Variational Bayesian surrogate modelling with application to robust design optimisation

Thomas A. Archbold, Ieva Kazlauskaite, Fehmi Cirak*

Department of Engineering, University of Cambridge, Cambridge, CB2 1PZ, UK

Abstract

Surrogate models provide a quick-to-evaluate approximation to complex computational models and are essential for multi-query problems like design optimisation. The inputs of current computational models are usually high-dimensional and uncertain. We consider Bayesian inference for constructing statistical surrogates with input uncertainties and intrinsic dimensionality reduction. The surrogates are trained by fitting to data from prevalent deterministic computational models. The assumed prior probability density of the surrogate is a Gaussian process. We determine the respective posterior probability density and parameters of the posited statistical model using variational Bayes. The non-Gaussian posterior is approximated by a simpler trial density with free variational parameters and the discrepancy between them is measured using the Kullback-Leibler (KL) divergence. We employ the stochastic gradient method to compute the variational parameters and other statistical model parameters by minimising the KL divergence. We demonstrate the accuracy and versatility of the proposed reduced dimension variational Gaussian process (RDVGP) surrogate on illustrative and robust structural optimisation problems with cost functions depending on a weighted sum of the mean and standard deviation of model outputs.

Keywords: Surrogate modelling; Multi-query problems; Bayesian inference; Variational Bayes; Gaussian processes; Robust optimisation

1. Introduction

The industrial design of engineering products relies on complex computational models typically consisting of several submodels responsible for different aspects of design and analysis, like geometric modelling, finite element analysis, etc. Prevailing computational models are costly to evaluate and quickly become a bottleneck in multi-query problems such as design optimisation, uncertainty quantification, inverse problems and model calibration. For surrogate modelling, the computational model is interpreted as a black-box function that can be interrogated by recording its outputs for given inputs. The quick-to-evaluate surrogate model is learned by fitting to data collected from the black-box function. Common surrogate modelling techniques include polynomial chaos expansions [1–4], neural networks [5–8] and Gaussian process (GP) regression [9–13]. In our primary application, robust design optimisation (RDO) [14–18], the model inputs consist of design variables like geometric dimensions and fixed (immutable) variables like forcing or constitutive parameters. The model outputs are quantities of interest like the maximum stress or compliance.

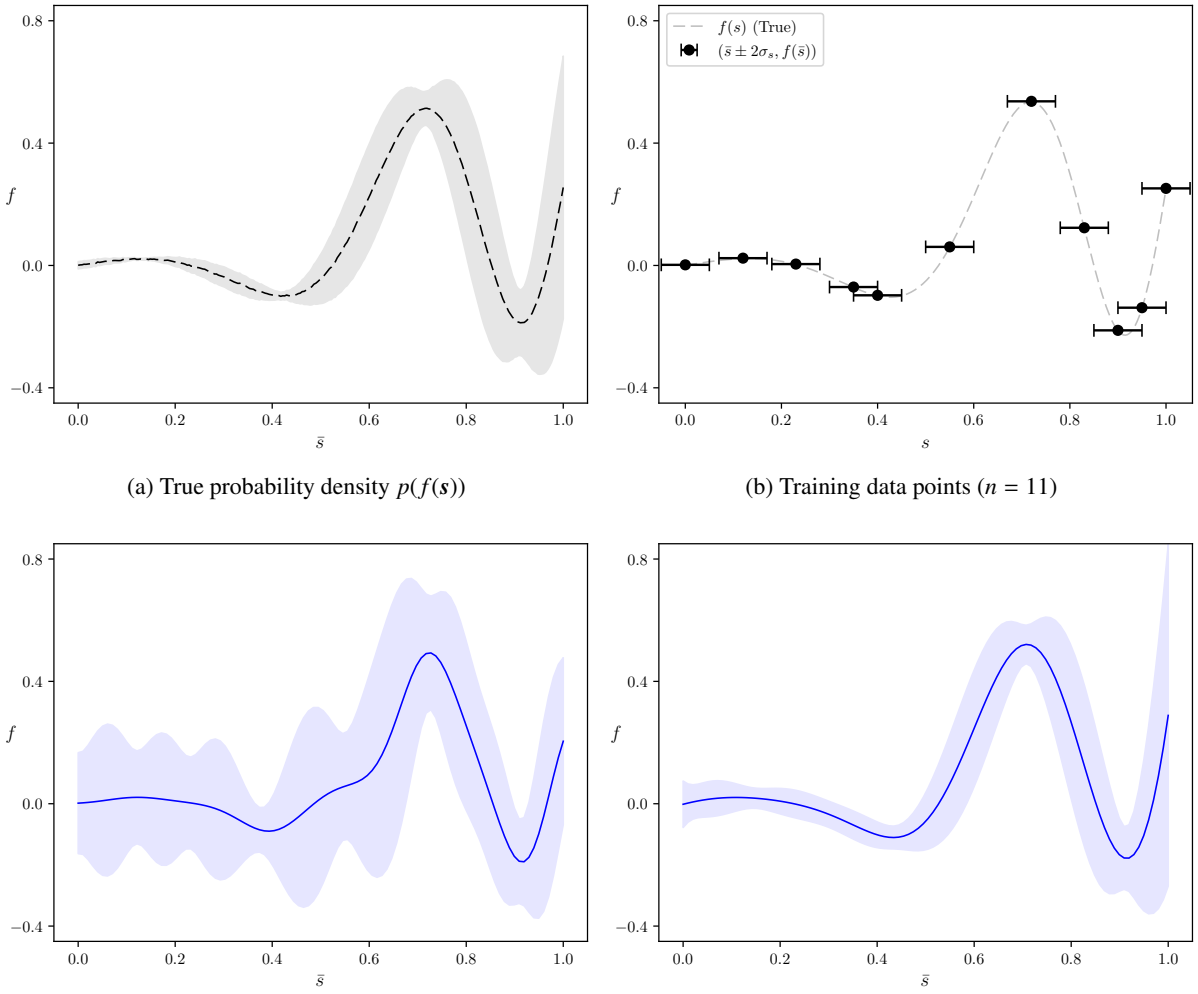
*Corresponding author
Email address: f.cirak@eng.cam.ac.uk (Fehmi Cirak)

Although most prevailing computational models are deterministic, it is critical to consider the input uncertainties (see Figure 1) to achieve more efficient and robust designs [19, 20]. The uncertainties typically arise from variations in manufacturing processes and insufficiently known operational conditions. Such intrinsic aleatoric uncertainties are critical for RDO, necessitating their inclusion in surrogate modelling, see e.g. [21, 22] for a discussion on aleatoric and epistemic uncertainties. Robust optimisation yields a design which is less sensitive to the variations in the inputs by considering a cost function consisting of the weighted sum of the expectation and standard deviation of typical objective and constraint functions used in deterministic optimisation. Surrogate-based techniques are widely used in RDO, see e.g. [23–29].

