathcal{X} \subset \mathbb{R}^d$. The most common formulations include:

Multi-Objective Formulation Multi-objective optimization is a common formulation, for instance, in finances (Iwazaki et al., 2021) and engineering (Tang & Périaux, 2012). Under this formulation, RBO is performed by treating the mean and the variance as a bi-objective optimization

[4] In case of $N_f > N_D$, more efficient computation can be realized through the Woodbury matrix identity (see e.g., Rasmussen (2003)) or pathwise update strategy (Wilson et al., 2020).

problem,

$$\underset{\boldsymbol{x}\in\mathcal{X}}{\text{maximize}}\ (\mathbb{J}(\boldsymbol{x}), -\mathbb{V}(\boldsymbol{x})) \tag{5}$$

where one seeks a *Pareto frontier* $\mathcal{F}^*_{MV} := \{\boldsymbol{f}_{MV} \in \mathbb{F}_{MV} | \{\boldsymbol{f}'_{MV} \in \mathbb{F}_{MV} | \boldsymbol{f}'_{MV} \succ \boldsymbol{f}_{MV}\} = \varnothing\}$, where $\mathbb{F}_{MV} := \{\boldsymbol{f}_{MV} \in \mathbb{R}^2 | \boldsymbol{f}_{MV} = [\mathbb{J}(\boldsymbol{x}), -\mathbb{V}(\boldsymbol{x})], \boldsymbol{x} \in \mathcal{X}\}$, $\succ$ represents the standard dominance operator (see, e.g., Section 1 of Daulton et al. (2020)).

Variance as Constraint Another scenario is performing the optimization, while make sure that the deviation of the final optimal solution with respect to input uncertainty is acceptable. This can be formulated as:

$$\begin{aligned} &\underset{x\in\mathcal{X}}{\text{maximize}}\ \mathbb{J}(\boldsymbol{x}) \\ &\text{s.t. } \mathbb{V}(\boldsymbol{x}) \le C_v \end{aligned} \tag{6}$$

where $C_v \in \mathbb{R}$ is a pre-specified constraint threshold, the goal is to locate the feasible maximum $\mathbb{J}^*_{\text{Fea}}$.

Scalarization Formulation Scalarization (also known as *Multi-Task* (Iwazaki et al., 2021)) aggregates multiple objectives into a single cost function. For instance, a straightforward aggregation is linear scalarization (e.g., Paria et al. (2020)):

$$\psi_{MV} := \underset{x\in\mathcal{X}}{\text{maximize}}\ \mathbb{J}(\boldsymbol{x}) - \alpha\mathbb{V}(\boldsymbol{x}) \tag{7}$$

where $\alpha \ge 0$ is the pre-specified scalarization coefficient.

Finally, we like to point out the interchangeability between the above formulations through the following remark:

Remark 2.1. Given $f, \boldsymbol{\xi}$'s distribution and their corresponding $\mathcal{F}^*_{MV}$:

1. $\forall \boldsymbol{f}_{MV} \in \mathcal{F}^*_{MV}, \exists C'_v \ge 0$ s.t. $\mathbb{J}^*_{\text{Fea}} | C'_v = \boldsymbol{f}_{MV}$;
2. If $\mathcal{F}^*_{MV}$ is concave: $\forall \boldsymbol{f}_{MV} \in \mathcal{F}^*_{MV}, \exists \alpha' \ge 0$ s.t. $\psi_{MV} | \alpha' = \boldsymbol{f}_{MV}$.

3. Spectral Representation of Robustness Measures

Motivation The calculation of robustness measures in Eq. 1 expects a continuous f to be handled appropriately by the expectation operation. Nevertheless, the standard GP predictive posterior depends on a finite number of discrete test input which makes the analytical calculation of the expectation non-trivial. Hence, we propose to degenerate the GP to its spectral representation of Eq. 3. This sheds light on an analytical expression of Eq. 1.

Definition 3.2. (Spectral representation of robustness measure) With the Fourier feature based approximation of the kernel, we are able to reach the following Bayesian robustness measures approximation $\tilde{\mathbb{J}}(\mathcal{GP}(f)) | \boldsymbol{\phi}, \boldsymbol{\theta}$ and $\tilde{\mathbb{V}}(\mathcal{GP}(f)) | \boldsymbol{\phi}, \boldsymbol{\theta}$:

$$\begin{aligned} \mathbb{J}(\mathcal{GP}(f)) &\dot{\sim} \tilde{\mathbb{J}}(\mathcal{GP}(f)) | \boldsymbol{\phi}, \boldsymbol{\theta} = \mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})^T]\boldsymbol{\theta} \\ \mathbb{V}(\mathcal{GP}(f)) &\dot{\sim} \tilde{\mathbb{V}}(\mathcal{GP}(f)) | \boldsymbol{\phi}, \boldsymbol{\theta} \\ &= \boldsymbol{\theta}^T \mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})^T]\boldsymbol{\theta} \\ &\quad - \left[\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})^T]\boldsymbol{\theta}\right]^2 \end{aligned} \tag{8}$$

It can be seen that under finite parameterization, the mean expression is a normal distribution. The variance conditional distribution is the difference of two correlated generalized χ^2 distributions.

With these spectral representations. we are also able to provide their first moments through the following theorem:

Theorem 3.3. *Given* $\mathcal{GP}(f) \dot{\sim} \boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{\theta}$, *where* $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_\theta, \Sigma_{\boldsymbol{\theta}})$, *the robustness measure's first moments are:*

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\theta}}\left[\tilde{\mathbb{J}}(\mathcal{GP}(f))\right] &= \mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})^T]\boldsymbol{\mu}_{\boldsymbol{\theta}} \\ \mathbb{E}_{\boldsymbol{\theta}}\left[\tilde{\mathbb{V}}(\mathcal{GP}(f))\right] &= \boldsymbol{\mu}_{\boldsymbol{\theta}}^T \mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})^T]\boldsymbol{\mu}_{\boldsymbol{\theta}} \\ &\quad - \left[\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})^T]\boldsymbol{\mu}_{\boldsymbol{\theta}}\right]^2 \\ &\quad + \operatorname{tr}\left(\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})^T]\Sigma_{\boldsymbol{\theta}}\right) \\ &\quad - \operatorname{tr}\left(\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})^T]\Sigma_{\boldsymbol{\theta}}\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})]\right) \end{aligned} \tag{9}$$

Theorem 3.3 provides the first moments of the parametric robustness measures[5]. It also indicates why the degenerate approximation of the variance on the GP posterior mean is not unbiased with regard to its first moment, which we further illustrate through the following proposition.

Proposition 3.4. *Denoting the* $\mathcal{GP}$ *posterior mean as* $\mu(f)$, *define* $\mathbb{E}_{\boldsymbol{\theta}}[\mathbb{V}(\mathcal{GP}(f))] = \lim_{N_f\to\infty} \tilde{\mathbb{V}}(\mathcal{GP}(f))$ *to represent the variance's mean under the* $\mathcal{GP}$, *one has* $\mathbb{E}_{\boldsymbol{\theta}}[\mathbb{V}(\mathcal{GP}(f))] > \mathbb{V}(\mu(f))$.

This states that the degenerate variance approximation is an underestimation compared to the variance's first moment under a GP. In fact, this holds for arbitrary Bayesian linear models with a normal prior on $\boldsymbol{\theta}$ (see appendix B.1).

In the following, we discuss two different approaches for constructing the Fourier feature mapping $\boldsymbol{\phi}$, resulting in two different spectral representing forms. Each with its advantages and drawbacks.

Table 1. Random Fourier feature based robustness measures

$\boldsymbol{\xi}$ distribution	$\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})_m^T]$ (Mean)	$\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})_m \boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})_n^T]$
$\mathcal{N}(\mathbf{0}, \Sigma)$	$\sqrt{\frac{2\sigma^2}{N_f}} \cos[\omega_m^T L\boldsymbol{x} + b_m] \cdot e^{-\frac{1}{2}\boldsymbol{\omega}_m^T L^T \Sigma L \boldsymbol{\omega}_m}$	$\sum_{\bigcirc \in \{+,-\}} \left[\frac{\sigma^2}{N_f} \cos[(\boldsymbol{\omega}_m^T \bigcirc \boldsymbol{\omega}_n^T) L\boldsymbol{x} + (b_m \bigcirc b_n)] \cdot e^{-\frac{1}{2}(\boldsymbol{\omega}_m^T \bigcirc \boldsymbol{\omega}_n^T) L^T \Sigma L (\boldsymbol{\omega}_m \bigcirc \boldsymbol{\omega}_m)} \right]$
$U(-\boldsymbol{\delta}, \boldsymbol{\delta})$	$\sqrt{\frac{2\sigma^2}{N_f}} \cos(\boldsymbol{\omega}_m^T L\boldsymbol{x} + b_m) \cdot \frac{\sin \boldsymbol{\omega}_m^T L\boldsymbol{\delta}}{\boldsymbol{\omega}_m^T L\boldsymbol{\delta}}$	$\sum_{\bigcirc \in \{+,-\}} \left[\frac{\sigma^2}{N_f} \cos\left[(\boldsymbol{\omega}_m^T \bigcirc \boldsymbol{\omega}_n^T) L\boldsymbol{x} + (b_m \bigcirc b_n)\right] \cdot \frac{\sin\left((\omega_m^T \bigcirc \omega_n^T) L\boldsymbol{\delta}\right)}{(\omega_m^T \bigcirc \omega_n^T) L\boldsymbol{\delta}} \right]$

3.1. Fourier Feature based Robustness Measures

3.1.1. Random Fourier Features

In order to generate the feature mapping function, a common approach is to draw the Fourier features randomly, known as Random Fourier Features (RFF) (Rahimi et al., 2007). Consider the following inner product form:

$$\begin{aligned} & k(\boldsymbol{x}, \boldsymbol{x}') \\ & = \sigma^2 \mathbb{E}_{\boldsymbol{\omega}, b}\left[\cos(\boldsymbol{\omega}^T L\boldsymbol{x} + b) \cos(\boldsymbol{\omega}^T L\boldsymbol{x}' + b)\right] \\ & \approx \boldsymbol{\phi}^T(\boldsymbol{x}) \boldsymbol{\phi}(\boldsymbol{x}') \end{aligned} \tag{10}$$

For the mth feature function, we can decompose $\boldsymbol{x}$ and $\boldsymbol{x}'$ into an inner product form as $\phi(\boldsymbol{x})_m = \sqrt{\frac{2\sigma^2}{N_f}} \cos(\boldsymbol{\omega}_m^T L\boldsymbol{x}) + b_m) \in \mathbb{R}$ using the sum of angle formula, where $\boldsymbol{x} \in \mathbb{R}^d$, σ^2 represents the kernel variance, $\boldsymbol{\omega}_m \in \mathbb{R}^d \sim p(\boldsymbol{\omega})$, $b_m \sim U(0, 2\pi)$.

Definition 3.5. (RFF-Mean, RFF-Variance) Given the RFF based kernel approximation using the decomposition of Eq. 10, the spectral robustness measures, using the expressions in Table 1, are called RFF-Mean and RFF-Variance.

RFF-Mean and RFF-Variance share the advantage of flexibility regarding the specification of their Fourier feature number N_f. Unfortunately, RFF is known to have *variance starvation* (Wang et al., 2018; Mutnỳ & Krause, 2019; Wilson et al., 2020): its kernel approximation can lead to inaccurate GP posterior approximations if the training data size is huge. This inaccuracy will affect the robustness measures' posterior as well (see RFF based robustness measure trajectory samples using algorithm 1 in Fig. 1). To mitigate this potential issue, we can utilize quasi-Monte Carlo to sample from common kernels' spectral densities (Yang et al., 2014). In order to provide a more accurate inference of the robustness measures, we introduce a numerical quadrature-based Fourier features constructing technique and demonstrate how we can build robustness measures on top of it.

[5]The second (central) moment is also available in appendix A.2.

3.1.2. Quadrature Fourier Features

In order to mitigate the variance starvation problem, Mutnỳ & Krause (2019) propose to make use of a numerical quadrature technique. For stationary kernels with decomposable spectral density, one can expect a much more accurate approximation of Eq. 2 with the same N_f, by utilizing numerical quadrature based Fourier features. This can be achieved by an alternative decomposition of the kernel (Rahimi et al., 2007) and can be written as:

$$\begin{aligned} & k(\boldsymbol{x}, \boldsymbol{x}') \\ & = \sigma^2 \mathbb{E}_{\boldsymbol{\omega}}\left[\begin{bmatrix} \cos(\boldsymbol{\omega}^T L\boldsymbol{x}) \\ \sin(\boldsymbol{\omega}^T L\boldsymbol{x}) \end{bmatrix}^T \begin{bmatrix} \cos(\boldsymbol{\omega}^T L\boldsymbol{x}') \\ \sin(\boldsymbol{\omega}^T L\boldsymbol{x}') \end{bmatrix}\right] \\ & \approx \boldsymbol{\phi}^T(\boldsymbol{x}) \boldsymbol{\phi}(\boldsymbol{x}') \end{aligned} \tag{11}$$

For the common Squared Exponential (SE) kernel, its normalized spectral density is decomposable as product of independent normal factors. Hence, the multidimensional integration in Eq. 11 can be approximated using a multivariate Gauss-Hermit quadrature, which results in the feature mapping expression $\boldsymbol{\phi} \in \mathbb{R}^{2N_f}$ represented as:

$$\begin{aligned} & \boldsymbol{\phi}_j(\boldsymbol{x}) \\ & = \begin{cases} \sqrt{\sigma^2 v(\boldsymbol{\omega}_j)} \cos(\boldsymbol{\omega}_j^T L\boldsymbol{x}) & j \le N_f, \\ \sqrt{\sigma^2 v(\boldsymbol{\omega}_{j-N_f})} \sin(\boldsymbol{\omega}_{j-N_f}^T L\boldsymbol{x}) & N_f < j \le 2N_f \end{cases} \end{aligned} \tag{12}$$

where $v(\boldsymbol{\omega}_j), v(\boldsymbol{\omega}_{j-N_f})$ represents the quadrature weights on integration node $\boldsymbol{\omega}_j$, $\boldsymbol{\omega}_{j-N_f}$ respectively, which has been generated on a Cartesian product grid: assuming $N_f = \overline{m}^d$, $\overline{m} \in \mathbb{N}$, defining the set $\zeta = \{v(\omega_1), ..., v(\omega_{\overline{m}})\}$ representing one-dimensional quadrature weights, we have $v(\boldsymbol{\omega}_j) \in \zeta^d$, where ζ^d represents d times Cartesian product over ζ. We refer to (Mutnỳ & Krause, 2019) for a more thorough discussion and show our robustness measure's expression based on QFF below.