We construct a statistical surrogate model by fitting to data from the costly-to-evaluate deterministic computational model. As common in probabilistic approaches, most prominently GP regression, we pose surrogate model construction as a Bayesian inference problem [30–34]. That is, the posterior probability density representing a surrogate $f(\boldsymbol{s})$ is determined by updating a prior probability density in light of a training data set $\mathcal{D}$ using the Bayes' rule. The prior probability density for $f(\boldsymbol{s})$ is a GP with a prescribed mean and covariance with several hyperparameters. The training data set $\mathcal{D} = \{(\boldsymbol{s}_i, y_i)\}_{i=1}^n$ collects the n evaluations of the computational model with each pair consisting of an input $\boldsymbol{s}_i \in \mathbb{R}^{d_s}$ and its output $y_i \in \mathbb{R}$. The assumed prior probability density is a GP defining a distribution over functions with a certain mean and covariance (smoothness). The required likelihood density is obtained from the posited statistical observation model $y = f(\boldsymbol{s}) + \epsilon_y$ with an additive noise or error term ϵ_y. A crucial difference of the sketched approach from standard GP regression is that the input vector $\boldsymbol{s}$ has intrinsic aleatoric uncertainties, e.g. due to manufacturing. Consequently, the posterior probability density representing the surrogate $f(\boldsymbol{s})$ encodes in addition to epistemic also intrinsic aleatoric uncertainties that do not vanish in the limit of infinite data. In standard GP regression with a deterministic input $\boldsymbol{s}$, all uncertainty is epistemic and tends toward zero when more data is taken into account [9, 21]. The key challenge in Bayesian inference with an uncertain input $\boldsymbol{s}$ is that the posterior density is not a GP with a closed-form solution (even when both the prior probability density and the likelihood density are Gaussians). Hence, it is necessary to use approximation techniques, like Markov chain Monte Carlo (MCMC), Laplace approximation, or variational Bayes (VB) [32].

In design optimisation, as in most other multi-query applications, the dimension of the input variable $\boldsymbol{s}$ must be restricted for computational tractability. The dimension of the input variable is typically large from the outset, so a suitable dimensionality reduction technique must be employed. Dimensionality reduction is an extensively studied topic, and many techniques are available [35–39]. Two-step approaches for sequentially combining dimensionality reduction with standard GP regression have also been proposed [40–42]. Departing from earlier approaches and adopting a fully Bayesian viewpoint, we implement dimensionality reduction by positing the statistical observation model $y = f(\boldsymbol{z}) + \epsilon_y$ with the low-dimensional (latent) input vector $\boldsymbol{z} = \boldsymbol{W}^\mathsf{T}\boldsymbol{s}$ and the non-square (tall) orthogonal matrix $\boldsymbol{W}$, i.e. $\boldsymbol{W}^\mathsf{T}\boldsymbol{W} = \boldsymbol{I}$. A reduction in input dimension is possible because in most applications the intrinsic dimensionality of the input-output manifold is usually low. We assume here that the solution lies on a hyperplane in the high-dimensional input-output space. Furthermore, as noted before, the input vector $\boldsymbol{s}$ and consequently, the latent vector $\boldsymbol{z}$ have intrinsic aleatoric uncertainties whereas the matrix $\boldsymbol{W}$ is deterministic. The posterior probability density of $f(\boldsymbol{s})$ and all statistical model parameters including the dimension of $\boldsymbol{z}$ and entries of $\boldsymbol{W}$, can be obtained by consequent application of the Bayes' rule. Expectedly, the posterior probability densities and the marginal likelihood density for determining the model parameters have no analytically tractable solution and must be approximated.

VB provides an optimisation-based formulation for approximating the posterior density and parameters of statistical models [43–46]. In VB the true posterior density is approximated by a trial density and the distance between the two is measured using the Kullback-Leibler (KL) divergence. The chosen trial density



(a) True probability density $p(f(\boldsymbol{s}))$

(b) Training data points ($n = 11$)

(c) Posterior density of the standard GP surrogate

(d) Posterior density of the proposed RDVGP surrogate

Figure 1: Comparison of standard GP and the proposed RDVGP surrogates for a function $f(s) = -0.5s\sin(3\pi s^2) + 0.25s$ and the normally distributed input variable with $p(s) = \mathcal{N}(\bar{s}, \sigma_s^2)$; see also Section 4.1. In the plots the lines indicate the mean and the shaded areas the 95% confidence intervals. Shown is the (a) true probability density $p(f(\bar{s})) = \int p(f|\boldsymbol{s})p(\boldsymbol{s})\,d\boldsymbol{s}$ obtained by MC sampling, (b) the $n = 11$ training data points comprising the training set $\mathcal{D}$, (c) the inferred GP posterior probability density $p(f(\bar{s})|\mathcal{D})$, and (d) the inferred RDVGP posterior probability density $p(f(\bar{s})|\mathcal{D})$. Comparing (a), (c), and (d) it is evident that the RDVGP posterior is much closer to the true probability density than the standard GP posterior.

is a parameterised probability density with free variational parameters. It can be shown that minimising the KL divergence is equivalent to maximising the evidence lower bound (ELBO). As the name suggests, the ELBO is a lower bound for the log-evidence also called the log-marginal likelihood, which quantifies how well the posited statistical observation model explains the training data. In our statistical surrogate model described above, the approximated density is the posterior density of the surrogate $f(\boldsymbol{s})$ and the model hyperparameters consist of the entries of the projection matrix $\boldsymbol{W}$ and the hyperparameters of the prior. We select as the trial density a Gaussian with a diagonal covariance structure and choose its mean and variance as the variational parameters. All the model hyperparameters and the variational parameters are determined by maximising the ELBO. In the proposed implementation, we maximise the ELBO using a

stochastic gradient method [47], specifically the Cayley ADAM optimiser [48], which automatically ensures the orthogonality of $\boldsymbol{W}$. Unfortunately, the evaluation of the ELBO has the computational complexity $O(n^3)$, where n is the size of the training data set $\mathcal{D}$. Therefore, we introduce a sparse approximation of $f(\boldsymbol{z})$ using m pseudo training data points reducing complexity to $O(nm^2)$ [49, 50]. Although the number m is user prescribed, the coordinates of the pseudo input points are determined as part of VB. The overall solution approach implemented is akin to the latent variable GP models widely used in machine learning [51–53]. For brevity, we refer to this as the reduced dimension variational Gaussian process (RDVGP) surrogate.

This paper is structured as follows. In Section 2, we define RDO problems in terms of objective and constraint functions describing the behavior of complex engineering systems, and review GP regression for emulating these functions. In Section 3, we present the statistical surrogate model by postulating a graphical model with intrinsic statistical assumptions and deriving the ELBO from the joint probability density using VB, later extending to include sparsity. We detail the algorithm used to train the RDVGP surrogate, and provide the metrics used to verify its accuracy when compared to the true solution estimated using Monte Carlo (MC) sampling. Four examples are introduced in Section 4 to demonstrate the accuracy and versatility of the predicted marginal posterior probability densities for solving RDO problems. Finally, Section 5 concludes the paper and provides promising directions for further research.

2. Robust design optimisation and Gaussian processes

In this section, we review RDO and GP regression. We introduce the problem in terms of inputs with uncertainties as set out by prescribed probability densities and we show how the problem is formulated in terms of objective and constraint functions from the computational model.

2.1. Review of robust design optimisation

The input variables $\boldsymbol{s} \in \mathbb{R}^{d_s}$ of deterministic computational models may be composed of immutable variables $\boldsymbol{s}_f \in \mathbb{R}^{d_f}$ that cannot be varied (such as the partial differential equation (PDE) source terms), and design variables $\boldsymbol{s}_d \in \mathbb{R}^{d_d}$ (such as CAD geometry parameters) such that $\boldsymbol{s} = (\boldsymbol{s}_d^\mathsf{T}\ \boldsymbol{s}_f^\mathsf{T})^\mathsf{T}$.

In many applications, the computational model output is not directly the quantity of interest, but an argument provided to an objective or constraint function representing model behavior, and may be computed from a finite element (FE) model. For example, in structural mechanics, compliance objectives and stress constraints are both functions of the PDE solution. For brevity, the objective $J : \mathbb{R}^{d_s} \to \mathbb{R}$ and constraint $H^{(j)} : \mathbb{R}^{d_s} \to \mathbb{R}$ may be denoted as functions of the input variables $\boldsymbol{s}$. A total of d_y functions, consisting (without loss of generality) of a single objective and $(d_y - 1)$ constraint functions are considered.

This paper assumes the input variable probability density is prescribed a priori, which could be Gaussian $p(\boldsymbol{s}) = \mathcal{N}(\bar{\boldsymbol{s}}, \boldsymbol{\Sigma}_s)$, where $\bar{\boldsymbol{s}} \in \mathbb{R}^{d_s}$ is the mean and $\boldsymbol{\Sigma}_s \in \mathbb{R}^{d_s \times d_s}$ is the covariance matrix. For example, a CAD embodiment of engineering products may consist of components with geometric variation represented by a Gaussian probability density, where an engineer models the product in CAD using the mean design variables $\bar{\boldsymbol{s}}$. Since the input variables are random, the objective $J(\boldsymbol{s})$ and constraint $H^{(j)}(\boldsymbol{s})$ functions are also random and the RDO problem may be solved using the weighted sum cost function

$$
\begin{aligned}
\min_{\bar{\boldsymbol{s}} \in D_{\bar{\boldsymbol{s}}}} \quad & \frac{(1-\alpha)}{\bar{\mu}} \mathbb{E}\left(J(\boldsymbol{s})\right) + \frac{\alpha}{\bar{\sigma}} \sqrt{\mathrm{var}\left(J(\boldsymbol{s})\right)}, \\
\text{s.t.} \quad & \mathbb{E}\left(H^{(j)}(\boldsymbol{s})\right) + \beta^{(j)} \sqrt{\mathrm{var}\left(H^{(j)}(\boldsymbol{s})\right)} \leq 0, \qquad j \in \{1, 2, ..., d_y - 1\}, \\
& \sqrt{\mathrm{var}\left(H^{(j)}(\boldsymbol{s})\right)} \leq \sigma^{(j)}, \\
& D_{\bar{\boldsymbol{s}}} = \left\{ \bar{\boldsymbol{s}} \in \mathbb{R}^{d_s} \,\middle|\, \bar{\boldsymbol{s}}^{(l)} \leq \bar{\boldsymbol{s}} \leq \bar{\boldsymbol{s}}^{(u)} \right\},
\end{aligned}
\tag{1}
$$

where $\alpha \in \mathbb{R}^+$ is a prescribed weighting factor ($0 \leq \alpha \leq 1$), $\bar{\mu} \in \mathbb{R}$ and $\bar{\sigma} \in \mathbb{R}$ are normalisation constants, $\beta^{(j)} \in \mathbb{R}^+$ is a prescribed feasibility index, and $\sigma^{(j)} \in \mathbb{R}^+$ is an upper limit on the standard deviation of the j^{th} constraint. The mean input variables $\bar{\boldsymbol{s}}$ are bounded between lower and upper limits $\bar{\boldsymbol{s}}^{(l)} = ((\bar{\boldsymbol{s}}_d^{(l)})^\mathsf{T}\ \bar{\boldsymbol{s}}_f^\mathsf{T})^\mathsf{T}$ and $\bar{\boldsymbol{s}}^{(u)} = ((\bar{\boldsymbol{s}}_d^{(u)})^\mathsf{T}\ \bar{\boldsymbol{s}}_f^\mathsf{T})^\mathsf{T}$ respectively. The objective of RDO is to optimise the mean design variables $\bar{\boldsymbol{s}}_d$.

2.2. Review of Gaussian processes

In RDO, the objective $J(\boldsymbol{s})$ and constraint $H^{(j)}(\boldsymbol{s})$ functions are emulated by a GP surrogate denoted $f(\boldsymbol{s})$. A single observation from each function denoted $y \in \mathbb{R}$, follows a statistical observation model, see the graphical model in Figure 2,

$$y = f(\boldsymbol{s}) + \epsilon_y, \qquad \epsilon_y \sim \mathcal{N}(0, \sigma_y^2), \tag{2}$$

where $\sigma_y \in \mathbb{R}^+$ denotes the noise or error standard deviation (such as FE discretisation error), and the GP $f : \mathbb{R}^{d_s} \to \mathbb{R}$ maps between the input variables $\boldsymbol{s}$ and observation y, and has the prior

$$f(\boldsymbol{s}) \sim \mathcal{GP}\left(\bar{f}(\boldsymbol{s})\,,\, c(\boldsymbol{s}, \boldsymbol{s}')\right). \tag{3}$$

Without loss of generality, we choose a zero-mean prior $\bar{f}(\boldsymbol{s}) = 0$ and squared exponential covariance function $c : \mathbb{R}^{d_s} \times \mathbb{R}^{d_s} \to \mathbb{R}$ (other types of kernels may be used when the objective is less smooth),

$$c(\boldsymbol{s}, \boldsymbol{s}') = \sigma_f^2 \exp\left(-\sum_{i=1}^{d_s} \frac{(s_i - s_i')^2}{2\ell_i}\right), \tag{4}$$

where $\sigma_f \in \mathbb{R}^+$ is a scaling parameter, $\ell_i \in \mathbb{R}^+$ is a length scale parameter, and s_i denotes the i^{th} component of input vector $\boldsymbol{s}$. The hyperparameters of the GP are summarised as $\boldsymbol{\theta} = \{\sigma_f, \ell_1, \ell_2, ..., \ell_{d_s}\}$.

Let $\mathcal{D} = \{(\boldsymbol{s}_i, y_i)\,|\,i \in \mathbb{N}, i \leq n\}$ denote the training data set, which is collected into the input variable matrix $\boldsymbol{S} = (\boldsymbol{s}_1\ \boldsymbol{s}_2\ ...\ \boldsymbol{s}_n)^\mathsf{T} \in \mathbb{R}^{n \times d_s}$ and observation vector $\boldsymbol{y} = (y_1\ y_2\ ...\ y_n)^\mathsf{T} \in \mathbb{R}^n$. The joint probability density of observations $\boldsymbol{y}$ and target output variables $\boldsymbol{f} = (f_1\ f_2\ ...\ f_n)^\mathsf{T} \in \mathbb{R}^n$ can be factorised as

$$p_\Theta(\boldsymbol{y}, \boldsymbol{f}) = p_{\sigma_y}(\boldsymbol{y}|\boldsymbol{f}) p_\theta(\boldsymbol{f}), \tag{5}$$

which is expanded using the postulated conditional independence with the likelihood (observations are assumed to be independent and identically distributed (iid))

$$p_{\sigma_y}(\boldsymbol{y}|\boldsymbol{f}) = \prod_{i=1}^{n} p_{\sigma_y}(y_i|f_i) = \mathcal{N}\left(\boldsymbol{f}, \sigma_y^2 \boldsymbol{I}\right), \tag{6}$$

where y_i and f_i denote the observation and target output variable for the i^{th} training data sample, and the GP prior probability density

$$p_\theta(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{C}_{SS}). \tag{7}$$

The entries of the covariance $\boldsymbol{C}_{SS} \in \mathbb{R}^{n \times n}$ consist of the covariance function $c(\boldsymbol{s}, \boldsymbol{s}')$ where $\boldsymbol{s}$ and $\boldsymbol{s}'$ are rows of $\boldsymbol{S}$. The marginal likelihood

$$p_\Theta(\boldsymbol{y}) = \int p_{\sigma_y}(\boldsymbol{y}|\boldsymbol{f}) p_\theta(\boldsymbol{f})\, d\boldsymbol{f} = \mathcal{N}\left(\boldsymbol{0}\,,\, \boldsymbol{C}_{SS} + \sigma_y^2 \boldsymbol{I}\right), \tag{8}$$