Definition 3.6. (QFF-Mean, QFF-Variance) Substituting Eq. 12 into Eq. 8, the mean and the variance robustness measures using the expressions in Table 2 (in the appendix D.1) are defined as QFF-Mean and QFF-Variance.

We demonstrate the QFF-Mean and Variance based sample trajectories using algorithm 1 in Fig. 1. It can be seen that, with this kind of technique, the variance starvation can be effectively suppressed and hence provide an accurate approximation of uncertainty, even with densely sampled training data.

Algorithm 1 Sampling Robustness Measure's Trajectories

1: **Input:** Stacked feature function Φ, training data D, sample size N.
2: Calculate $\boldsymbol{\theta}$'s posterior distribution parameter: $\boldsymbol{\mu}_{\boldsymbol{\theta}}, \Sigma_{\boldsymbol{\theta}}$ (Eq. 4).
3: **for** $i = 1$ **to** N **do**
4: $\quad \boldsymbol{\theta}_i \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \Sigma_{\boldsymbol{\theta}})$.
5: $\quad$ Construct $\tilde{\mathbb{J}}_i\left(\mathcal{GP}(f)\right)|\boldsymbol{\phi}, \boldsymbol{\theta}_i$ trajectory sample. (RFF: Eq. 1 use Table 1, QFF: Eq. 1 use Table 2)
6: $\quad$ Construct $\tilde{\mathbb{V}}_i\left(\mathcal{GP}(f)\right)|\boldsymbol{\phi}, \boldsymbol{\theta}_i$ trajectory sample. (RFF: Eq. 1 use Table 1, QFF: Eq. 1 use Table 2)
7: **end for**

We remark the main limitations of QFF based robustness measures are twofold: 1) the quadrature utilized in (Mutnỳ & Krause, 2019) depends on the decomposable assumption of the kernel's spectral density. 2) the Fourier feature number N_f grows exponentially with the input dimensionality due to the Cartesian form of integration nodes, hence restricting the application of this strategy to small to medium scale (see appendix F for a discussion of computational complexity).

We summarize that a fully Bayesian posterior can be obtained by marginalizing out the kernel hyperparameters, which is the same as for the GPs (Simpson et al., 2020; Lalchand & Rasmussen, 2020; Filippone & Girolami, 2014). Meanwhile, we highlight one advantage of this spectral robustness measure representation is its utilization in posterior sampling, as deterministic posterior trajectory samples are available through sampling the weights $\boldsymbol{\theta}$. This property is readily useful in RBO through different forms of acquisition functions, we will discuss several of the main forms in the next section.

4. Application in Robust Bayesian Optimization

We elaborate on how the Bayesian inference of the robustness measures can be utilized in different RBO formulations. More specifically, we provide modified myopic acquisition functions (Daulton et al., 2021; Letham et al., 2019; Jones et al., 1998) to take the robustness measures into account.

Mean-Variance based Multi-Objective Approach We provide Fourier Feature based, Mean Variance considered, Expected HyperVolume Improvement acquisition function (FF-MV-EHVI) to search for the optimal Pareto front $\mathcal{F}^*_{MV}$.

$$\begin{aligned}
&\alpha_{\text{FF-MV-EHVI}} \\
&= \int_{\boldsymbol{F}_{MV}} \int_{\mathbb{J},\mathbb{V}} \text{HVI}(\mathbb{J}, \mathbb{V}, \mathcal{F}^*_{MV}) p(\mathbb{J}, \mathbb{V}|\boldsymbol{F}_{MV}) \\
&\quad p(\boldsymbol{F}_{MV}|\boldsymbol{X}) d\mathbb{J} d\mathbb{V} d\boldsymbol{F}_{MV} \\
&\approx \frac{1}{N} \sum_{i=1}^{N} \text{HVI}(\tilde{\mathbb{J}}_i, \tilde{\mathbb{V}}_i, \tilde{\mathcal{F}}^*_{MV_i} | \boldsymbol{\theta}_i, D)
\end{aligned} \tag{13}$$

where $\boldsymbol{X} : \{\boldsymbol{x}_i\}_{i \leq N_D}$ represents the training data's input, $\boldsymbol{F}_{MV}$ represents the corresponding output in $\mathbb{F}_{MV}$, HVI represents the hypervolume improvement (see e.g., Definition 2 of Daulton et al. (2020)) based on the current Pareto frontier $\mathcal{F}^*_{MV}$. Since this quantity is not directly observed in our scenario, we propose using a Monte Carlo approach to approximately integrate $\mathcal{F}^*_{MV}$ similar as Daulton et al. (2021), this can be easily achieved within the Fourier feature based robustness measures, as jointly sample $\tilde{\mathbb{J}}_i, \tilde{\mathbb{V}}_i, \tilde{\mathcal{F}}^*_{MV_i}$ can be achieved by simply sampling $\boldsymbol{\theta}$[6].

Variance as Constraint Assuming the objective is to perform RBO by considering the variance to be lower than a specified threshold C_v. Based on (Letham et al., 2019), we propose the Fourier Feature based, Mean Variance considered, Expected Constrained Improvement acquisition function (FF-MV-ECI):

$$\begin{aligned}
&\alpha_{\text{FF-MV-ECI}} \\
&= \int_{\mathbb{J}^*_{Fea}} \int_{\mathbb{J},\mathbb{V}} l(\mathbb{J}, \mathbb{J}^*_{Fea}) \mathbb{I}^-(\mathbb{V} - C_v) \\
&\quad p(\mathbb{J}, \mathbb{V}|\mathbb{J}^*_{Fea}) p(\mathbb{J}^*_{Fea}|D) d\mathbb{J} d\mathbb{V} d\mathbb{J}^*_{Fea} \\
&\approx \frac{1}{N} \sum_{i=1}^{N} \left[l(\tilde{\mathbb{J}}_i, \tilde{\mathbb{J}}^*_{Fea_i} | \boldsymbol{\theta}_i, D) \sigma_s \left(\frac{C_v - \tilde{\mathbb{V}}_i | \boldsymbol{\theta}_i}{\tau} \right) \right]
\end{aligned} \tag{14}$$

[6] One can also use a vanilla approach (Gramacy & Lee, 2010) to simplify the problem by utilizing the model posterior mean to infer the current Pareto frontier $\overline{\mathcal{F}^*_{MV}}$ based on evaluated candidates. We note this is also applicable for the other two acquisition functions.

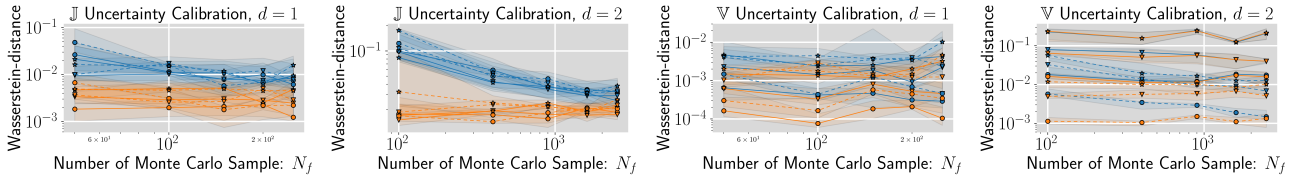

Figure 2. Uncertainty calibration comparison between RFF and QFF based robustness measures. Dashed lines represents normal input uncertainty: $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$ and solid lines represents uniform: $\boldsymbol{\xi} \sim U(-\mathbf{1}\delta, \mathbf{1}\delta)$. We use {orange , blue} to represent {QFF , RFF}, respectively. We use $\{\bigcirc, \bigtriangledown, \bigstar\}$ to represent $\sigma^2 = \{0.001, 0.005, 0.01\}$ for normal and $\delta = \{0.05, 0.1, 0.2\}$ for uniform uncertainty respectively.

The utility function l is defined as : $l(\tilde{\mathbb{J}}_i, \tilde{\mathbb{J}}^*_{Fea_i} | \boldsymbol{\theta}_i) := \max(\tilde{\mathbb{J}}_i - \tilde{\mathbb{J}}^*_{Fea_i}, 0)$. For the constraint handling, we relax the original indicator function $\mathbb{I}^-$ to a sigmoid function $\sigma_s(\cdot)$ with a small positive temperature parameter τ to ensure differentiability (Wilson et al., 2018; Maddison et al., 2016).

Scalarization based Approach Finally, the Fourier Feature based, Mean Variance considered, Expected Improvement acquisition function (FF-MV-EI) is:

$$\begin{aligned} &\alpha_{\text{FF-MV-EI}} \\ &= \int_{\psi^*} \int_{\mathbb{J},\mathbb{V}} l(\mathbb{J} - \alpha\mathbb{V}, \psi^*) p(\mathbb{J}, \mathbb{V} | \psi^*) p(\psi^* | D) d\psi^* d\mathbb{J} d\mathbb{V} \\ &\approx \frac{1}{N} \sum_{i=1}^{N} l(\tilde{\mathbb{J}}_i - \alpha\tilde{\mathbb{V}}_i, \tilde{\psi}^*_i | \boldsymbol{\theta}_i, D) \end{aligned} \tag{15}$$

where $\tilde{\psi}^*_i$ is extracted from the training data, $l(\tilde{\mathbb{J}} - \alpha\tilde{\mathbb{V}}, \tilde{\psi}^*, \boldsymbol{\theta}_i) := \max(\tilde{\mathbb{J}} - \alpha\tilde{\mathbb{V}} - \tilde{\psi}^*, 0)$.

Other Extensions We restricted ourselves to three examples using the Bayesian robustness measures. Nevertheless, the strength of this work is that one could easily extend it to any acquisition function like Thompson sampling (Bradford et al., 2018) or non-myopic acquisition functions. We discuss in appendix G that a batch extension of the above acquisition functions can be easily achieved.

5. Experiments

Benchmarking of Uncertainty Calibration

We conduct multiple experiments to investigate the accuracy of the spectral representation of the robustness measures[7]. We also benchmark their performance for RBO, using the three proposed acquisition functions. The code is implemented using the open-source library *Trieste* (Berkeley et al., 2021)[8].

We first present an empirical uncertainty calibration study to investigate the inference accuracy of the RFF and QFF based robustness measures for guiding real-life applications. A data set of $10d$ samples is drawn from a GP prior based on a SE kernel. Afterwards, a GP is constructed and we compare the robustness measure distributions (based on RFF and QFF) at different input locations with an exhaustive Monte Carlo approach, which is regarded as the ground-truth. The difference between distributions is measured using the 1-Wasserstein distance (Ramdas et al., 2017), and we report the average of the 1-Wasserstein distance across eight runs in Fig. 2. As expected, given the same Fourier Feature numbers, QFF provides more accurate distributions.

Robust Bayesian Optimization The proposed FF-MV-EHVI, FF-MV-ECI and FF-SMV-EI acquisition functions are benchmarked on different synthetic functions and real-life problems. For each problem, $5d$ initial randomly generated data are used. For multi-objective formulation, we compare against MO-MVA-BO (Iwazaki et al., 2021). For the variance as constraint formulation, we compare against Co-MVA-BO (Iwazaki et al., 2021). Finally, in the scalarization approach we benchmark against MT-MBA-BO (Iwazaki et al., 2021)[9]. These acquisition functions are known demonstrating the state-of-the-art performance in active learning. Besides, we also compare with Uncertainty Sampling (US) (Iwazaki et al., 2021) and random search for all three formulations. We adopt the log-hypervolume difference (Daulton et al., 2020), utility gap (Hernández-Lobato et al., 2016) and simple regret for multi-objective, constraint and scalarization formulation, respectively, to indicate the performance. We conduct the optimal solutions recommendation based on model inference. For the comparing methods, we use their corresponding optimal solution extracting strategies. For the GP constructing, the SE ARD kernel is used with a log-normal prior on lengthscales, where the kernel hyperparameter is inferred using maximum a posteriori estimation. Each experiment is repeated 30 times.

[7] We provide an additional benchmark comparing the point estimation's (cfr. Theorem 3.3) accuracy with direct Monte Carlo sample $\boldsymbol{\xi}$ in $\mathcal{X}$ in appendix H.1.

[8] Our code is available at https://github.com/TsingQAQ/gp_mean_var_rbo.

[9] Since Iwazaki et al. (2021) use the standard deviation for the problem formulations, we reformulate it to be variance for the same objective setting, see appendix H.2 for details.

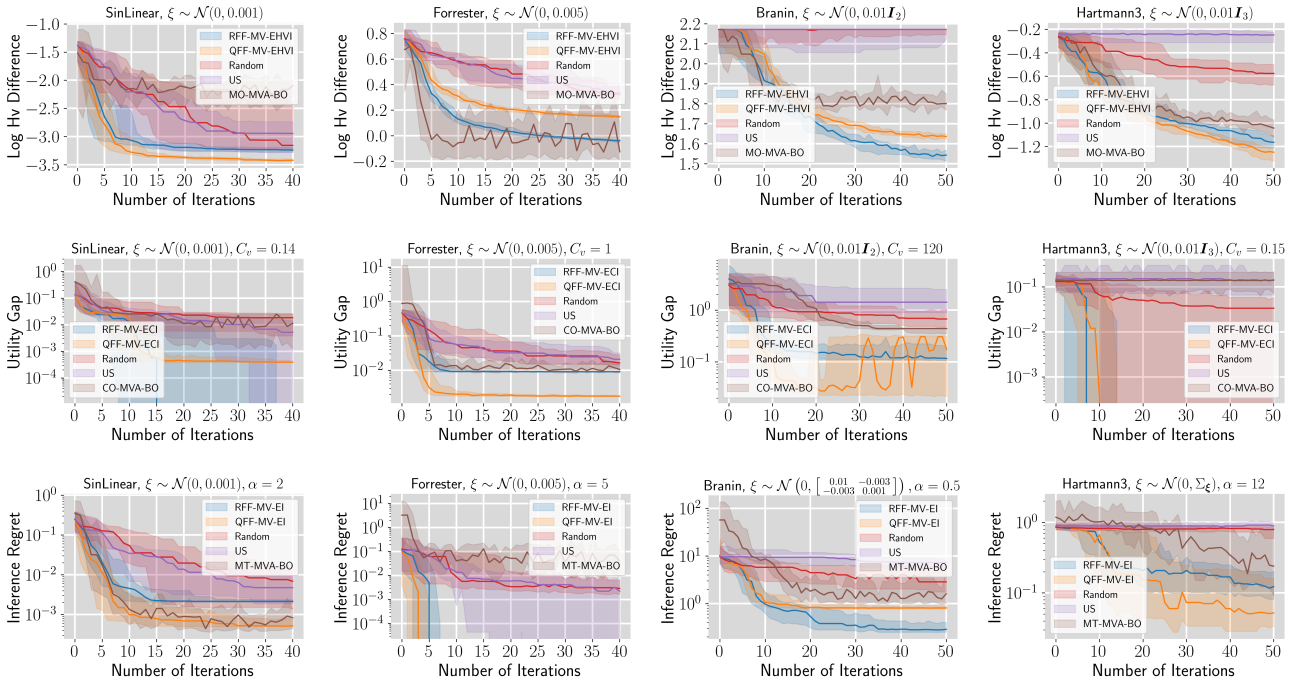

Figure 3. Numerical experiment results on synthetic functions shown through median performances and interquartile ranges, where $\Sigma_{\boldsymbol{\xi}} = \begin{bmatrix} 0.01 & 0.009 & -0.009 \\ 0.009 & 0.01 & -0.008 \\ -0.009 & -0.008 & 0.01 \end{bmatrix}$. Each row represents one formulation of the objective functions.