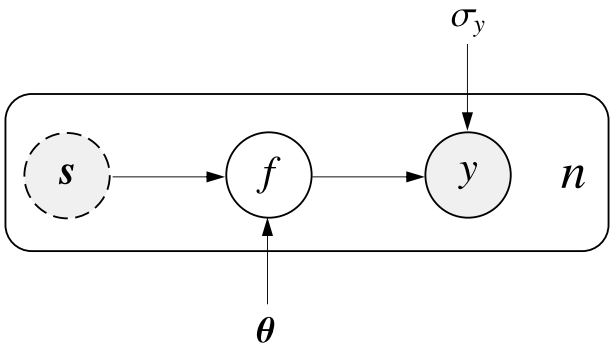


Figure 2: Graphical model of a standard GP where n is the size of the training data set $\mathcal{D}$, $\boldsymbol{s}$ is the vector of (deterministic) input variables, f is the (random) target output variable, y is the (random) noisy observation, and $\boldsymbol{\theta}$ and σ_y are model hyperparameters.

is derived by marginalising over the target output variables $\boldsymbol{f}$. In empirical Bayes' methods, the optimum hyperparameters $\boldsymbol{\Theta} = \boldsymbol{\theta} \cup \{\sigma_y\}$ are computed by maximising the log marginal likelihood

$$\boldsymbol{\Theta}^* = \arg\max_{\Theta} \ln p_{\Theta}(\boldsymbol{y}), \tag{9}$$

where

$$\ln p_{\Theta}(\boldsymbol{y}) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln\left|\boldsymbol{C}_{SS} + \sigma_y^2\boldsymbol{I}\right| - \frac{1}{2}\boldsymbol{y}^{\mathsf{T}}\left(\boldsymbol{C}_{SS} + \sigma_y^2\boldsymbol{I}\right)^{-1}\boldsymbol{y}, \tag{10}$$

and $|\cdot|$ denotes the determinant. For any new test input variables $\boldsymbol{S}_* \in \mathbb{R}^{n_* \times d_s}$, the posterior probability density of the test output variables $\boldsymbol{f}_* \in \mathbb{R}^{n_*}$ is given by

$$p_{\Theta}(\boldsymbol{f}_*|\boldsymbol{y}) = \int p_{\theta}(\boldsymbol{f}_*|\boldsymbol{f}) p_{\Theta}(\boldsymbol{f}|\boldsymbol{y})\, d\boldsymbol{f}, \tag{11}$$

which assumes conditional independence between the test output variables $\boldsymbol{f}_*$ and target output variables $\boldsymbol{f}$ when observations $\boldsymbol{y}$ are given. The posterior probability densities $p_{\theta}(\boldsymbol{f}_*|\boldsymbol{f})$ and $p_{\Theta}(\boldsymbol{f}|\boldsymbol{y})$ can be determined using Bayes rule

$$p_{\theta}(\boldsymbol{f}_*|\boldsymbol{f}) = \frac{p_{\theta}(\boldsymbol{f}|\boldsymbol{f}_*)p_{\theta}(\boldsymbol{f}_*)}{p_{\theta}(\boldsymbol{f})}, \qquad p_{\Theta}(\boldsymbol{f}|\boldsymbol{y}) = \frac{p_{\sigma_y}(\boldsymbol{y}|\boldsymbol{f})p_{\theta}(\boldsymbol{f})}{p_{\Theta}(\boldsymbol{y})}. \tag{12}$$

Alternatively because all involved probability densities are Gaussian, the posterior probability density $p_{\Theta}(\boldsymbol{f}_*|\boldsymbol{y})$ is derived using the joint probability density

$$p_{\Theta}(\boldsymbol{f}_*, \boldsymbol{y}) = \mathcal{N}\left(\begin{pmatrix}\boldsymbol{0}\\ \boldsymbol{0}\end{pmatrix}, \begin{pmatrix}\boldsymbol{C}_{S_*S_*} & \boldsymbol{C}_{S_*S}\\ \boldsymbol{C}_{SS_*} & \boldsymbol{C}_{SS} + \sigma_y^2\boldsymbol{I}\end{pmatrix}\right). \tag{13}$$

The posterior probability density of test output variables $\boldsymbol{f}_*$ is given by

$$p_{\Theta}(\boldsymbol{f}_*|\boldsymbol{y}) = \mathcal{N}\left(\boldsymbol{C}_{S_*S}\left(\boldsymbol{C}_{SS} + \sigma_y^2\boldsymbol{I}\right)^{-1}\boldsymbol{y}\,,\, \boldsymbol{C}_{S_*S_*} - \boldsymbol{C}_{S_*S}\left(\boldsymbol{C}_{SS} + \sigma_y^2\boldsymbol{I}\right)^{-1}\boldsymbol{C}_{SS_*}\right), \tag{14}$$

which is derived by conditioning the joint probability density $p_{\Theta}(\boldsymbol{f}_*, \boldsymbol{y})$ on observations $\boldsymbol{y}$. The evaluation of the posterior probability density becomes computationally expensive for large training data sets due to

the inversion of the covariance matrix $\boldsymbol{C}_{SS}$, which involves $O(n^3)$ operations.

So far the input variables $\boldsymbol{S}$ have been assumed deterministic; however, if they are random and sampled from a prescribed probability density $p(\boldsymbol{S})$, the posterior probability density of the target output variables

$$p_\Theta(\boldsymbol{f}|\boldsymbol{y}) = \frac{p_{\sigma_y}(\boldsymbol{y}|\boldsymbol{f}) \int p_\theta(\boldsymbol{f}|\boldsymbol{S}) p(\boldsymbol{S})\, d\boldsymbol{S}}{p_\Theta(\boldsymbol{y})}, \tag{15}$$

becomes analytically intractable because the covariance $\boldsymbol{C}_{SS}$ of the GP prior $p_\theta(\boldsymbol{f}|\boldsymbol{S})$ depends non-linearly according to (4), on the input variables $\boldsymbol{S}$.

3. Statistical surrogate model

In this section, we generalise standard GP regression to random input variables with intrinsic dimensionality reduction using a latent variable formulation. This is later extended to a sparse formulation of the latent variables, which reduces the computational expense of performing inference with input variable uncertainty. Finally, an algorithm for training the model and metrics for measuring accuracy are provided.

3.1. Reduced dimension variational Gaussian process

We introduce a low-dimensional latent variable representation for each input variable $\boldsymbol{s}$, which is computed using the statistical observation model given by

$$\boldsymbol{z} = \boldsymbol{W}^\mathsf{T}\boldsymbol{s}, \qquad \boldsymbol{s} \sim \mathcal{N}(\bar{\boldsymbol{s}}, \boldsymbol{\Sigma}_s), \tag{16a}$$

$$y = f(\boldsymbol{z}) + \epsilon_y, \qquad \epsilon_y \sim \mathcal{N}(0, \sigma_y^2). \tag{16b}$$

The low-dimensional latent variables $\boldsymbol{z} \in \mathbb{R}^{d_z}$, $(d_z < d_s)$ are mapped by a GP surrogate $f(\boldsymbol{z})$ to an observation y following the posited statistical observation model. The linear projection matrix $\boldsymbol{W} \in \mathbb{R}^{d_s \times d_z}$ is composed of basis vectors of the low-dimensional subspace, and the mean $\bar{\boldsymbol{s}}$ and covariance $\boldsymbol{\Sigma}_s$ of the input variable probability density $p(\boldsymbol{s})$ are prescribed a priori. For a training data set $\mathcal{D}$ of size n, the joint probability density of the observations $\boldsymbol{y}$, target output variables $\boldsymbol{f}$, latent variables $\boldsymbol{Z} = (\boldsymbol{z}_1\ \boldsymbol{z}_2\ \ldots\ \boldsymbol{z}_n)^\mathsf{T} \in \mathbb{R}^{n \times d_z}$, and input variables $\boldsymbol{S}$, is factorised using the postulated conditional independence, see Figure 3, such that

$$p_\Theta(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{Z}, \boldsymbol{S}) = p_{\sigma_y}(\boldsymbol{y}|\boldsymbol{f}) p_\theta(\boldsymbol{f}|\boldsymbol{Z}) p_W(\boldsymbol{Z}|\boldsymbol{S}) p(\boldsymbol{S}). \tag{17}$$

The likelihood $p_{\sigma_y}(\boldsymbol{y}|\boldsymbol{f})$ is the same as in GP regression (6). The zero-mean GP prior probability density

$$p_\theta(\boldsymbol{f}|\boldsymbol{Z}) = \mathcal{N}(\boldsymbol{0}, \boldsymbol{C}_{ZZ}), \tag{18}$$

has the covariance matrix $\boldsymbol{C}_{ZZ} \in \mathbb{R}^{n \times n}$ with the entries $c(\boldsymbol{z}, \boldsymbol{z}')$ as defined in (4), where $\boldsymbol{z}$ and $\boldsymbol{z}'$ are rows of $\boldsymbol{Z}$. The conditional probability density

$$p_W(\boldsymbol{Z}|\boldsymbol{S}) = \prod_{i=1}^{n} \delta\left(\boldsymbol{z}_i - \boldsymbol{W}^\mathsf{T}\boldsymbol{s}_i\right), \tag{19}$$

readily follows from (16a), where $\delta(\cdot)$ is the Dirac delta function. The input variable probability density

$$p(\boldsymbol{S}) = \prod_{i=1}^{n} \mathcal{N}(\bar{\boldsymbol{s}}_i, \boldsymbol{\Sigma}_s), \tag{20}$$

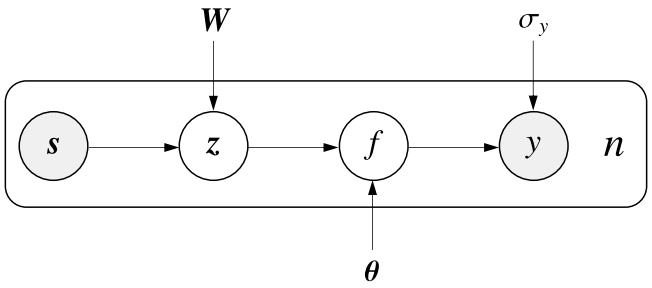


Figure 3: Graphical model of the RDVGP surrogate where n is the size of the training data set $\mathcal{D}$, $\boldsymbol{s}$ is the random input variable vector, $\boldsymbol{z}$ is the low-dimensional unobserved latent variable vector, f is the target output variable, y is the noisy observations, and $\boldsymbol{W}$, $\boldsymbol{\theta}$ and σ_y are model hyperparameters.

is prescribed a priori. We assume the input variables are uncorrelated and the covariance $\boldsymbol{\Sigma}_s$ does not depend on the sample index i. The input variables $\boldsymbol{S}$ are marginalised out of the joint probability density, yielding

$$p_\Theta(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{Z}) = p_{\sigma_y}(\boldsymbol{y}|\boldsymbol{f}) p_\theta(\boldsymbol{f}|\boldsymbol{Z}) p_W(\boldsymbol{Z}), \tag{21}$$

where

$$p_W(\boldsymbol{Z}) = \int p_W(\boldsymbol{Z}|\boldsymbol{S}) p(\boldsymbol{S})\, d\boldsymbol{S} = \prod_{i=1}^{n} \mathcal{N}\left(\boldsymbol{W}^\mathsf{T} \bar{\boldsymbol{s}}_i\,,\, \boldsymbol{W}^\mathsf{T} \boldsymbol{\Sigma}_s \boldsymbol{W}\right). \tag{22}$$

The model hyperparameters $\boldsymbol{\Theta} = \boldsymbol{\theta} \cup \{\boldsymbol{W}, \sigma_y\}$ include in addition to the hyperparameters of the GP prior $\boldsymbol{\theta}$, the observation noise or error standard deviation σ_y and entries of the linear projection matrix $\boldsymbol{W}$. Although the posterior probability density of all unobserved variables

$$p_\Theta(\boldsymbol{f}, \boldsymbol{Z}|\boldsymbol{y}) = \frac{p_\Theta(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{Z})}{p_\Theta(\boldsymbol{y})}, \tag{23}$$

follows from Bayes' rule, it is not available in closed-form without sampling.

In VB, the posterior is approximated with a trial probability density $q_{\theta,\psi}(\boldsymbol{f}, \boldsymbol{Z})$ where the KL divergence between the two probability densities is given by

$$\begin{aligned} D_{KL}\left(q_{\theta,\psi}(\boldsymbol{f}, \boldsymbol{Z}) \,\|\, p_\Theta(\boldsymbol{f}, \boldsymbol{Z}|\boldsymbol{y})\right) &= \int q_{\theta,\psi}(\boldsymbol{f}, \boldsymbol{Z}) \ln\left(\frac{q_{\theta,\psi}(\boldsymbol{f}, \boldsymbol{Z})}{p_\Theta(\boldsymbol{f}, \boldsymbol{Z}|\boldsymbol{y})}\right) d\boldsymbol{Z} d\boldsymbol{f} \\ &= \int q_{\theta,\psi}(\boldsymbol{f}, \boldsymbol{Z}) \ln\left(\frac{q_{\theta,\psi}(\boldsymbol{f}, \boldsymbol{Z}) p_\Theta(\boldsymbol{y})}{p_\Theta(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{Z})}\right) d\boldsymbol{Z} d\boldsymbol{f} \\ &= \int q_{\theta,\psi}(\boldsymbol{f}, \boldsymbol{Z}) \ln\left(\frac{q_{\theta,\psi}(\boldsymbol{f}, \boldsymbol{Z})}{p_\Theta(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{Z})}\right) d\boldsymbol{Z} d\boldsymbol{f} + \ln p_\Theta(\boldsymbol{y}) \\ &= -\mathcal{F}(\boldsymbol{y}) + \ln p_\Theta(\boldsymbol{y}). \end{aligned} \tag{24}$$

Since KL divergence is non-negative $D_{KL}(\cdot\|\cdot) \geq 0$,

$$\ln p_\Theta(\boldsymbol{y}) \geq \int q_{\theta,\psi}(\boldsymbol{f}, \boldsymbol{Z}) \ln\left(\frac{p_\Theta(\boldsymbol{y}, \boldsymbol{f}, \boldsymbol{Z})}{q_{\theta,\psi}(\boldsymbol{f}, \boldsymbol{Z})}\right) d\boldsymbol{Z} d\boldsymbol{f}. \tag{25}$$

The KL divergence is minimised by maximising the ELBO $\mathcal{F}(\boldsymbol{y})$. We assume a trial probability density

$$q_{\theta,\psi}(\boldsymbol{f},\boldsymbol{Z}) = p_\theta(\boldsymbol{f}|\boldsymbol{Z})q_\psi(\boldsymbol{Z}), \tag{26}$$

with the Gaussian prior probability density $p_\theta(\boldsymbol{f}|\boldsymbol{Z})$ given by (18), and the trial probability density

$$q_\psi(\boldsymbol{Z}) := \prod_{i=1}^{n} q_\psi(\boldsymbol{z}_i) := \prod_{i=1}^{n} \mathcal{N}\left(\tilde{\boldsymbol{\mu}}_{z,i}\,,\,\tilde{\boldsymbol{\Sigma}}_{z,i}\right). \tag{27}$$