Synthetic Experiments We conduct RBO on common synthetic benchmark functions and report the normal input uncertainty results in Fig. 3, as well as the uniform input uncertainty results in appendix H.5. It can be seen that both QFF and RFF based acquisition functions demonstrate competitive performance. Intriguingly, while the QFF robustness measure is expected to provide a more accurate robustness measure, the RFF based approach also seems to have competitive performance on most problems.

Conceptual Low Drag Wing Design

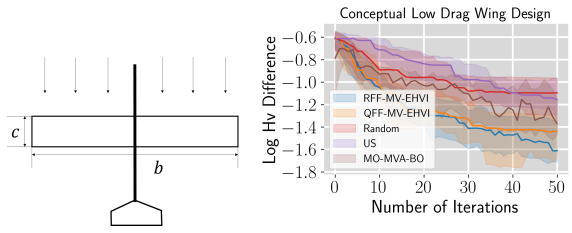

Figure 4. Conceptual Aircraft Low Drag Wing Design

We perform RBO on a two-dimensional conceptual design problem of an aircraft's rectangular wing in level flight condition. The goal is to find the optimal wing geometry, i.e., the wing span b and chord length c (see Fig. 4), to achieve a realistic wing geometry with high aerodynamic efficiency.

In practice, this geometric parameters can easily get shifted not only because of fabrication errors, but also the deformation due to aerodynamic force during flight. Hence, in order to have a wing configuration that endows high aerodynamic efficiency even under slight geometry change, we consider a multi-objective formulation on this problem in the input space: $(b, c) \in [1, 3] \times [0.05, 1]$ under the input uncertainty of the form $(b, c) \sim \mathcal{N}(0, \left[\begin{smallmatrix} 0.004 & 0.0 \\ 0.0 & 0.00045125 \end{smallmatrix}\right])$. The objective function definition is provided in appendix H.4. we remark that the multi-objective formulation provides a lot of flexibility to the practitioner, which is especially useful in engineering design as the designer is able to choose among the Pareto optimal solutions according to his own preference. The log-Hypervolume convergence curve is shown in Fig. 4, both QFF and RFF based MV-EHVI demonstrate a faster convergence speed.

Pre-Image Learning in Robot Pushing

Lastly, a deterministic version of pre-image learning for a robot pushing task (Kaelbling & Lozano-Pérez, 2017) is used in a RBO setting[10]. The goal of the robot pushing prob-

[10]The objective function is available at https://github.com/zi-w/Max-value-Entropy-Search.

lem is to find a good pre-image to stop at a target location (Wang & Jegelka, 2017). The three-dimensional problem takes the initial object's position $(r_{\boldsymbol{x}_1}, r_{\boldsymbol{x}_2}) \in [-5, 5]^2$, as well as the pushing time $r_t \in [1, 30]$ as input, and the objective function is the Euclidean distance between the object stop location and the target location $(g_{\boldsymbol{x}_1}, g_{\boldsymbol{x}_2})$. A normal input noise $\boldsymbol{\xi} \sim \mathcal{N}(0, \mathrm{diag}(0.004, 0.004, 0.002))$ is specified on the input parameters, and the goal location is set as $(g_{\boldsymbol{x}_1}, g_{\boldsymbol{x}_2}) = [4, 3]$.

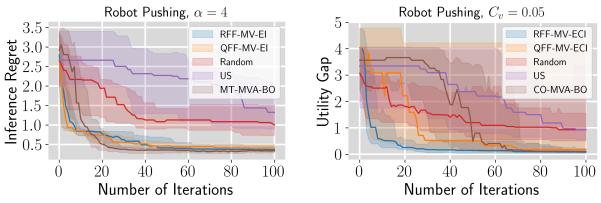

Figure 5. Results of robot pushing problem using scalarization and variance as constraint formulations.

We consider both variance as constraint and scalarization formulations for this problem, The results are shown in Fig. 5. It can be seen that for both formulations, our methods tend to converge well in practice. For the scalarization formulation, MT-MVA-BO demonstrates a faster convergence speed after 10 iterations. For the variance as constraint formulation, both QFF and RFF based EI demonstrate clear advantages.

6. Conclusions

A spectral approximation of two robustness measures is introduced based on the GP's weight-space approximation, in which we provide two versions of the robustness measures using different kernel approximations. We compare these two formulations on uncertainty calibration. We also elaborate on how these robustness measures can be utilized in robust Bayesian Optimization by extending three well-known acquisition functions, and we demonstrate their competitive performance through synthetic problems and real-life applications.

Limitations and Future work Since the spectral representation is based on kernel approximation through the finite number of feature mapping function, a larger feature numbers are expected in higher dimensions for an accurate approximation, hence our robustness measures. As a unfavorable result, practitioners have to make a trade off between approximation accuracy, computation time and memory footprint. Future work will be focused on two aspects: (i) provide a more accurate and efficient approximation of the robustness measures, which is crucial for scaling to a higher number of input dimensions, and (ii) different acquisition functions and robustness measures shall be considered, tailored to specific robust Bayesian Optimization problems and formulations.

Acknowledgement

This research receives funding from the Flemish Government under the "Onderzoeksprogramma Artifciële Intelligentie (AI) Vlaanderen" programme and the "Fonds Wetenschappelijk Onderzoek (FWO)" programme, and Chinese Scholarship Council (grant number 201906290032). We sincerely thank Xin Wang and Bin Chen for useful discussions. We also gratefully thank our colleague Nasrulloh Loka and anonymous reviewers for providing extensive comments for improving the paper.

References

Balandat, M., Karrer, B., Jiang, D. R., Daulton, S., Letham, B., Wilson, A. G., and Bakshy, E. Botorch: A framework for efficient monte-carlo bayesian optimization. *arXiv preprint arXiv:1910.06403*, 2019.

Beland, J. J. and Nair, P. B. Bayesian optimization under uncertainty. In *NIPS BayesOpt 2017 workshop*, 2017.

Berkeley, J., Moss, H. B., Artemev, A., Pascual-Diaz, S., Granta, U., Stojic, H., Couckuyt, I., Qing, J., Satrio, L., and Picheny, V. Trieste, 8 2021. URL https://github.com/secondmind-labs/trieste.

Beyer, H.-G. and Sendhoff, B. Robust optimization–a comprehensive survey. *Computer methods in applied mechanics and engineering*, 196(33-34):3190–3218, 2007.

Bogunovic, I. and Krause, A. Misspecified gaussian process bandit optimization. *Advances in Neural Information Processing Systems*, 34, 2021.

Bogunovic, I., Scarlett, J., Jegelka, S., and Cevher, V. Adversarially robust optimization with gaussian processes. *arXiv preprint arXiv:1810.10775*, 2018.

Bradford, E., Schweidtmann, A. M., and Lapkin, A. Efficient multiobjective optimization employing gaussian processes, spectral sampling and a genetic algorithm. *Journal of global optimization*, 71(2):407–438, 2018.

Cakmak, S., Astudillo, R., Frazier, P., and Zhou, E. Bayesian optimization of risk measures. *arXiv preprint arXiv:2007.05554*, 2020.

Cui, L.-j., Sun, M.-y., Cao, Y.-l., Zhao, Q.-j., Zeng, W.-h., and Guo, S.-r. A novel tolerance geometric method based on machine learning. *Journal of Intelligent Manufacturing*, 32(3):799–821, 2021.

Daulton, S., Balandat, M., and Bakshy, E. Differentiable expected hypervolume improvement for parallel multi-objective bayesian optimization. *arXiv preprint arXiv:2006.05078*, 2020.

Daulton, S., Balandat, M., and Bakshy, E. Parallel bayesian optimization of multiple noisy objectives with expected hypervolume improvement. *arXiv preprint arXiv:2105.08195*, 2021.

Filippone, M. and Girolami, M. Pseudo-marginal bayesian inference for gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(11):2214–2226, 2014.

Frazier, P. I. A tutorial on bayesian optimization. *arXiv preprint arXiv:1807.02811*, 2018.

Fröhlich, L. P., Klenske, E. D., Vinogradska, J., Daniel, C., and Zeilinger, M. N. Noisy-input entropy search for efficient robust bayesian optimization. *arXiv preprint arXiv:2002.02820*, 2020.

Girard, A., Rasmussen, C., Candela, J. Q., and Murray-Smith, R. Gaussian process priors with uncertain inputs application to multiple-step ahead time series forecasting. *Advances in neural information processing systems*, 15, 2002.

Gramacy, R. B. and Lee, H. K. H. Optimization under unknown constraints, 2010.

Graybill, F. A. *Matrices with applications in statistics*. Number 512.896 G7 1983. 1983.

Hensman, J., Durrande, N., Solin, A., et al. Variational fourier features for gaussian processes. *J. Mach. Learn. Res.*, 18(1):5537–5588, 2017.

Hernández-Lobato, J. M., Hoffman, M. W., and Ghahramani, Z. Predictive entropy search for efficient global optimization of black-box functions. *arXiv preprint arXiv:1406.2541*, 2014.

Hernández-Lobato, J. M., Gelbart, M. A., Adams, R. P., Hoffman, M. W., and Ghahramani, Z. A general framework for constrained bayesian optimization using information-based search. 2016.

Hoque, M. E. and Low, S.-W. Industry risk factors and stock returns of malaysian oil and gas industry: A new look with mean semi-variance asset pricing framework. *Mathematics*, 8(10):1732, 2020.

Iwazaki, S., Inatsu, Y., and Takeuchi, I. Mean-variance analysis in bayesian optimization under uncertainty. In *International Conference on Artificial Intelligence and Statistics*, pp. 973–981. PMLR, 2021.

Jones, D. R., Schonlau, M., and Welch, W. J. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.

Kaelbling, L. P. and Lozano-Pérez, T. Learning composable models of parameterized skills. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 886–893. IEEE, 2017.

Kapinchev, K., Bradu, A., Barnes, F., and Podoleanu, A. Gpu implementation of cross-correlation for image generation in real time. In *2015 9th International Conference on Signal Processing and Communication Systems (ICSPCS)*, pp. 1–6. IEEE, 2015.

Keane, A., Forrester, A., and Sobester, A. *Engineering design via surrogate modelling: a practical guide*. American Institute of Aeronautics and Astronautics, Inc., 2008.

Kirschner, J., Bogunovic, I., Jegelka, S., and Krause, A. Distributionally robust bayesian optimization. In *International Conference on Artificial Intelligence and Statistics*, pp. 2174–2184. PMLR, 2020.

Knudde, N., van der Herten, J., Dhaene, T., and Couckuyt, I. Gpflowopt: A bayesian optimization library using tensorflow. *arXiv preprint arXiv:1711.03845*, 2017.

Lalchand, V. and Rasmussen, C. E. Approximate inference for fully bayesian gaussian process regression. In *Symposium on Advances in Approximate Bayesian Inference*, pp. 1–12. PMLR, 2020.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Lázaro-Gredilla, M., Quinonero-Candela, J., Rasmussen, C. E., and Figueiras-Vidal, A. R. Sparse spectrum gaussian process regression. *The Journal of Machine Learning Research*, 11:1865–1881, 2010.

Letham, B., Karrer, B., Ottoni, G., Bakshy, E., et al. Constrained bayesian optimization with noisy experiments. *Bayesian Analysis*, 14(2):495–519, 2019.

Maddison, C. J., Mnih, A., and Teh, Y. W. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

Makarova, A., Usmanova, I., Bogunovic, I., and Krause, A. Risk-averse heteroscedastic bayesian optimization. *Advances in Neural Information Processing Systems*, 34, 2021.

Manfredi, P. and Trinchero, R. A probabilistic machine learning approach for the uncertainty quantification of electronic circuits based on gaussian process regression. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, 2021.

Markowitz, H. M. and Todd, G. P. *Mean-variance analysis in portfolio choice and capital markets*, volume 66. John Wiley & Sons, 2000.

McHutchon, A. and Rasmussen, C. E. Gaussian process training with input noise. In *Advances in Neural Information Processing Systems*, pp. 1341–1349, 2011.

Mutnỳ, M. and Krause, A. Efficient high dimensional bayesian optimization with additivity and quadrature fourier features. *Advances in Neural Information Processing Systems 31*, pp. 9005–9016, 2019.

Neiswanger, W. and Ramdas, A. Uncertainty quantification using martingales for misspecified gaussian processes. In *Algorithmic Learning Theory*, pp. 963–982. PMLR, 2021.

Nguyen, Q. P., Dai, Z., Low, B. K. H., and Jaillet, P. Optimizing conditional value-at-risk of black-box functions. *Advances in Neural Information Processing Systems*, 34, 2021a.

Nguyen, Q. P., Dai, Z., Low, B. K. H., and Jaillet, P. Value-at-risk optimization with gaussian processes. *arXiv preprint arXiv:2105.06126*, 2021b.

Nogueira, J., Martinez-Cantin, R., Bernardino, A., and Jamone, L. Unscented bayesian optimization for safe robot grasping. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1967–1972. IEEE, 2016.

O'Hagan, A. Bayes–hermite quadrature. *Journal of statistical planning and inference*, 29(3):245–260, 1991.

Oliveira, R., Ott, L., and Ramos, F. Bayesian optimisation under uncertain inputs. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1177–1184. PMLR, 2019.

O'Hagan, A. Uncertainty analysis: the variance of the variance. *Managing Uncertainty in Complex Models, Research Councils UK, Swindon, UK. http://www. mucm. ac. uk/Pages/Downloads/Technical% 20Reports/10-12% 20AOH% 20UAVarVar*, 200611, 2011.

Papoulis, A. and Pillai, S. *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill series in electrical engineering: Communications and signal processing. McGraw-Hill, 2002. ISBN 9780070486584. URL https://books.google.be/books?id=g6eUoWOlcQMC.

Paria, B., Kandasamy, K., and Póczos, B. A flexible framework for multi-objective bayesian optimization using random scalarizations. In *Uncertainty in Artificial Intelligence*, pp. 766–776. PMLR, 2020.

Rahimi, A., Recht, B., et al. Random features for large-scale kernel machines. In *NIPS*, volume 3, pp. 5. Citeseer, 2007.