Variational parameters $\boldsymbol{\psi} = \left\{(\tilde{\boldsymbol{\mu}}_{z,i}, \tilde{\boldsymbol{\Sigma}}_{z,i}) \,\middle|\, i \in \mathbb{N}, i \leq n\right\}$ consist of the components of the mean vector $\tilde{\boldsymbol{\mu}}_{z,i} \in \mathbb{R}^{d_z}$ and the entries of the diagonal covariance matrix $\tilde{\boldsymbol{\Sigma}}_{z,i} \in \mathbb{R}^{d_z \times d_z}$. The ELBO is expanded by introducing the factorised joint probability density $p_\Theta(\boldsymbol{y},\boldsymbol{f},\boldsymbol{Z})$ that follows from (21). The trial probability density $q_{\theta,\psi}(\boldsymbol{f},\boldsymbol{Z})$ such that

$$\begin{aligned}
\mathcal{F}(\boldsymbol{y}) &= \int p_\theta(\boldsymbol{f}|\boldsymbol{Z})q_\psi(\boldsymbol{Z}) \ln\left(\frac{p_{\sigma_y}(\boldsymbol{y}|\boldsymbol{f})p_\theta(\boldsymbol{f}|\boldsymbol{Z})p_W(\boldsymbol{Z})}{p_\theta(\boldsymbol{f}|\boldsymbol{Z})q_\psi(\boldsymbol{Z})}\right) d\boldsymbol{Z}d\boldsymbol{f} \\
&= \int p_\theta(\boldsymbol{f}|\boldsymbol{Z})q_\psi(\boldsymbol{Z}) \left(\ln p_{\sigma_y}(\boldsymbol{y}|\boldsymbol{f}) + \ln\left(\frac{p_W(\boldsymbol{Z})}{q_\psi(\boldsymbol{Z})}\right)\right) d\boldsymbol{Z}d\boldsymbol{f} \\
&= \mathbb{E}_{q_\psi(\boldsymbol{Z})}\left(\mathbb{E}_{p_\theta(\boldsymbol{f}|\boldsymbol{Z})}\left(\ln p_{\sigma_y}(\boldsymbol{y}|\boldsymbol{f})\right)\right) - D_{KL}\left(q_\psi(\boldsymbol{Z}) \,\|\, p_W(\boldsymbol{Z})\right).
\end{aligned} \tag{28}$$

The KL divergence term in the ELBO is analytically tractable given that $q_\psi(\boldsymbol{Z})$ and $p_W(\boldsymbol{Z})$ belong to the exponential family of probability densities (Appendix A). The expectations can be evaluated by MC sampling with $\boldsymbol{Z} \sim q_\psi(\boldsymbol{Z})$ and $\boldsymbol{f} \sim p_\theta(\boldsymbol{f}|\boldsymbol{Z})$. This is computationally expensive for large sample sizes (when n is large) since sampling of $\boldsymbol{f}$ involves $O(n^3)$ operations. The ELBO $\mathcal{F}(\boldsymbol{y})$ can be extended to multiple observation vectors using the mean field approximation (Appendix B).

3.2. Sparse formulation

To reduce the computational expense while retaining the properties of the RDVGP surrogate, the training data set $\mathcal{D}$ is augmented with m (where $m \ll n$) pseudo output variables $\tilde{\boldsymbol{f}} = (\tilde{f}_1\ \tilde{f}_2\ \ldots\ \tilde{f}_m)^\mathsf{T} \in \mathbb{R}^m$. The pseudo output variables are trainable model parameters that when sampled at the pseudo latent variables $\tilde{\boldsymbol{Z}} = (\tilde{\boldsymbol{z}}_1\ \tilde{\boldsymbol{z}}_2\ \ldots\ \tilde{\boldsymbol{z}}_m)^\mathsf{T} \in \mathbb{R}^{m \times d_z}$, accurately explain the observations, see e.g. [50]. The pseudo latent variables also called inducing points, are also considered as hyperparameters. The joint probability density $p_\Theta(\boldsymbol{y},\boldsymbol{f},\boldsymbol{Z})$ given by (21) is augmented with the pseudo output variables $\tilde{\boldsymbol{f}}$ such that

$$p_\Theta\left(\boldsymbol{y},\boldsymbol{f},\tilde{\boldsymbol{f}},\boldsymbol{Z}\right) = p_{\sigma_y}(\boldsymbol{y}|\boldsymbol{f})p_\theta\left(\boldsymbol{f}\,|\,\tilde{\boldsymbol{f}},\boldsymbol{Z}\right)p_\theta\left(\tilde{\boldsymbol{f}}\right)p_W(\boldsymbol{Z}), \tag{29}$$

using the assumed conditional independence structure (Figure 4). The new GP prior probability density

$$p_\theta\left(\boldsymbol{f}\,|\,\tilde{\boldsymbol{f}},\boldsymbol{Z}\right) = \mathcal{N}\left(\boldsymbol{C}_{Z\tilde{Z}}\boldsymbol{C}_{\tilde{Z}\tilde{Z}}^{-1}\tilde{\boldsymbol{f}}\,,\,\boldsymbol{C}_{ZZ} - \boldsymbol{C}_{Z\tilde{Z}}\boldsymbol{C}_{\tilde{Z}\tilde{Z}}^{-1}\boldsymbol{C}_{\tilde{Z}Z}\right), \tag{30}$$

is derived by conditioning the joint GP prior probability density $p_\theta(\boldsymbol{f},\tilde{\boldsymbol{f}}\,|\,\boldsymbol{Z}) = p_\theta(\boldsymbol{f}\,|\,\boldsymbol{Z})p_\theta(\tilde{\boldsymbol{f}})$ on the pseudo output variables $\tilde{\boldsymbol{f}}$, where

$$p_\theta(\boldsymbol{f},\tilde{\boldsymbol{f}}\,|\,\boldsymbol{Z}) = \mathcal{N}\left(\begin{pmatrix}\boldsymbol{0}\\ \boldsymbol{0}\end{pmatrix}, \begin{pmatrix}\boldsymbol{C}_{ZZ} & \boldsymbol{C}_{Z\tilde{Z}}\\ \boldsymbol{C}_{\tilde{Z}Z} & \boldsymbol{C}_{\tilde{Z}\tilde{Z}}\end{pmatrix}\right). \tag{31}$$

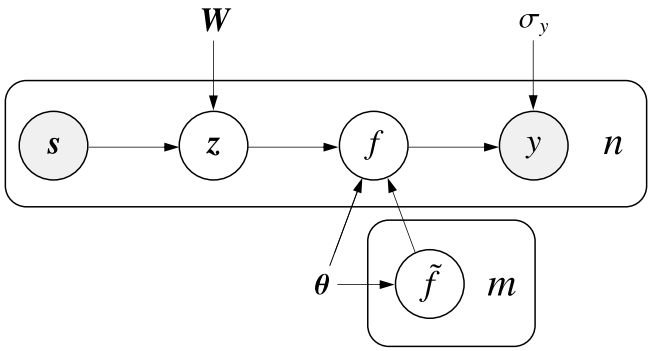


Figure 4: Graphical model of the sparse RDVGP surrogate where n is the size of the training data set $\mathcal{D}$, $\boldsymbol{s}$ is the random input variable vector, $\boldsymbol{z}$ is the low-dimensional unobserved latent variable vector, f is the target output variable, $\tilde{f}$ is the pseudo output variable with m realisations, y is the noisy observation, and $\boldsymbol{W}$, $\boldsymbol{\theta}$ (which includes pseudo latent variables $\tilde{\boldsymbol{Z}}$) and σ_y are model hyperparameters.