Rahimian, H. and Mehrotra, S. Distributionally robust optimization: A review. *arXiv preprint arXiv:1908.05659*, 2019.

Ramdas, A., Trillos, N. G., and Cuturi, M. On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

Rasmussen, C. E. Gaussian processes in machine learning. In *Summer school on machine learning*, pp. 63–71. Springer, 2003.

Rivier, M. and Congedo, P. M. *Surrogate-Assisted Bounding-Box Approach Applied to Constrained Multi-Objective Optimisation Under Uncertainty*. PhD thesis, Inria Saclay Ile de France, 2018.

Schölkopf, B., Smola, A. J., Bach, F., et al. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

Seber, G. A. and Lee, A. J. *Linear regression analysis*, volume 329. John Wiley & Sons, 2012.

Sessa, P. G., Bogunovic, I., Kamgarpour, M., and Krause, A. Mixed strategies for robust optimization of unknown objectives. In *International Conference on Artificial Intelligence and Statistics*, pp. 2970–2980. PMLR, 2020.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and De Freitas, N. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104 (1):148–175, 2015.

Simpson, F., Lalchand, V., and Rasmussen, C. E. Marginalised gaussian processes with nested sampling. *arXiv preprint arXiv:2010.16344*, 2020.

Stein, M. L. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 1999.

Suzuki, S., Takeno, S., Tamura, T., Shitara, K., and Karasuyama, M. Multi-objective bayesian optimization using pareto-frontier entropy. In *International Conference on Machine Learning*, pp. 9279–9288. PMLR, 2020.

Tang, Z. and Périaux, J. Uncertainty based robust optimization method for drag minimization problems in aerodynamics. *Computer methods in applied mechanics and engineering*, 217:12–24, 2012.

Torossian, L., Picheny, V., and Durrande, N. Bayesian quantile and expectile optimisation. *arXiv preprint arXiv:2001.04833*, 2020.

Wang, H., Yuan, J., and Ng, S. H. Gaussian process based optimization algorithms with input uncertainty. *IISE Transactions*, 52(4):377–393, 2020a.

Wang, J., Clark, S. C., Liu, E., and Frazier, P. I. Parallel bayesian global optimization of expensive functions. *Operations Research*, 68(6):1850–1865, 2020b.

Wang, Z. and Jegelka, S. Max-value entropy search for efficient bayesian optimization. In *International Conference on Machine Learning (ICML)*, 2017.

Wang, Z., Gehring, C., Kohli, P., and Jegelka, S. Batched large-scale bayesian optimization in high-dimensional spaces. In *International Conference on Artificial Intelligence and Statistics*, pp. 745–754. PMLR, 2018.

Wauters, J., Couckuyt, I., Knudde, N., Dhaene, T., and Degroote, J. Multi-objective optimization of a wing fence on an unmanned aerial vehicle using surrogate-derived gradients. *Structural and Multidisciplinary Optimization*, 61(1):353–364, 2020.

Weichert, D. and Kister, A. Bayesian optimization for min max optimization. *arXiv preprint arXiv:2107.13772*, 2021.

Wilson, J., Borovitskiy, V., Terenin, A., Mostowsky, P., and Deisenroth, M. Efficiently sampling functions from gaussian process posteriors. In *International Conference on Machine Learning*, pp. 10292–10302. PMLR, 2020.

Wilson, J. T., Hutter, F., and Deisenroth, M. P. Maximizing acquisition functions for bayesian optimization. *arXiv preprint arXiv:1805.10196*, 2018.

Wu, X., Marshall, L., and Sharma, A. Quantifying input error in hydrologic modeling using the bayesian error analysis with reordering (bear) approach. *Journal of Hydrology*, 598:126202, 2021.

Yang, J., Sindhwani, V., Avron, H., and Mahoney, M. Quasi-monte carlo feature maps for shift-invariant kernels. In *International Conference on Machine Learning*, pp. 485–493. PMLR, 2014.

Appendix Contents

A. The Moments of the Spectral Robustness Measures

A.1. Deriving the First Moments

Lemma A.1. *Let* $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma) \in \mathbb{R}^d$*, let* $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ *is a symmetric matrix, one have* $\mathbb{E}\left[\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}\right] = \boldsymbol{\mu}^T\boldsymbol{A}\boldsymbol{\mu} + \mathrm{tr}(\boldsymbol{A}\Sigma)$

Proof see Theorem 1.5 of Seber & Lee (2012).

Now we begin the proof of Theorem 3.3:

Proof.

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\theta}}\Big[\boldsymbol{\theta}^T\mathbb{E}_{\boldsymbol{\xi}}[\phi(\boldsymbol{x}+\boldsymbol{\xi})\phi(\boldsymbol{x}+\boldsymbol{\xi})^T]\boldsymbol{\theta} - [\mathbb{E}_{\boldsymbol{\xi}}[\phi(\boldsymbol{x}+\boldsymbol{\xi})^T]\boldsymbol{\theta}]^2\Big]\\ &= \mathbb{E}_{\boldsymbol{\theta}}\Big[\boldsymbol{\theta}^T\mathbb{E}_{\boldsymbol{\xi}}[\phi(\boldsymbol{x}+\boldsymbol{\xi})\phi(\boldsymbol{x}+\boldsymbol{\xi})^T]\boldsymbol{\theta}\Big] - \mathbb{E}_{\boldsymbol{\theta}}\Big[\Big[\mathbb{E}_{\boldsymbol{\xi}}[\phi(\boldsymbol{x}+\boldsymbol{\xi})^T]\boldsymbol{\theta}\Big]^2\Big]\end{aligned} \tag{16}$$

The first part is a quadratic form of $\boldsymbol{\theta}$, with Lemma A.1 its expectation can hence be written as :

$$\mathbb{E}_{\boldsymbol{\theta}}\Big[\boldsymbol{\theta}^T\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})^T]\boldsymbol{\theta}\Big] = \mathrm{tr}(\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})^T]\Sigma_{\boldsymbol{\theta}}) + \boldsymbol{\mu}_\theta^T\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})^T]\boldsymbol{\mu}_\theta \tag{17}$$

For the second term, since $\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})^T]\boldsymbol{\theta} \sim \mathcal{N}\left(\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})^T]\boldsymbol{\mu}_{\boldsymbol{\theta}}, \mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})^T]\Sigma_{\boldsymbol{\theta}}\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})]\right)$, with the same lemma it is:

$$\mathbb{E}_{\boldsymbol{\theta}}\left[\left[\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})^T]\boldsymbol{\theta}\right]^2\right] = \left[\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})^T]\boldsymbol{\mu}_{\boldsymbol{\theta}}\right]^2 + \mathrm{tr}\left(\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})^T]\Sigma_{\boldsymbol{\theta}}\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})]\right) \tag{18}$$

Hence the proof complete. □

A.2. Second Moments

The second moments for both robustness measures are provided through the following theorem.

Theorem A.2. *Given $\mathcal{GP}(f) \approx \boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{\theta}$, where $\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \Sigma_{\boldsymbol{\theta}})$, the second moments of the robustness measures are:*

$$\mathbb{V}_{\boldsymbol{\theta}}[\tilde{\mathbb{J}}(f)] = \mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})^T]\Sigma_{\boldsymbol{\theta}}\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})] \tag{19}$$

$$\begin{aligned}\mathbb{V}_{\boldsymbol{\theta}}[\tilde{\mathbb{V}}(f)] &= \mathbb{V}_{\boldsymbol{\theta}}\left[\boldsymbol{\theta}^T\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})^T]\boldsymbol{\theta} - \boldsymbol{\theta}^T\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})]\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})^T]\boldsymbol{\theta}\right] \\ &= 2\,\mathrm{tr}[\Lambda_1\Sigma_{\boldsymbol{\theta}}\Lambda_1\Sigma_{\boldsymbol{\theta}}] + 4\boldsymbol{\mu}_{\boldsymbol{\theta}}^T\Lambda_1\Sigma_{\boldsymbol{\theta}}\Lambda_1\boldsymbol{\mu}_{\boldsymbol{\theta}} + 2\,\mathrm{tr}[\Lambda_2\Sigma_{\boldsymbol{\theta}}\Lambda_2\Sigma_{\boldsymbol{\theta}}] + 4\boldsymbol{\mu}_{\boldsymbol{\theta}}^T\Lambda_2\Sigma_{\boldsymbol{\theta}}\Lambda_2\boldsymbol{\mu}_{\boldsymbol{\theta}} - \\ &\quad 4\,\mathrm{tr}[\Lambda_1\Sigma_{\boldsymbol{\theta}}\Lambda_2\Sigma_{\boldsymbol{\theta}}] - 8\boldsymbol{\mu}_{\boldsymbol{\theta}}^T\Lambda_1\Sigma_{\boldsymbol{\theta}}\Lambda_2\boldsymbol{\mu}_{\boldsymbol{\theta}}\end{aligned} \tag{20}$$

where $\Lambda_1 := \mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})^T], \Lambda_2 := \mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})]\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})^T]$.

We remark the second moment of the variance measure (Eq. 20) is calculated as the variance of difference of two quadrature forms with respect to random variable $\boldsymbol{\theta}$, which is known to have an analytical form (see e.g., Theorem 10.9.11 of Graybill (1983)).

B. Proof of Proposition. 3.4

Lemma B.1. *For any Bayesian linear regression model: $f = \boldsymbol{\phi}^T(\boldsymbol{x})\boldsymbol{\theta}$, $\boldsymbol{\phi} \in \mathbb{R}^{N_f}, N_f \in N^+$, if $\boldsymbol{\theta} \sim \mathcal{N}(\mu_{\boldsymbol{\theta}}, \Sigma_{\boldsymbol{\theta}})$. Define its mean function as $\mu = \boldsymbol{\phi}^T(\boldsymbol{x})\boldsymbol{\mu}_{\boldsymbol{\theta}}$ and for random variable $\boldsymbol{\xi}$ we have $\mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{V}_{\boldsymbol{\xi}}(f(\boldsymbol{x}+\boldsymbol{\xi}))\right] \geq \mathbb{V}_{\boldsymbol{\xi}}(\mu(\boldsymbol{x}+\boldsymbol{\xi}))$.*

Proof.

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{V}_{\boldsymbol{\xi}}(f(\boldsymbol{x}+\boldsymbol{\xi}))\right] - \mathbb{V}_{\boldsymbol{\xi}}(\mu(\boldsymbol{x}+\boldsymbol{\xi})) \\ &= \boldsymbol{\mu}_{\boldsymbol{\theta}}^T\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{\xi})\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})^T]\boldsymbol{\mu}_{\boldsymbol{\theta}} - \left[\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})^T]\boldsymbol{\mu}_{\boldsymbol{\theta}}\right]^2 - \left[\mathbb{E}_{\boldsymbol{\xi}}[\mu^2] - \mathbb{E}_{\boldsymbol{\xi}}[\mu]^2\right] \\ &\quad + \left[\mathrm{tr}\left(\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})^T]\Sigma_{\boldsymbol{\theta}}\right) - \mathrm{tr}\left(\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})^T]\Sigma_{\boldsymbol{\theta}}\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})]\right)\right] \\ &= \left[\mathrm{tr}\left(\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})^T]\Sigma_{\boldsymbol{\theta}}\right) - \mathrm{tr}\left(\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})^T]\Sigma_{\boldsymbol{\theta}}\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})]\right)\right] \\ &= \mathrm{tr}\left(\left[\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})^T] - \mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})]\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})^T]\right]\Sigma_{\boldsymbol{\theta}}\right)\end{aligned} \tag{21}$$

Define $Z \in \mathbb{R}^{N_f}$, and $\lambda = Z^T\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})$. Then we have,

$$\begin{aligned}\mathbb{V}_{\boldsymbol{\xi}}[\lambda] &= \mathbb{E}_{\boldsymbol{\xi}}\left[(Z^T\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi}))^2\right] - \left[\mathbb{E}_{\boldsymbol{\xi}}\left[(Z^T\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi}))\right]\right]^2 \\ &= Z^T\mathbb{E}_{\boldsymbol{\xi}}\left[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})^T\right]Z - Z^T\mathbb{E}_{\boldsymbol{\xi}}\left[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})\right]\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})]^T Z\end{aligned} \tag{22}$$

By substituting $Z = \mathbf{1}$, Eq. 21 results in $\mathrm{tr}\left(\mathbb{V}_{\boldsymbol{\xi}}[\lambda]\,\mathrm{tr}(\Sigma_{\boldsymbol{\theta}})\right)$, hence the prove complete. We remark the equality can only be achieved when all the entries of $\boldsymbol{\phi}$ are constant, which is not the case for almost all the realistic Bayesian linear models. □

This also leads to the proof of proposition. 3.4. As the first line of r.h.s. of Eq. 21 is zero when $N_f \to \infty$. The remaining proof is the same as above.

C. Spectral Robustness Measures based on Random Fourier Features

C.1. Derivation for Normal Distributed Input Uncertainty

Lemma C.1. *Given* $\boldsymbol{\xi} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma), \boldsymbol{\xi} \in \mathbb{R}^D, C \in \mathbb{R}$*, the following expression holds:*

$$\mathbb{E}_{\boldsymbol{\xi}}\left[\sin(\boldsymbol{t}^T\boldsymbol{\xi} + C)\right] = \left[\sin(\boldsymbol{\mu t})\cos(C) + \cos(\boldsymbol{\mu t})\sin(C)\right] e^{-\frac{1}{2}\boldsymbol{t}^T\Sigma\boldsymbol{t}} \tag{23}$$

$$\mathbb{E}_{\boldsymbol{\xi}}\left[\cos(\boldsymbol{t}^T\boldsymbol{\xi} + C)\right] = \left[\cos(\boldsymbol{\mu t})\cos(C) - \sin(\boldsymbol{\mu t})\sin(C)\right] e^{-\frac{1}{2}\boldsymbol{t}^T\Sigma\boldsymbol{t}} \tag{24}$$

where $\boldsymbol{t} \in \mathbb{R}^D$.