This is similar to the derivation of standard GP regression, where $\boldsymbol{\theta} = \left\{\tilde{\boldsymbol{Z}}, \sigma_f, \ell_1, \ell_2, ..., \ell_{d_z}\right\}$ are hyperparameters of $p_\theta(\boldsymbol{f}, \tilde{\boldsymbol{f}} \,|\, \boldsymbol{Z})$. The covariance $\boldsymbol{C}_{\tilde{Z}\tilde{Z}} \in \mathbb{R}^{m\times m}$ consists of the entries $c(\tilde{\boldsymbol{z}}, \tilde{\boldsymbol{z}}')$ (according to (4)) where $\tilde{\boldsymbol{z}}$ and $\tilde{\boldsymbol{z}}'$ are rows of $\tilde{\boldsymbol{Z}}$. Similarly, the covariance $\boldsymbol{C}_{Z\tilde{Z}} \in \mathbb{R}^{n\times m}$ consists of entries $c(\boldsymbol{z}, \tilde{\boldsymbol{z}})$ where $\boldsymbol{z}$ is a row of $\boldsymbol{Z}$. The posterior probability density of all unobserved variables

$$p_\Theta\left(\boldsymbol{f}, \tilde{\boldsymbol{f}}, \boldsymbol{Z} \,|\, \boldsymbol{y}\right) = \frac{p_\Theta\left(\boldsymbol{y}, \boldsymbol{f}, \tilde{\boldsymbol{f}}, \boldsymbol{Z}\right)}{p_\Theta(\boldsymbol{y})} = p_\theta\left(\boldsymbol{f} \,|\, \tilde{\boldsymbol{f}}, \boldsymbol{Z}\right) p_\Theta\left(\tilde{\boldsymbol{f}} \,|\, \boldsymbol{y}\right) p_W(\boldsymbol{Z}), \tag{32}$$

is analytically intractable and is therefore approximated by a Gaussian trial probability density

$$q_{\theta,\Psi}\left(\boldsymbol{f}, \tilde{\boldsymbol{f}}, \boldsymbol{Z}\right) = p_\theta\left(\boldsymbol{f} \,|\, \tilde{\boldsymbol{f}}, \boldsymbol{Z}\right) q_\omega\left(\tilde{\boldsymbol{f}}\right) q_\psi(\boldsymbol{Z}), \tag{33}$$

where $\boldsymbol{\Psi} = \boldsymbol{\psi} \cup \boldsymbol{\omega}$ are variational parameters. We approximate the analytically intractable posterior probability density $p_\Theta(\tilde{\boldsymbol{f}} \,|\, \boldsymbol{y})$ with the Gaussian trial density

$$q_\omega\left(\tilde{\boldsymbol{f}}\right) := \mathcal{N}\left(\tilde{\boldsymbol{\mu}}_{\tilde{f}}, \tilde{\boldsymbol{\Sigma}}_{\tilde{f}}\right), \tag{34}$$

where $\boldsymbol{\omega} = \{\tilde{\boldsymbol{\mu}}_{\tilde{f}}, \tilde{\boldsymbol{\Sigma}}_{\tilde{f}}\}$ are the free variational parameters. The mean vector $\tilde{\boldsymbol{\mu}}_{\tilde{f}} \in \mathbb{R}^m$ and diagonal covariance matrix $\tilde{\boldsymbol{\Sigma}}_{\tilde{f}} \in \mathbb{R}^{m\times m}$ are trainable parameters.

For determining the variational parameters, the KL divergence between the posterior probability density and Gaussian trial probability density

$$D_{KL}\left(q_{\theta,\Psi}\left(\boldsymbol{f}, \tilde{\boldsymbol{f}}, \boldsymbol{Z}\right) \,\|\, p_\Theta\left(\boldsymbol{f}, \tilde{\boldsymbol{f}}, \boldsymbol{Z} \,|\, \boldsymbol{y}\right)\right) = \int q_{\theta,\Psi}\left(\boldsymbol{f}, \tilde{\boldsymbol{f}}, \boldsymbol{Z}\right) \ln\left(\frac{q_{\theta,\Psi}(\boldsymbol{f}, \tilde{\boldsymbol{f}}, \boldsymbol{Z})}{p_\Theta(\boldsymbol{y}, \boldsymbol{f}, \tilde{\boldsymbol{f}}, \boldsymbol{Z})}\right) d\boldsymbol{Z} d\tilde{\boldsymbol{f}} d\boldsymbol{f} + \ln p_\Theta(\boldsymbol{y}), \tag{35}$$

is minimised. By factorising the KL divergence using the joint probability density $p_\Theta(\boldsymbol{y}, \boldsymbol{f}, \tilde{\boldsymbol{f}}, \boldsymbol{Z})$ and trial probability density $q_{\theta,\Psi}(\boldsymbol{f}, \tilde{\boldsymbol{f}}, \boldsymbol{Z})$ as in (29) and (33) respectively, we obtain the ELBO

$$\mathcal{F}(\boldsymbol{y}) = \mathbb{E}_{q_\psi(Z)}\left(\hat{\mathcal{F}}(\boldsymbol{y}, \boldsymbol{Z})\right) - D_{KL}\left(q_\omega\left(\tilde{\boldsymbol{f}}\right) \,\|\, p_\theta\left(\tilde{\boldsymbol{f}}\right)\right) - D_{KL}\left(q_\psi(\boldsymbol{Z}) \,\|\, p_W(\boldsymbol{Z})\right), \tag{36}$$

where $\hat{\mathcal{F}}(\boldsymbol{y})$ according to [54], is given by

$$\begin{aligned}\hat{\mathcal{F}}(\boldsymbol{y}, \boldsymbol{Z}) &= \mathbb{E}_{q_\omega(\tilde{f})}\left(\mathbb{E}_{p_\theta(f\,|\,\tilde{f},Z)}\left(\ln p_{\sigma_y}(\boldsymbol{y}|\boldsymbol{f})\right)\right) \\ &= \ln \mathcal{N}\left(\boldsymbol{y}\,|\,\boldsymbol{C}_{Z\tilde{Z}}\boldsymbol{C}_{\tilde{Z}\tilde{Z}}^{-1}\tilde{\boldsymbol{\mu}}_{\tilde{f}},\, \sigma_y^2\boldsymbol{I}\right) - \frac{1}{2\sigma_y^2}\left(\mathrm{Tr}\left(\boldsymbol{C}_{ZZ} - \boldsymbol{C}_{Z\tilde{Z}}\boldsymbol{C}_{\tilde{Z}\tilde{Z}}^{-1}\boldsymbol{C}_{\tilde{Z}Z}\right) + \mathrm{Tr}\left(\tilde{\boldsymbol{\Sigma}}_{\tilde{f}}\boldsymbol{C}_{\tilde{Z}\tilde{Z}}^{-1}\boldsymbol{C}_{\tilde{Z}Z}\boldsymbol{C}_{Z\tilde{Z}}\boldsymbol{C}_{\tilde{Z}\tilde{Z}}^{-1}\right)\right).\end{aligned} \tag{37}$$

MC sampling of $\boldsymbol{Z} \sim q_\psi(\boldsymbol{Z})$ is used to estimate the expectation in the ELBO $\mathcal{F}(\boldsymbol{y})$. The KL divergence terms are derived analytically given that all probability densities belong to the exponential family (Appendix A).