Proof. Recall that $\mathbb{E}[e^{i\boldsymbol{t}^T\boldsymbol{\xi}}] = \mathbb{E}_{\boldsymbol{\xi}}[\cos(\boldsymbol{t}^T\boldsymbol{\xi})] + i\mathbb{E}_{\boldsymbol{\xi}}[\sin(\boldsymbol{t}^T\boldsymbol{\xi})]$, and that the characteristic function of $\boldsymbol{\xi}$ is $\varphi(\boldsymbol{\xi}) = \mathbb{E}_{\boldsymbol{\xi}}[e^{i\boldsymbol{t}^T\boldsymbol{\xi}}] = e^{i\boldsymbol{t}^T\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^T\Sigma\boldsymbol{t}}$, we have:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\xi}}[e^{i\boldsymbol{t}^T\boldsymbol{\xi}}] &= e^{i\boldsymbol{t}^T\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^T\Sigma\boldsymbol{t}} \\ &= \cos(\boldsymbol{t}^T\boldsymbol{\xi})e^{-\frac{1}{2}\boldsymbol{t}^T\Sigma\boldsymbol{t}} + i\sin(\boldsymbol{t}^T\boldsymbol{\xi})e^{-\frac{1}{2}\boldsymbol{t}^T\Sigma\boldsymbol{t}} \\ &= \mathbb{E}_{\boldsymbol{\xi}}[\cos(\boldsymbol{t}^T\boldsymbol{\xi})] + i\mathbb{E}_{\boldsymbol{\xi}}[\sin(\boldsymbol{t}^T\boldsymbol{\xi})]\end{aligned} \tag{25}$$

then the proof can be finished by using the sum of angle identity. □

Now we are able to give the main expression for RFF-Mean and RFF-Variance when $\boldsymbol{\xi}$ is normal distributed:

$$\mathbb{E}_{\boldsymbol{\xi}}[\phi_m(\boldsymbol{x} + \boldsymbol{\xi})^T] = \sqrt{\frac{2\sigma^2}{N_f}}\mathbb{E}_{\boldsymbol{\xi}}[\cos(\boldsymbol{\omega}_m^T L\boldsymbol{\xi} + \boldsymbol{\omega}_m^T L\boldsymbol{x} + b_m)] \tag{26}$$

With Lemma. C.1, we have:

$$\mathbb{E}_{\boldsymbol{\xi}}[\phi_m(\boldsymbol{x} + \boldsymbol{\xi})^T] = \sqrt{\frac{2\sigma^2}{N_f}}\cos(\boldsymbol{\omega}_m^T L\boldsymbol{x} + b_m)e^{-\frac{1}{2}\boldsymbol{\omega}_m^T L\Sigma L^T\boldsymbol{\omega}_m} \tag{27}$$

We now derive the first part of the variance expression in Eq. 8. For any m, n element of the feature function, its expectation with respect to input uncertainty can be calculated using the product-to-sum rule as:

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{\xi}}[\phi_m(\boldsymbol{x} + \boldsymbol{\xi})\phi_n(\boldsymbol{x} + \boldsymbol{\xi})] \\ &= \frac{2\sigma^2}{N_f}\mathbb{E}_{\boldsymbol{\xi}}[\cos(\boldsymbol{\omega}_m^T L\boldsymbol{\xi} + \boldsymbol{\omega}_m^T L\boldsymbol{x} + b_m)\cos(\boldsymbol{\omega}_n^T L\boldsymbol{\xi} + \boldsymbol{\omega}_n^T L\boldsymbol{x} + b_n)] \\ &= \frac{\sigma^2}{N_f}\mathbb{E}_{\boldsymbol{\xi}}[\cos((\boldsymbol{\omega}_m^T + \boldsymbol{\omega}_n^T)L\boldsymbol{\xi} + (\boldsymbol{\omega}_m^T + \boldsymbol{\omega}_n^T)L\boldsymbol{x} + (b_m + b_n))] + \\ &\quad \frac{\sigma^2}{N_f}\mathbb{E}_{\boldsymbol{\xi}}[\cos((\boldsymbol{\omega}_m^T - \boldsymbol{\omega}_n^T)L\boldsymbol{\xi} + (\boldsymbol{\omega}_m^T - \boldsymbol{\omega}_n^T)L\boldsymbol{x} + (b_m - b_n))] \\ &= \frac{\sigma^2}{N_f}\cos[(\boldsymbol{\omega}_m^T + \boldsymbol{\omega}_n^T)L\boldsymbol{x} + (b_m + b_n)]e^{-\frac{1}{2}(\boldsymbol{\omega}_m^T + \boldsymbol{\omega}_n^T)L\Sigma L^T(\boldsymbol{\omega}_m + \boldsymbol{\omega}_m)} + \\ &\quad \frac{\sigma^2}{N_f}\cos[(\boldsymbol{\omega}_m^T - \boldsymbol{\omega}_n^T)L\boldsymbol{x} + (b_m - b_n)]e^{-\frac{1}{2}(\boldsymbol{\omega}_m^T - \boldsymbol{\omega}_n^T)L\Sigma L^T(\boldsymbol{\omega}_m - \boldsymbol{\omega}_m)}\end{aligned} \tag{28}$$

C.2. Derivation for Uniform Distributed Input Uncertainty

We derive the analytical expression of expectation through Fourier feature mapping under uniform uncertainty. As each of the input dimensionality is independent, we perform the integration in each dimension separately using the same derivation trick of (Fröhlich et al., 2020):

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\xi}}[\phi_m(\boldsymbol{x}+\boldsymbol{\xi})^T] \\
&= \mathbb{E}_{\boldsymbol{\xi}_{\neg d}}\mathbb{E}_{\boldsymbol{\xi}_d}\left[\cos\left((\boldsymbol{\omega}_m^T L)_d(\boldsymbol{x}_d+\boldsymbol{\xi}_d) + \underbrace{(\boldsymbol{\omega}_m^T L)_{\neg d}\boldsymbol{x}_{\neg d} + b_m}_{c_d}\right)\right]
\end{aligned} \tag{29}$$

where d represents the dth input dimensionality and $\neg d$ is its complement. Note that the inner expectation is a cross-correlation between ϕ and the complex conjugate of $\boldsymbol{\xi}_d$'s probability density function p: $(\overline{p} \star \phi_d)(\boldsymbol{x}_d)$. As the conjugate of p is itself, the integral can be calculated with the property of cross-correlation (Kapinchev et al., 2015):

$$\begin{aligned}
(p \star \phi_{m,d})(\boldsymbol{x}_d) &= \mathcal{F}^{-1}\{\mathcal{F}(\overline{p(-\boldsymbol{\xi}_d)}) \cdot \mathcal{F}(\phi_{m,d})\} \\
&= \mathcal{F}^{-1}\{\mathcal{F}(p) \cdot \mathcal{F}(\phi_{m,d})\}
\end{aligned} \tag{30}$$

where $\mathcal{F}$ represents a Fourier transformation and $\mathcal{F}^{-1}$ is its inverse.

The standard Fourier transformation (i.e., $\int f(t)e^{-j\omega t}dt$) of a uniform distribution: $u \sim [\delta_{\boldsymbol{l}_d}, \delta_{\boldsymbol{u}_d}]$, by assuming the distribution is symmetric (i.e., $\delta_{\boldsymbol{u}_d} = -\delta_{\boldsymbol{l}_d} := \delta_d$), can be given as:

$$\mathcal{F}\left(\frac{1}{2\delta_d}\right) = \frac{\sin(\omega\delta_d)}{\omega\delta_d} \tag{31}$$

For convenience, the standard Fourier transform of ϕ is provided here as well:

$$\begin{aligned}
&\mathcal{F}(\phi_{m,d}) \\
&= \sqrt{\frac{2\sigma^2}{N_f}}\mathcal{F}(\cos((\omega_{m,d}L_{d,d})\boldsymbol{x}_d + c_d)) \\
&= \sqrt{\frac{2\sigma^2}{N_f}}\int \frac{e^{jc_d}e^{j(\omega_{m,d}L_{d,d}\boldsymbol{x}_d)} + e^{-jc_d}e^{-j(\omega_{m,d}L_{d,d}\boldsymbol{x}_d)}}{2} e^{-j\omega\boldsymbol{x}_d}\boldsymbol{x}_d \\
&= \sqrt{\frac{2\sigma^2}{N_f}}\pi\left(e^{jc_d}\delta(\omega-\omega_{m,d}L_{d,d}) + e^{-jc_d}\delta(\omega+\omega_{m,d}L_{d,d})\right)
\end{aligned} \tag{32}$$

Finally, multiplying Eq. 31 and Eq. 32 and applying the inverse Fourier transformation leads to:

$$\begin{aligned}
&\int \mathcal{F}\left(\frac{1}{2\delta_d}\right) \cdot \mathcal{F}(\phi_{m,d}) exp(j\omega x_d) d\omega \\
&= \frac{1}{2\pi}\int \sqrt{\frac{2\sigma^2}{N_f}}\pi\left(e^{jc_d}\delta(\omega-\omega_{m,d}L_{d,d}) + e^{-jc_d}\delta(\omega+\omega_{m,d}L_{d,d})\right)\frac{\sin(\omega\delta_d)}{\omega\delta_d}e^{j\omega x_d}d\omega \\
&= \sqrt{\frac{2\sigma^2}{N_f}}\frac{\sin\omega_{m,d}L_{d,d}\delta_d}{\omega_{m,d}L_{d,d}\delta_d}\frac{1}{2}\left[e^{(j(\omega_{m,d}L_{d,d}x_d+c_d))} + e^{(-j(\omega_{m,d}L_{d,d}x_d+c_d))}\right] \\
&= \frac{\sin\omega_{m,d}L_{d,d}\delta_d}{\omega_{m,d}L_{d,d}\delta_d}\sqrt{\frac{2\sigma^2}{N_f}}\cos(\omega_{m,d}L_{d,d}x_d + c_d) \\
&= \frac{\sin\omega_{m,d}L_{d,d}\delta_d}{\omega_{m,d}L_{d,d}\delta_d}\phi_{m,d}
\end{aligned} \tag{33}$$

extending to an arbitrary input dimensionality D, this leads to:

$$\mathbb{E}_{\boldsymbol{\xi}}[\phi_m(\boldsymbol{x}+\boldsymbol{\xi})^T] = \phi_m(\boldsymbol{x}) \prod_{d=1}^{D} \frac{\sin\left(\omega_{m,d} L_{d,d} \delta_d\right)}{\omega_{m,d} L_{d,d} \delta_d} = \phi_m \frac{\sin\left(\omega_m L \boldsymbol{\delta}\right)}{\omega_m L \boldsymbol{\delta}} \tag{34}$$

Similarly, we can calculate the following integration when $\boldsymbol{\xi}$ is uniformly distributed:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\xi}}[\phi_m(\boldsymbol{x}+\boldsymbol{\xi})\phi_n(\boldsymbol{x}+\boldsymbol{\xi})] \\
&= \frac{2\sigma^2}{N_f}\mathbb{E}_{\boldsymbol{\xi}}[\cos(\boldsymbol{\omega}_m^T L\boldsymbol{\xi} + \boldsymbol{\omega}_m^T L\boldsymbol{x} + b_m)\cos(\boldsymbol{\omega}_n^T L\boldsymbol{\xi} + \boldsymbol{\omega}_n^T L\boldsymbol{x} + b_n)] \\
&= \frac{\sigma^2}{N_f}\mathbb{E}_{\boldsymbol{\xi}}[\cos((\boldsymbol{\omega}_m^T + \boldsymbol{\omega}_n^T)L\boldsymbol{\xi} + (\boldsymbol{\omega}_m^T + \boldsymbol{\omega}_n^T)L\boldsymbol{x} + (b_m + b_n))] + \\
&\quad \frac{\sigma^2}{N_f}\mathbb{E}_{\boldsymbol{\xi}}[\cos((\boldsymbol{\omega}_m^T - \boldsymbol{\omega}_n^T)L\boldsymbol{\xi} + (\boldsymbol{\omega}_m^T - \boldsymbol{\omega}_n^T)L\boldsymbol{x} + (b_m - b_n))] \\
&= \frac{\sigma^2}{N_f}\cos((\boldsymbol{\omega}_m^T - \boldsymbol{\omega}_n^T)L\boldsymbol{x} + (b_m - b_n))\frac{\sin\left((\omega_m^T - \omega_n^T)L\boldsymbol{\delta}\right)}{(\omega_m^T - \omega_n^T)L\boldsymbol{\delta}} + \\
&\quad \frac{\sigma^2}{N_f}\cos((\boldsymbol{\omega}_m^T + \boldsymbol{\omega}_n^T)L\boldsymbol{x} + (b_m + b_n))\frac{\sin\left((\omega_m^T + \omega_n^T)L\boldsymbol{\delta}\right)}{(\omega_m^T + \omega_n^T)L\boldsymbol{\delta}}
\end{aligned} \tag{35}$$

D. Spectral Robustness Measures based on Quadrature Fourier Features

D.1. Quadrature Fourier Feature based Robustness Measure

Table 2: Expressions of quadrature Fourier features based robustness measures

$\boldsymbol{\xi}$ distribution	$\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})_m^T]$ (Mean)	$\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})_m \boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})_n^T]$
$\mathcal{N}(\mathbf{0}, \Sigma)$	$1^\circ m \le N_f$:	$1^\circ m \le N_f, n \le N_f$:
	$\sigma\sqrt{\prod_{i=1}^{d} v(\omega_{m,d})}\cos(\boldsymbol{\omega}_m^T L\boldsymbol{x})\cdot e^{-\frac{1}{2}\boldsymbol{\omega}_m^T L\Sigma L^T \boldsymbol{\omega}_m}$	$\frac{\sigma^2}{2}\sqrt{\prod_{i=1}^{d} v(\omega_{m,d})v(\omega_{n,d})}\cdot \cos[(\boldsymbol{\omega}_m^T + \boldsymbol{\omega}_n^T)L\boldsymbol{x}]\cdot e^{-\frac{1}{2}(\boldsymbol{\omega}_m^T + \boldsymbol{\omega}_n^T)L\Sigma L^T(\boldsymbol{\omega}_m + \boldsymbol{\omega}_m)} + \cos[(\boldsymbol{\omega}_m^T - \boldsymbol{\omega}_n^T)L\boldsymbol{x}]\cdot e^{-\frac{1}{2}(\boldsymbol{\omega}_m^T - \boldsymbol{\omega}_n^T)L\Sigma L^T(\boldsymbol{\omega}_m - \boldsymbol{\omega}_m)}$
	$2^\circ N_f < m \le 2N_f$:	$2^\circ N_f < m \le 2N_f, n \le N_f$:
	$\sigma\sqrt{\prod_{i=1}^{d} v(\omega_{m-N_f,d})}\cdot \sin(\omega_{m-N_f}^T L\boldsymbol{x})\cdot e^{-\frac{1}{2}\boldsymbol{\omega}_{m-N_f}^T L\Sigma L^T \boldsymbol{\omega}_{m-N_f}}$	$\frac{\sigma^2}{2}\sqrt{\prod_{i=1}^{d} v(\omega_{m-N_f,d})v(\omega_{n,d})}\Big[\sin((\boldsymbol{\omega}_{m-N_f}^T + \boldsymbol{\omega}_n^T)L\boldsymbol{x})\cdot e^{-\frac{1}{2}(\boldsymbol{\omega}_{m-N_f}^T + \boldsymbol{\omega}_n^T)L\Sigma L^T(\boldsymbol{\omega}_{m-N_f} + \boldsymbol{\omega}_n)} + \sin((\boldsymbol{\omega}_{m-N_f}^T - \boldsymbol{\omega}_n^T)L\boldsymbol{x})\cdot e^{-\frac{1}{2}(\boldsymbol{\omega}_{m-N_f}^T - \boldsymbol{\omega}_n^T)L\Sigma L^T(\boldsymbol{\omega}_{m-N_f} - \boldsymbol{\omega}_n)}\Big]$
		$3^\circ m \le N_f, N_f < n \le 2N_f$:

		$\frac{\sigma^2}{2}\sqrt{\prod_{i=1}^{d} v(\omega_{m,d})v(\omega_{n-N_f,d})}\Big[$ $\sin((\boldsymbol{\omega}_m^T + \boldsymbol{\omega}_{n-N_f}^T)L\boldsymbol{x})\cdot$ $e^{-\frac{1}{2}(\boldsymbol{\omega}_m^T+\boldsymbol{\omega}_{n-N_f}^T)L^T\Sigma L(\boldsymbol{\omega}_m+\boldsymbol{\omega}_{n-N_f})}+$ $\sin((-\boldsymbol{\omega}_m^T + \boldsymbol{\omega}_{n-N_f}^T)L\boldsymbol{x})\cdot$ $e^{-\frac{1}{2}(-\boldsymbol{\omega}_m^T+\boldsymbol{\omega}_{n-N_f}^T)L\Sigma L^T(-\boldsymbol{\omega}_m+\boldsymbol{\omega}_{n-N_f})}\Big]$ $4^\circ N_f < m \le 2N_f, N_f < n \le 2N_f:$ $\frac{\sigma^2}{2}\sqrt{\prod_{i=1}^{d} v(\omega_{m-N_f,d})v(\omega_{n-N_f,d})}\cdot\Big[$ $\cos((\boldsymbol{\omega}_{m-N_f}^T - \boldsymbol{\omega}_{n-N_f}^T)L\boldsymbol{x})\cdot$ $e^{-\frac{1}{2}(\boldsymbol{\omega}_{m-N_f}^T-\boldsymbol{\omega}_{n-N_f}^T)L^T\Sigma L(\boldsymbol{\omega}_{m-N_f}-\boldsymbol{\omega}_{n-N_f})}-$ $\cos((\boldsymbol{\omega}_{m-N_f}^T + \boldsymbol{\omega}_{n-N_f}^T)L\boldsymbol{x})\cdot$ $e^{-\frac{1}{2}(\boldsymbol{\omega}_{m-N_f}^T+\boldsymbol{\omega}_{n-N_f}^T)L\Sigma L^T(\boldsymbol{\omega}_{m-N_f}+\boldsymbol{\omega}_{n-N_f})}\Big]$
$U(-\boldsymbol{\delta},\boldsymbol{\delta})$	$1^\circ m < N_f$ $\sigma\sqrt{\prod_{i=1}^{d} v(\omega_{m,d})}\cdot$ $\cos(\boldsymbol{\omega}_m^T L\boldsymbol{x})\frac{\sin(\omega_m L\boldsymbol{\delta})}{\omega_m L\boldsymbol{\delta}}$ $2^\circ N_f < m \le 2N_f:$ $\sigma\sqrt{\prod_{i=1}^{d} v(\omega_{m-N_f,d})}\cdot$ $\sin(\omega_{m-N_f}^T L\boldsymbol{x})\cdot$ $\frac{\sin\left(\omega_{m-N_f} L\boldsymbol{\delta}\right)}{\omega_{m-N_f} L\boldsymbol{\delta}}$	$1^\circ m \le N_f, n \le N_f:$ $\frac{\sigma^2}{2}\sqrt{\prod_{i=1}^{d} v(\omega_{m-N_f,d})v(\omega_{n,d})}\cdot\Big[$ $\cos((\boldsymbol{\omega}_m^T - \boldsymbol{\omega}_n^T)L\boldsymbol{x})\cdot\frac{\sin\left((\omega_m^T - \omega_n^T)L\delta\right)}{(\omega_m^T - \omega_n^T)L\delta}+$ $\cos((\boldsymbol{\omega}_m^T + \boldsymbol{\omega}_n^T)L\boldsymbol{x})\frac{\sin\left((\omega_m^T + \omega_n^T)L\delta\right)}{(\omega_m^T + \omega_n^T)L\delta}\Big]$ $2^\circ m \le N_f, N_f < n \le 2N_f:$ $\frac{\sigma^2}{2}\sqrt{\prod_{i=1}^{d} v(\omega_{m-N_f,d})v(\omega_{n,d})}\cdot\Big[$ $\sin\left((\boldsymbol{\omega}_{m-N_f}^T + \boldsymbol{\omega}_n^T)L\boldsymbol{x}\right)$ $\frac{\sin\left((\boldsymbol{\omega}_{m-N_f}^T + \boldsymbol{\omega}_n^T)L\boldsymbol{\delta}\right)}{(\boldsymbol{\omega}_{m-N_f}^T + \boldsymbol{\omega}_n^T)L\boldsymbol{\delta}}+$ $\sin\left((\boldsymbol{\omega}_{m-N_f}^T - \boldsymbol{\omega}_n^T)L\boldsymbol{x}\right)$ $\frac{\sin\left((\boldsymbol{\omega}_{m-N_f}^T - \boldsymbol{\omega}_n^T)L\boldsymbol{\delta}\right)}{(\boldsymbol{\omega}_{m-N_f}^T - \boldsymbol{\omega}_n^T)L\boldsymbol{\delta}}\Big]$ $3^\circ N_f < m \le 2N_f, n \le N_f:$

		$\frac{\sigma^2}{2}\sqrt{\prod_{i=1}^{d} v(\omega_{m-N_f,d})v(\omega_{n,d})}\cdot\Bigg[\sin\left((\boldsymbol{\omega}_{m-N_f}^T+\boldsymbol{\omega}_n^T)L\boldsymbol{x}\right) \frac{\sin\left((\boldsymbol{\omega}_{m-N_f}^T+\boldsymbol{\omega}_n^T)L\boldsymbol{\delta}\right)}{(\boldsymbol{\omega}_{m-N_f}^T+\boldsymbol{\omega}_n^T)L\boldsymbol{\delta}} + \sin\left((-\boldsymbol{\omega}_{m-N_f}^T+\boldsymbol{\omega}_n^T)L\boldsymbol{x}\right)\cdot \frac{\sin\left((-\boldsymbol{\omega}_{m-N_f}^T+\boldsymbol{\omega}_n^T)L\boldsymbol{\delta}\right)}{(-\boldsymbol{\omega}_{m-N_f}^T+\boldsymbol{\omega}_n^T)L\boldsymbol{\delta}}\Bigg]$
		$4^\circ\ N_f < m \leq 2N_f, N_f < n \leq 2N_f:$
		$\frac{\sigma^2}{2}\sqrt{\prod_{i=1}^{d} v(\omega_{m-N_f,d})v(\omega_{n,d})}\cdot\Bigg[\cos\left((\boldsymbol{\omega}_{m-N_f}^T-\boldsymbol{\omega}_{n-N_f}^T)L\boldsymbol{x}\right)\cdot \frac{\sin\left((\boldsymbol{\omega}_{m-N_f}^T-\boldsymbol{\omega}_{n-N_f}^T)L\boldsymbol{\delta}\right)}{(\boldsymbol{\omega}_{m-N_f}^T-\boldsymbol{\omega}_{n-N_f}^T)L\boldsymbol{\delta}} - \cos\left((\boldsymbol{\omega}_{m-N_f}^T+\boldsymbol{\omega}_{n-N_f}^T)L\boldsymbol{x}\right) \frac{\left(\sin(\boldsymbol{\omega}_{m-N_f}^T+\boldsymbol{\omega}_{n-N_f}^T)L\boldsymbol{\delta}\right)}{(\boldsymbol{\omega}_{m-N_f}^T+\boldsymbol{\omega}_{n-N_f}^T)L\boldsymbol{\delta}}\Bigg]$

D.2. Derivation for Normal Distributed Input Uncertainty

To get the mean expression, when $m \leq N_f$ in Eq. 11, the expression from Eq. 27 can be reused:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\xi}}[\phi_m(\boldsymbol{x}+\boldsymbol{\xi})^T] &= \sigma\sqrt{\prod_{i=1}^{d} v(\omega_{m,d})}\,\mathbb{E}_{\boldsymbol{\xi}}\left[\cos\left(\omega_m^T L\boldsymbol{x}+\boldsymbol{\omega}_m^T L\boldsymbol{\xi}\right)\right] \\ &= \sigma\sqrt{\prod_{i=1}^{d} v(\omega_{m,d})}\,\cos(\boldsymbol{\omega}_m^T L\boldsymbol{x})e^{-\frac{1}{2}\boldsymbol{\omega}_m^T L\Sigma L^T\boldsymbol{\omega}_m}\end{aligned} \tag{36}$$

When $m > N_f$:

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\xi}}[\phi_m(\boldsymbol{x}+\boldsymbol{\xi})^T] &= \sigma\sqrt{\prod_{i=1}^{d} v(\omega_{m-N_f,d})}\,\mathbb{E}_{\boldsymbol{\xi}}\left[\sin\left(\omega_{m-N_f}^T L\boldsymbol{x}+\boldsymbol{\omega}_{m-N_f}^T L\boldsymbol{\xi}\right)\right] \\ &= \sigma\sqrt{\prod_{i=1}^{d} v(\omega_{m-N_f,d})}\,\sin(\omega_{m-N_f}^T L\boldsymbol{x})e^{-\frac{1}{2}\boldsymbol{\omega}_{m-N_f}^T L\Sigma L^T\boldsymbol{\omega}_{m-N_f}}\end{aligned} \tag{37}$$

For the expectation of $\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})_m\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})_n^T]$, when $N_f < m \leq 2N_f$ and $n \leq N_f$, we have:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\xi}}[\phi_m(\boldsymbol{x}+\boldsymbol{\xi})\phi_n(\boldsymbol{x}+\boldsymbol{\xi})] \\
&= \sigma^2 \sqrt{\prod_{i=1}^{d} v(\omega_{m-N_f,d})v(\omega_{n,d})} \cdot \mathbb{E}_{\boldsymbol{\xi}}\left[\sin(\boldsymbol{\omega}_{m-N_f}^T L\boldsymbol{\xi} + \boldsymbol{\omega}_{m-N_f}^T L\boldsymbol{x})\cos(\boldsymbol{\omega}_n^T L\boldsymbol{\xi} + \boldsymbol{\omega}_n^T L\boldsymbol{x})\right] \\
&= \frac{\sigma^2}{2}\sqrt{\prod_{i=1}^{d} v(\omega_{m-N_f,d})v(\omega_{n,d})}\sin\left((\boldsymbol{\omega}_{m-N_f}^T + \boldsymbol{\omega}_n^T)L\boldsymbol{x}\right)e^{-\frac{1}{2}(\boldsymbol{\omega}_{m-N_f}^T+\boldsymbol{\omega}_n^T)L\Sigma L^T(\boldsymbol{\omega}_{m-N_f}+\boldsymbol{\omega}_n)} + \\
&\frac{\sigma^2}{2}\sqrt{\prod_{i=1}^{d} v(\omega_{m-N_f,d})v(\omega_{n,d})}\sin\left((\boldsymbol{\omega}_{m-N_f}^T - \boldsymbol{\omega}_n^T)L\boldsymbol{x}\right)e^{-\frac{1}{2}(\boldsymbol{\omega}_{m-N_f}^T-\boldsymbol{\omega}_n^T)L\Sigma L^T(\boldsymbol{\omega}_{m-N_f}-\boldsymbol{\omega}_n)}
\end{aligned} \tag{38}$$

The other three conditions can be derived in the same fashion.

D.3. Derivation for Uniform Distributed Input Uncertainty

We need the Fourier transformation of $\cos(\cdot)\sin(\cdot)$ and $\sin(\cdot)\sin'(\cdot)$, $\cos(\cdot)\cos'(\cdot)$. By making use of the prodct-to-summation rule, we, in fact, only need the Fourier transformation of $\sin$ function. Starting from Eq. 36 and Eq. 37:

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\xi}}[\phi_m(\boldsymbol{x}+\boldsymbol{\xi})^T] &= \sigma\sqrt{\prod_{i=1}^{d} v(\omega_{m,d})}\mathbb{E}_{\boldsymbol{\xi}}\left[\cos\left(\omega_m^T L\boldsymbol{x} + \boldsymbol{\omega}_m^T L\boldsymbol{\xi}\right)\right] \\
&= \sigma\sqrt{\prod_{i=1}^{d} v(\omega_{m,d})}\cos(\boldsymbol{\omega}_m^T L\boldsymbol{x})\mathbb{E}_{\boldsymbol{\xi}}\left[\cos(\boldsymbol{\omega}_m^T L\boldsymbol{\xi})\right] \\
&= \sigma\sqrt{\prod_{i=1}^{d} v(\omega_{m,d})}\cos(\boldsymbol{\omega}_m^T L\boldsymbol{x})\frac{\sin(\omega_m L\boldsymbol{\delta})}{\omega_m L\boldsymbol{\delta}}
\end{aligned} \tag{39}$$

When $m > N_f$:

$$\mathbb{E}_{\boldsymbol{\xi}}[\phi_m(\boldsymbol{x}+\boldsymbol{\xi})^T] = \sigma\sqrt{\prod_{i=1}^{d} v(\omega_{m-N_f,d})}\sin(\omega_{m-N_f}^T L\boldsymbol{x})\frac{\sin(\omega_m L\boldsymbol{\delta})}{\omega_m L\boldsymbol{\delta}} \tag{40}$$

For the expectation of $\mathbb{E}_{\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})_m\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi})_n^T]$, with $N_f < m \leq 2N_f$ and $n \leq N_f$, we have:

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\xi}}[\phi_m(\boldsymbol{x}+\boldsymbol{\xi})\phi_n(\boldsymbol{x}+\boldsymbol{\xi})] \\
&= \sigma^2 \sqrt{\prod_{i=1}^{d} v(\omega_{m-N_f,d})v(\omega_{n,d})} \cdot \mathbb{E}_{\boldsymbol{\xi}}\left[\sin(\boldsymbol{\omega}_{m-N_f}^T L\boldsymbol{\xi} + \boldsymbol{\omega}_{m-N_f}^T L\boldsymbol{x})\cos(\boldsymbol{\omega}_n^T L\boldsymbol{\xi} + \boldsymbol{\omega}_n^T L\boldsymbol{x})\right] \\
&= \frac{\sigma^2}{2}\sqrt{\prod_{i=1}^{d} v(\omega_{m-N_f,d})v(\omega_{n,d})}\sin\left((\boldsymbol{\omega}_{m-N_f}^T + \boldsymbol{\omega}_n^T)L\boldsymbol{x}\right)\frac{\sin\left((\boldsymbol{\omega}_{m-N_f}^T + \boldsymbol{\omega}_n^T)L\boldsymbol{\delta}\right)}{(\boldsymbol{\omega}_{m-N_f}^T + \boldsymbol{\omega}_n^T)L\boldsymbol{\delta}} + \\
&\frac{\sigma^2}{2}\sqrt{\prod_{i=1}^{d} v(\omega_{m-N_f,d})v(\omega_{n,d})}\sin\left((\boldsymbol{\omega}_{m-N_f}^T - \boldsymbol{\omega}_n^T)L\boldsymbol{x}\right)\frac{\sin\left(\boldsymbol{\omega}_{m-N_f}^T - \boldsymbol{\omega}_n^T)L\boldsymbol{\delta}\right)}{(\boldsymbol{\omega}_{m-N_f}^T - \boldsymbol{\omega}_n^T)L\boldsymbol{\delta}}
\end{aligned} \tag{41}$$

The rest of the cases can be obtained in a similar fashion.

E. Spectral Robustness Measures Supporting Continuous Environmental Variable

In real-life applications, another commonly investigated problem is the robustness with respect to environmental variables (e.g., Beland & Nair (2017)). Let $f : \mathcal{X} \times \mathbf{\Gamma} \to \mathbb{R}$, where $\boldsymbol{\gamma}$ represents the environmental variable defined in a bounded space $\mathbf{\Gamma}$. In these kinds of problems, We are interested in the following extended robustness measures:

$$\begin{aligned}\mathbb{J}_{\boldsymbol{\gamma},\boldsymbol{\xi}}(\mathcal{GP}(f)) &\approx \tilde{\mathbb{J}}(\mathcal{GP}(f))|\boldsymbol{\phi},\boldsymbol{\theta} = \mathbb{E}_{\boldsymbol{\gamma},\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{\gamma},\boldsymbol{x}+\boldsymbol{\xi})^T]\boldsymbol{\theta} \\ \mathbb{V}_{\boldsymbol{\gamma},\boldsymbol{\xi}}(\mathcal{GP}(f)) &\approx \tilde{\mathbb{V}}(\mathcal{GP}(f))|\boldsymbol{\phi},\boldsymbol{\theta} \\ &= \boldsymbol{\theta}^T\mathbb{E}_{\boldsymbol{\gamma},\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{\gamma},\boldsymbol{x}+\boldsymbol{\xi})\boldsymbol{\phi}(\boldsymbol{\gamma},\boldsymbol{x}+\boldsymbol{\xi})^T]\boldsymbol{\theta} - \left[\mathbb{E}_{\boldsymbol{\gamma},\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{\gamma},\boldsymbol{x}+\boldsymbol{\xi})^T]\boldsymbol{\theta}\right]^2\end{aligned} \tag{42}$$

Assuming we are able query $f(\boldsymbol{x},\boldsymbol{\gamma})$ exactly, e.g., in a simulating scenario. It turns out that this problem can be regarded as a special case of Eq. 1, where the environmental variable is reformulated as a design variable located at the center of $\mathbf{\Gamma}$ with uniform distributed input uncertainty $\boldsymbol{\xi}$.

Hence, for RFF based robustness measures, we give their expressions in Table 3.

Table 3. Expression of the RFF based robustness measures supporting environmental variables

$\boldsymbol{\gamma}$ distribution	$\mathbb{E}_{\boldsymbol{\gamma},\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi},\boldsymbol{\gamma})_m^T]$ (Mean)	$\mathbb{E}_{\boldsymbol{\gamma},\boldsymbol{\xi}}[\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi},\boldsymbol{\gamma})_m\boldsymbol{\phi}(\boldsymbol{x}+\boldsymbol{\xi},\boldsymbol{\gamma})_n^T]$
$U(\mathbf{\Gamma}_l,\mathbf{\Gamma}_u)$	$\sqrt{\frac{2\sigma^2}{N_f}}\cos(\boldsymbol{\omega}_{m_{\boldsymbol{x}}}^T L_{\boldsymbol{x}}\boldsymbol{x} + \boldsymbol{\omega}_{m_{\boldsymbol{\gamma}}}^T L_{\boldsymbol{\gamma}}\overline{\mathbf{\Gamma}} + b_m)\cdot\frac{\sin\left(\omega_{m_{\boldsymbol{\gamma}}}L_{\boldsymbol{\gamma}}\boldsymbol{\delta}_{\boldsymbol{\gamma}} + \omega_{m_{\boldsymbol{x}}}L_{\boldsymbol{x}}\boldsymbol{\delta}_{\boldsymbol{x}}\right)}{\left(\omega_{m_{\boldsymbol{\gamma}}}L_{\boldsymbol{\gamma}}\boldsymbol{\delta}_{\boldsymbol{\gamma}} + \omega_{m_{\boldsymbol{x}}}L_{\boldsymbol{x}}\boldsymbol{\delta}_{\boldsymbol{x}}\right)}$	$\sum_{\bigcirc\in\{+,-\}}\left[\frac{\sigma^2}{N_f}\cos\left[(\boldsymbol{\omega}_{m_{\boldsymbol{x}}}^T\bigcirc\boldsymbol{\omega}_{n_{\boldsymbol{x}}}^T)L_{\boldsymbol{x}}\boldsymbol{x} + (\boldsymbol{\omega}_{m_{\boldsymbol{\gamma}}}^T\bigcirc\boldsymbol{\gamma}_{n_{\boldsymbol{x}}}^T)L_{\boldsymbol{\gamma}}\overline{\mathbf{\Gamma}} + (b_m\bigcirc b_n)\right]\cdot\frac{\sin[(\omega_{m_{\boldsymbol{\gamma}}}^T\bigcirc\omega_{n_{\boldsymbol{\gamma}}}^T)L_{\boldsymbol{\gamma}}\boldsymbol{\delta}_{\boldsymbol{\gamma}} + (\omega_{m_{\boldsymbol{x}}}^T\bigcirc\omega_{n_{\boldsymbol{x}}}^T)L_x\boldsymbol{\delta}_{\boldsymbol{x}}]}{(\omega_{m_{\boldsymbol{\gamma}}}^T\bigcirc\omega_{n_{\boldsymbol{\gamma}}}^T)L_{\boldsymbol{\gamma}}\boldsymbol{\delta}_{\boldsymbol{\gamma}} + (\omega_{m_{\boldsymbol{x}}}^T\bigcirc\omega_{n_{\boldsymbol{x}}}^T)L_x\boldsymbol{\delta}_{\boldsymbol{x}}}\right]$

where $\boldsymbol{\delta}_{\boldsymbol{\gamma}} = (\mathbf{\Gamma}_u - \mathbf{\Gamma}_l)/2$, $\overline{\mathbf{\Gamma}} = (\mathbf{\Gamma}_u + \mathbf{\Gamma}_l)/2$. We remark the same strategy can be utilized for QFF based robustness measures under environmental uncertainty.

F. Computational Complexity

Table 4. Inference complexity of RFF and QFF based robustness measures.

Category	Method	Computation Cost
Point Estimation (Theorem 3.3)		
	RFF-$\mathbb{E}_{\boldsymbol{\theta}}[\tilde{\mathbb{J}}]$	$N_f d$
	RFF-$\mathbb{E}_{\boldsymbol{\theta}}[\tilde{\mathbb{V}}]$	$N_f^2 d$
	QFF-$\mathbb{E}_{\boldsymbol{\theta}}[\tilde{\mathbb{J}}]$	$2N_f d$
	QFF-$\mathbb{E}_{\boldsymbol{\theta}}[\tilde{\mathbb{V}}]$	$4N_f^2 d$
Bayesian Inference		
	RFF-Mean	Init: $N_f^3 d$, Query: $N_f d$
	QFF-Mean	Init: $8N_f^3 d$, Query: $2N_f d$
	RFF-Variance	Init: $N_f^3 d$, Query: $N_f^2 d$
	QFF-Variance	Init: $8N_f^3 d$, Query: $4N_f^2 d$

We provide inference complexity on sample trajectories in Table 4. The Initialization cost is the Cholesky decomposition of $\boldsymbol{\theta}$ posterior covariance matrix (i.e., Eq. 4). At this stage, given the same Fourier feature number N_f, the QFF based strategies are 8 times more expensive than the corresponding RFF based strategies. For the query stage, the variance inference's complexity for both Fourier features scales quadratic with the number of Fourier features. Besides, we also remark that if

only utilizing the posterior mean of the robustness measure (i.e., Theorem 3.3), one could avoid the cubic initialization cost. We note a sequential strategy can be used through test data for memory-efficient inference.

G. Batch Query Support for the Acquisition Functions

Parallel computation resources can be available in real-life applications, where a practitioner can be interested in query more than one point per iteration to booster the optimization process. We remark that the joint posterior distribution condition on multiple candidate points is supported in the spectral robustness measure representation. Hence, all of the Fourier feature-based acquisition functions we described can easily be extended in this scenario, that is, one can easily obtain FF-MV-qEHVI, FF-MV-qECI and FF-MV-qEI acquisition functions (q is commonly given as a representation of acquisition function is supported for batch query, e.g., Daulton et al. (2020); Wang et al. (2020b)). For FF-MV-EHVI, the calculation of $\text{HVI}(\tilde{\mathbb{J}}_i, \hat{\mathbb{V}}_i, \tilde{\mathcal{F}}^*_{MV_i} | \boldsymbol{\theta}_i, D)$ behaves naturally as the cached box decomposition (Daulton et al., 2021). For FF-MV-ECI and FF-MV-EI, Fourier feature-based posterior supports batch evaluation on the differentiable trajectories, which acts similar to the sample average approximation (Balandat et al., 2019). We leave such an extension and thorough performance assessment as future work.

H. Additional Experiments

H.1. Comparison of the First Moments of the Robustness Measures

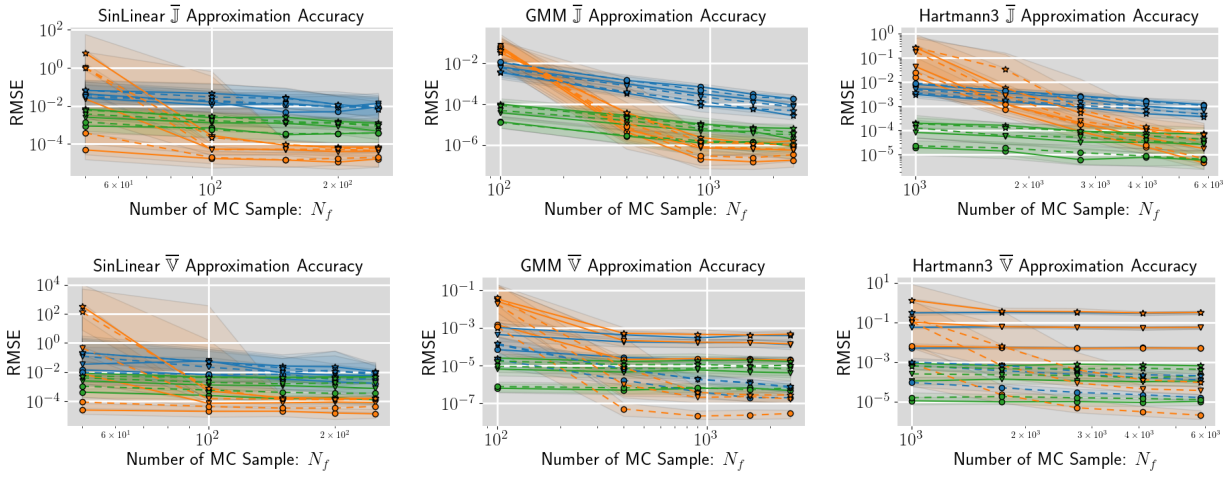

Figure 6. Comparison of robustness measure model accuracy through MC in input space vs on spectral density on the first moments $(\mathbb{J}, \mathbb{V})$. Dashed lines represents normal input uncertainty: $\boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma I)$ and the solid line rest represents uniform : $\boldsymbol{\xi} \sim U(-\mathbf{1}\delta, \mathbf{1}\delta)$. We use {orange, blue, green} to represent {QFF, RFF, MC} respectively. We use $\{\bigcirc, \bigtriangledown, \star\}$ to represent $\sigma = \{0.001, 0.005, 0.01\}$ for normal and $\delta = \{0.05, 0.1, 0.2\}$ for uniform uncertainty respectively.

Estimating the robustness measure through MC sampling in the input space on the GP posterior mean is a common non-Bayesian strategy (i.e., a point estimation) used in the robust optimization community (e.g., Rivier & Congedo (2018)). Since we also provide the first moment of robustness measures through Fourier features, we conduct empirical experiments comparing the model accuracy of robustness measures between the two approaches (i.e., we compare how well our approaches approximate $\overline{\mathbb{J}} := \mathbb{E}_f\left[\mathbb{J}\left(\mathcal{GP}(f)\right)\right]$ and $\overline{\mathbb{V}} := \mathbb{E}_f\left[\mathbb{V}\left(\mathcal{GP}(f)\right)\right]$). We conduct experiment on synthetic functions (GMM function is from (Nogueira et al., 2016)) ranging from $1-3$ input dimensionality. For each synthetic problem, we use $20d$ training data to construct a GP with SE kernel. This number is chosen to avoid either an inaccurate model (i.e., the GP posterior collapses into its prior) or an overconfident model (the model uncertainty vanishes). Then, we compare the modeling error of the robustness measures between MC in the input space and using spectral density (i.e., Fourier feature based approximation). For a fair comparison, we use the same MC sample numbers [11].

The benchmark results are provided in Fig. 6. For each experiment, we calculate the Root-Mean-Squared-Error (RMSE)

[11] For the Fourier feature based method, the MC sample number is used as N_f.

between the approximation and the ground truth extracted from the GP posterior [12] based on a $20 * 2^d$ test sample size. The experiments are conducted for a different number of Fourier features, and each experiment is repeated 64 times for robustness. It can be seen that, especially for the QFF robustness measure, its first moment can reach a better accuracy, provided there are enough MC samples.

H.2. Overview of the Acquisition Functions for Comparison

We provide an overview of the acquisition functions from Iwazaki et al. (2021) in this section. As the second objective F_2 (Eq. 3 of (Iwazaki et al., 2021)) represents negative standard deviation, we modify it to be the variance to compare in the same setting. As a result, the upper and lower bound for variance can be calculated as:

$$l_t^{F_2}(\boldsymbol{x}) = \int \tilde{l}_t^{sq}(\boldsymbol{x} + \boldsymbol{\xi}) p(\boldsymbol{\xi}) d\boldsymbol{\xi} \tag{43}$$

$$u_t^{F_2}(\boldsymbol{x}) = \int \tilde{u}_t^{sq}(\boldsymbol{x} + \boldsymbol{\xi}) p(\boldsymbol{\xi}) d\boldsymbol{\xi} \tag{44}$$

where $\tilde{l}_t^{sq}, \tilde{u}_t^{sq}$ are defined the same as in (Iwazaki et al., 2021), we use 200 MC samples to approximate the integration. As the code is not open-sourced yet, we implement their acquisition function for numerical comparison.

We remark that at iteration t, one is optimizing the following *two-stage* acquisition optimization process (see appendix B.2 of (Iwazaki et al., 2021)) and evaluate $f(\tilde{\boldsymbol{x}}_t + \tilde{\boldsymbol{\xi}}_t)$:

$$\tilde{\boldsymbol{x}}_t = \arg\max_{\boldsymbol{x} \in \mathcal{X}} \alpha_t(\boldsymbol{x}) \tag{45}$$

$$\tilde{\boldsymbol{\xi}}_t = \arg\max_{\boldsymbol{\xi} \in \Delta} \sigma_t(\tilde{\boldsymbol{x}}_t + \boldsymbol{\xi}) p(\boldsymbol{\xi}) \tag{46}$$

where σ_t represents the standard deviation of the GP posterior. For the definition of Δ, since the original paper does not mention an extension to the input uncertainty scenario, we define it as: $\Delta := [-\boldsymbol{\delta}, \boldsymbol{\delta}]$ when $\boldsymbol{\xi} \sim U[-\boldsymbol{\delta}, \boldsymbol{\delta}]$. When $\boldsymbol{\xi}$ is normal distributed, then $\Delta := [-\text{CI}_{.025}, \text{CI}_{.975}]$, $\text{CI}_{.975d} := Q(0.975|\boldsymbol{\xi}_d)$, where Q is the quantile function built upon marginal distribution $\boldsymbol{\xi}_d$. For all acquisition functions, we use $\beta_t = 2$ as suggested in the original paper.

MO-MVA-BO is originally proposed in the active learning settings (Iwazaki et al., 2021). In order to make use of it in a continuous input space, we discretize the input space $\mathcal{X}$ evenly using a space segmentation parameter $\tau = \{0.001, 0.05, 0.05\}$ for input dimensionality $d = \{1, 2, 3\}$ respectively. We make a post recommendation based on $\hat{\Pi}_t$, which is extracted from the discretized input set $\tilde{\mathcal{X}}$ as the same setting in the origin paper. Nevertheless, we remark the **comparison is not fair** as this recommendation is an out-of-sample strategy: which is potentially better, especially at the beginning as the candidate size can be larger than the in-sample data size.

CO-MVA-BO We follow the same discretization approach as for MO-MVA-BO. For both FF-MV-ECI and CO-MVA-BO acquisition functions, if the feasible candidates set is empty($St = \varnothing$), we use a small pseudo constant value as the current best value keeping the acquisition function operating as defined. We use $\arg\max_{\boldsymbol{x} \in \tilde{\mathcal{X}}} l_t^{F_1}$ for making recommendations.

Uncertainty Sampling For the 2nd stage, instead of performing Eq. 46, we only optimize the standard deviation σ_t. We make optimal recommendations based on in-sample evaluations that are in $\mathcal{X}$.

H.3. Experimental Setup for the Robust Bayesian Optimization

Number of Fourier features Since Fourier feature-based acquisition functions need to specify the number of features explicitly, we use $\{128, 900, 1000\}$ for $\{1d, 2d, 3d\}$ problems, respectively, for both RFF and QFF.

[12] The ground truth robustness measures' means are extracted by doing an exhausted MC on input space with min($2000/500 * 2^d$, 10000/6000) sample size for first moment accuracy measure/uncertainty calibration respectively, the maximum is taken to enforce a feasible Cholesky decomposition on sampling robustness measure. 256 samples are used extracted the mean of robustness (i.e., variance) from GP posterior samples

Recommendations for optimal solutions At the end of the optimization process, we need to provide the optimal candidate for the practitioner. In our problem scenario, both robustness measures are not observable. To extract the optimal solution, one can either perform an in-sample recommendation (i.e., extracting optimal recommendations based on queried data $\boldsymbol{X}$), or perform an out-of-sample recommendation (see, e.g., Daulton et al. (2021); Fröhlich et al. (2020)). Since out-of-sample strategies require performing another optimization process to extract optimal solutions, which makes the computation cost even higher, we focus on the in-sample strategy in this paper. We recommend the optimal candidate based on the inferred robustness measures for the scalarization and multi-objective formulations. For the variance as constraint approach, we use $\boldsymbol{x}^* = \arg\max_{\boldsymbol{x}\in\boldsymbol{F}} \mathbb{J}(\boldsymbol{x})$ for conservative recommendations in case the constraint is expected to be satisfied with high priority (e.g., the Utility Gap metric), where $\boldsymbol{F} := \{\boldsymbol{x} \in \boldsymbol{X} | Pr(\tilde{\mathbb{V}}(\boldsymbol{x}) \leq C_v) \geq p\}$. p can be specified by practitioners and this probability can be approximated by MC sampling of $\tilde{\mathbb{V}}(\boldsymbol{x})$, we use $p = 0.8$ in our numerical experiments. We alternatively recommend $\arg\max_{\boldsymbol{x}\in\boldsymbol{X}} Pr(\tilde{\mathbb{V}}(\boldsymbol{x}) \leq C_v)$ in case $\boldsymbol{F} = \varnothing$.

Reference point setting For hypervolume based multi-objective optimization approaches, a reference point is needed to partition the non-dominated region into hypercubes. In this paper, we follow the strategy by keeping an automatically updated reference point as an agnostic treatment of the Pareto front.

We use a fixed reference point extracted from a reference Pareto front to calculate the log-hypervolume metric. We provide the corresponding reference point setting in Table 5[13] which is calculated as Knudde et al. (2017): $\mathcal{F}_{MV_{ref}} = \min_r(\mathcal{F}^*_{MV}) - 2\left(\max_r(\mathcal{F}^*_{MV}) - \min_r(\mathcal{F}^*_{MV})\right)/|\mathcal{F}^*_{MV}|$, where the operator $\max_r() := [\max_{f_{MV}\in\mathcal{F}^*_{MV}} f^{(1)}_{MV}, \max_{f_{MV}\in\mathcal{F}^*_{MV}} f^{(2)}_{MV}]$, $\min_r() := [\min_{f_{MV}\in\mathcal{F}^*_{MV}} f^{(1)}_{MV}, \min_{f_{MV}\in\mathcal{F}^*_{MV}} f^{(2)}_{MV}]$.

Table 5. Reference Point Setting

Problem	Input Uncertainty $\boldsymbol{\xi}$ distribution	Reference point
SineLinear (Fröhlich et al., 2020)	$\mathcal{N}(0, 0.001)$	$[0.03003507, 0.07371671]$
	$U(-0.05, 0.05)$,	$[0.03320457, 0.02952188]$
Forrester (Keane et al., 2008)	$\mathcal{N}(0, 0.005)$	$[0.15948319, 5.8976241]$
	$U(-0.1, 0.1)$,	$[0.17751173, 1.79019393]$
Branin	$\mathcal{N}(\mathbf{0}, 0.01\boldsymbol{I}_2)$	$[15.77117159, 84.70987466]$
	$U(-[0.1, 0.01], [0.1, 0.01])$,	$[19.4079212, 10.30334723]$
Hartmann3	$\mathcal{N}(\mathbf{0}, 0.01\boldsymbol{I}_3)$	$[0.11863389, 0.46989925]$
	$U(-[0.15, 0.15, 0.15], [0.15, 0.15, 0.15])$,	$[0.12358115, 0.21843887]$
Conceptual Low Drag Wing Design	$\mathcal{N}(\mathbf{0}, \left[\begin{smallmatrix} 0.004 & 0.0 \\ 0.0 & 0.00045125 \end{smallmatrix}\right])$	$[3.10164206, 0.24411262]$
	$U(-[0.08, 0.02], [0.08, 0.02])$	$[3.13006765, 0.07853147]$

H.4. Details of the Conceptual Low Drag Wing Design Problem

We perform RBO on a two-dimensional conceptual low drag wing design problem in level flight conditions. The goal is to find the optimal wing geometry (i.e., wingspan b and chord length c) to achieve a low drag coefficient C_D, which is usually represented by lift-to-drag ratio $\frac{C_L}{C_D}$. In practice, besides the low drag requirement, we also favor a light aircraft, and the wing geometry cannot be too long span-wise (i.e., large b) while too short in chord-wise (i.e., small c), which may impose difficulties for structural design. Moreover, the lift coefficient C_L is preferably not too big, as otherwise it may impose difficulties for airfoil design. Hence, the objective function is defined as:

$$\arg\max_{\{b,c\in\mathcal{X}\}} -log\left[-\left(5\frac{C_L}{C_D} - \max(A-10, 0) - 100 * \max(C_L - 0.9, 0) - 0.4W - 5\right)\right] \tag{47}$$

where $A = \frac{b}{c}$ represents the aspect ratio. The drag coefficient is defined as:

$$C_D = \frac{0.03062702}{S} + kC_f\frac{S_{wet}}{S} + \frac{C_L^2}{\pi Ae} \tag{48}$$

[13]The reference point is specified for performing **minimization** as is the default in Trieste.

where the wing area $S = bc$, friction drag coefficient $C_f = \frac{0.074}{Re^{0.2}}$, Reynolds number $Re = \frac{\rho V c}{\mu}$. For design specified weight W and speed V, the lift coefficient required in steady level flight can be determined by:

$$C_L = \frac{2W}{\rho V^2 S} \tag{49}$$

where $\rho = 1.29$ is the air density. The aircraft weight is $W = W_0 + W_w$, where W_w represents the weight of the wing, which can be determined by:

$$W_w = 45.42S + 8.71 \times 10^{-5} \frac{N_{ult} b^3 \sqrt{W_0 W}}{S(t/c)} \tag{50}$$

The remaining parameters are defined as follows:

Table 6. Parameters for steady level flight

Parameter	Value	Unit	Notes
W_0	18	N	aircraft weight (excluding weight of the wing)
t/c	0.12	NA	average thickness-to-chord ratio
N_{ult}	2.5	NA	ultimate load factor
S_{wet}	2.05S	m^2	wing wetted area
μ	17.8×10^{-6}	kg/(m sec)	air viscosity
k	1.2	NA	form factor
e	0.96	NA	Oswald efficiency number

H.5. Results for Uniform Distributed Input Uncertainty

We provide the additional experiments for uniformly distributed input uncertainty $\boldsymbol{\xi}$ in Fig. 7. For the conceptual wing design problem, the settings for the input uncertainty are provided in Table 5. For the robot pushing problem, input uncertainty $\boldsymbol{\xi}$'s distribution is specified as $U(-[0.2, 0.2, 0.58], [0.2, 0.2, 0.58])$ and $U(-[0.3, 0.3, 0.29], [0.3, 0.3, 0.29])$ for scalarization and variance as constraint formulation, respectively.

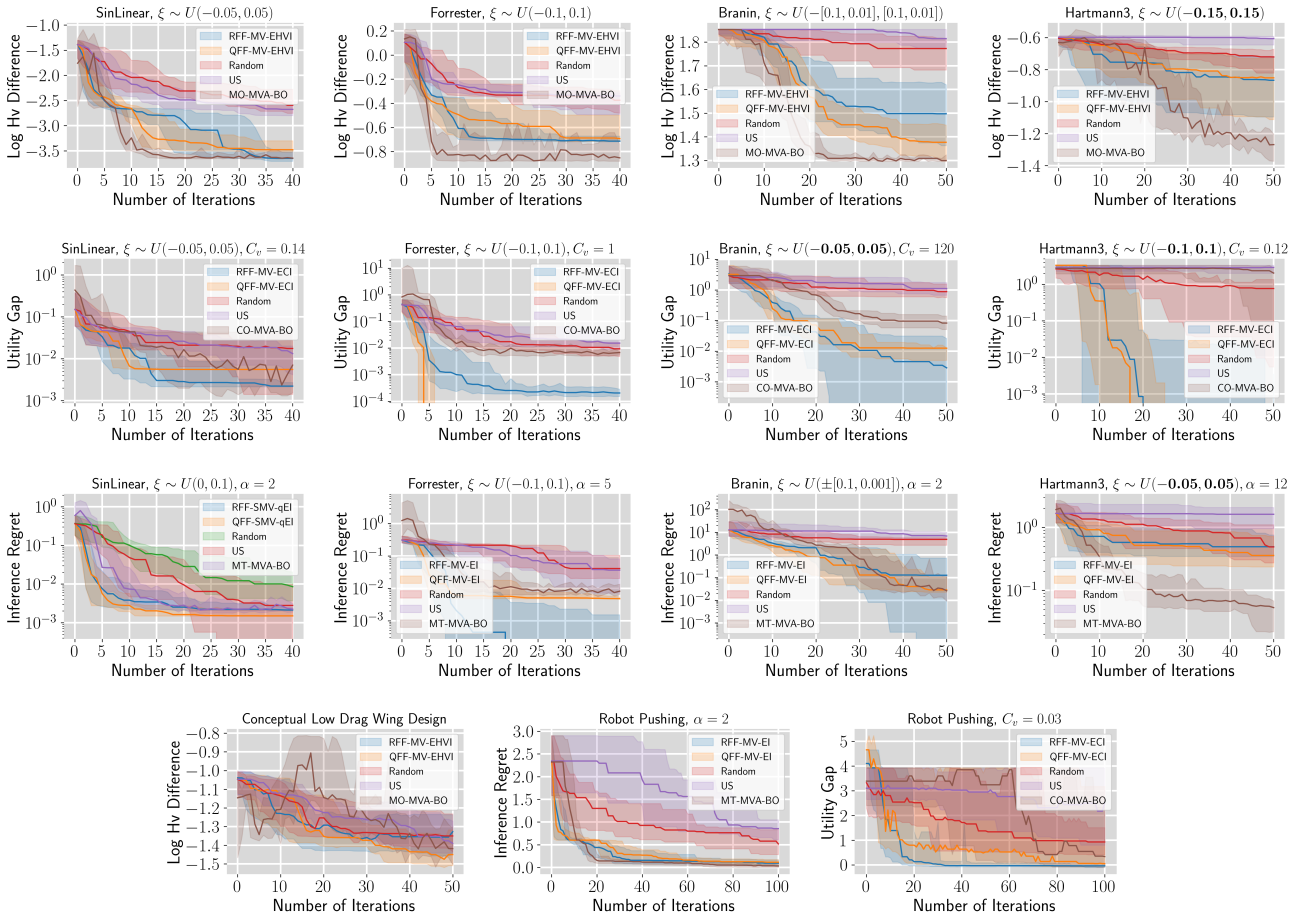

Figure 7. Experiment results on synthetic functions and real-life applications for uniform distributed input uncertainty.