The approximate posterior probability density of the test output variables $\boldsymbol{f}_*$ evaluated at test latent variables $\boldsymbol{Z}_* \in \mathbb{R}^{n_* \times d_z}$ given by

$$\begin{aligned}q_{\theta,\omega}(\boldsymbol{f}_*|\boldsymbol{Z}_*) &= \int p_\theta\left(\boldsymbol{f}_*\,|\,\tilde{\boldsymbol{f}}, \boldsymbol{Z}_*\right) q_\omega\left(\tilde{\boldsymbol{f}}\right) d\tilde{\boldsymbol{f}} \\ &= \mathcal{N}\left(\boldsymbol{C}_{Z_*\tilde{Z}}\boldsymbol{C}_{\tilde{Z}\tilde{Z}}^{-1}\tilde{\boldsymbol{\mu}}_{\tilde{f}},\, \boldsymbol{C}_{Z_*Z_*} - \boldsymbol{C}_{Z_*\tilde{Z}}\boldsymbol{C}_{\tilde{Z}\tilde{Z}}^{-1}\boldsymbol{C}_{\tilde{Z}Z_*} + \boldsymbol{C}_{Z_*\tilde{Z}}\boldsymbol{C}_{\tilde{Z}\tilde{Z}}^{-1}\tilde{\boldsymbol{\Sigma}}_{\tilde{f}}\boldsymbol{C}_{\tilde{Z}\tilde{Z}}^{-1}\boldsymbol{C}_{\tilde{Z}Z_*}\right),\end{aligned} \tag{38}$$

is derived by marginalising the conditional probability density $p_\theta(\boldsymbol{f}_*\,|\,\tilde{\boldsymbol{f}}, \boldsymbol{Z}_*)$ (of similar form to (30)) over the pseudo output variables $\tilde{\boldsymbol{f}}$ sampled from the trial probability density $q_\omega(\tilde{\boldsymbol{f}})$. By augmenting the training data set $\mathcal{D}$ with pseudo output variables $\tilde{\boldsymbol{f}}$ the computational complexity is reduced from $O(n^3)$ to $O(nm^2)$ (see also [49–51]), since the GP prior $p_\theta(\boldsymbol{f}|\boldsymbol{Z}) = \int p_\theta(\boldsymbol{f}\,|\,\tilde{\boldsymbol{f}}, \boldsymbol{Z})p_\theta(\tilde{\boldsymbol{f}})\,d\tilde{\boldsymbol{f}}$ now depends on the covariance matrix $\boldsymbol{C}_{\tilde{Z}\tilde{Z}}$ over the m pseudo latent variables $\tilde{\boldsymbol{Z}}$.

3.3. Inference with input uncertainty

It is computationally more efficient to optimise in RDO the RDVGP surrogate over the low-dimensional test latent variables $\boldsymbol{Z}_*$ instead of the high-dimensional test input variables $\boldsymbol{S}_*$. For any test latent variables $\boldsymbol{Z}_*$ and corresponding test output variable $\boldsymbol{f}_*$, the approximate marginal posterior probability density given by

$$q_{\theta,\omega,W}(\boldsymbol{f}_*) = \int q_{\theta,\omega}(\boldsymbol{f}_*|\boldsymbol{Z}_*)\, p_W(\boldsymbol{Z}_*)\, d\boldsymbol{Z}_*, \tag{39}$$

is obtained from the derived approximate posterior probability density $q_{\theta,\omega}(\boldsymbol{f}_*|\boldsymbol{Z}_*)$ in (38) and the marginal probability density

$$p_W(\boldsymbol{Z}_*) = \prod_{i_*=1}^{n_*} \mathcal{N}\left(\boldsymbol{W}^\mathsf{T}\bar{\boldsymbol{s}}_{i_*},\, \boldsymbol{W}^\mathsf{T}\boldsymbol{\Sigma}_s\boldsymbol{W}\right), \tag{40}$$

where the objective of RDO is to compute $\bar{\boldsymbol{z}}^* = \boldsymbol{W}^\mathsf{T}\bar{\boldsymbol{s}}^*$, which is mapped to the input domain using the Moore-Penrose inverse $\boldsymbol{W}^\dagger$. For any test latent variable $\boldsymbol{z}_*$ (that is any row of $\boldsymbol{Z}_*$), the mean

$$\mathbb{E}(f_*) = \mathbb{E}_{p_W(z_*)}\left(\mathbb{E}_{q_{\theta,\omega}(f_*|z_*)}(f_*)\right), \tag{41}$$

and variance

$$\mathrm{var}(f_*) = \mathrm{var}_{p_W(z_*)}\left(\mathbb{E}_{q_{\theta,\omega}(f_*|z_*)}(f_*)\right) + \mathbb{E}_{p_W(z_*)}\left(\mathrm{var}_{p_{\theta,\omega}(f_*|z_*)}(f_*)\right), \tag{42}$$

of the approximate marginal posterior probability density $q_{\theta,\omega,W}(f_*)$ given in (39), can be approximated using the laws of total expectation and total variance, with MC sampling.

3.4. Surrogate training

The RDVGP surrogate is trained by learning the set of hyperparameters $\boldsymbol{\Theta} = \boldsymbol{\theta} \cup \{\boldsymbol{W}, \sigma_y\}$ and variational parameters $\boldsymbol{\Psi} = \boldsymbol{\psi} \cup \boldsymbol{\omega}$, that maximise the ELBO $\mathcal{F}(\boldsymbol{y})$ given by (36), such that

$$\boldsymbol{\Theta}^*, \boldsymbol{\Psi}^* = \arg\max_{\Theta,\Psi} \mathcal{F}(\boldsymbol{y}). \tag{43}$$

We use the stochastic gradient method with the adaptive moment estimation algorithm (ADAM) [55], which evaluates the gradient of the expectation taken with respect to variational parameters of the latent variable trial probability density $q_\psi(\boldsymbol{Z})$ given by (27). Each ADAM iteration can be repeated n_t times until convergence. We use the default ADAM step size and decay rates [55]. As discussed in [56], to compute the gradient, the latent variable must use the reparameterisation given by

$$\boldsymbol{z}_i = \tilde{\boldsymbol{\mu}}_{z,i} + \tilde{\boldsymbol{L}}_{z,i}\boldsymbol{\epsilon}_I, \tag{44}$$

where $\boldsymbol{\epsilon}_I \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ and $\tilde{\boldsymbol{L}}_{z,i}$ is the lower-triangular matrix computed using the Cholesky decomposition of $\tilde{\boldsymbol{\Sigma}}_{z,i}$ (this is known as the reparameterisation trick). The expectation in the ELBO $\mathcal{F}(\boldsymbol{y})$ is estimated using MC sampling with n_l samples. We compute the ELBO gradient using a single MC sample estimate, which is sufficient for many practical applications [57].

Since the linear projection matrix $\boldsymbol{W}$ from the statistical observation model (16a) diverges from the Stiefel manifold (is no-longer orthogonal) if using ADAM, the Cayley ADAM algorithm [48] is used (we use the default step size and decay rates). An automatic relevance determination (ARD) procedure [58] is used to determine the dimension d_z of the latent variables $\boldsymbol{z}$.

Since the ELBO is non-convex, optimisation on the Stiefel manifold using gradient methods may converge to local minima [59], n_r restarts with different initial hyperparameters are used (Algorithm 1). Exploration of highly redundant random input variables $\boldsymbol{s}$ can be prevented using a sparse orthogonal initialisation of the projection matrix $\boldsymbol{W}$; however, a random initialisation (using QR decomposition [60]) is more appropriate for input variables $\boldsymbol{s}$ with low redundancy (we use a combination of both). The covariance matrices of the trial probability densities $\tilde{\boldsymbol{\Sigma}}_{\tilde{f}}$ and $\tilde{\boldsymbol{\Sigma}}_{z,i}$ are sampled from the space of positive definite diagonal matrices.

Training data set $\mathcal{D}$ is proposed using Latin hypercube sampling (LHS) [61]. Input variables $\boldsymbol{s} = (\boldsymbol{s}_d^\mathsf{T}\ \boldsymbol{s}_f^\mathsf{T})^\mathsf{T}$ consist of immutable variables sampled from the domain $\boldsymbol{s}_f \in \left\{\boldsymbol{s}_f \in \mathbb{R}^{d_f} \,\middle|\, \bar{\boldsymbol{s}}_f - 2\boldsymbol{\sigma}_{s_f} \leq \boldsymbol{s}_f \leq \bar{\boldsymbol{s}}_f + 2\boldsymbol{\sigma}_{s_f}\right\}$, $\bar{\boldsymbol{s}}_f$ and $\boldsymbol{\sigma}_{s_f}$ are vectors of the mean and standard deviations corresponding to each dimension. Similarly, the design variables are sampled from the domain $\boldsymbol{s}_d \in \left\{\boldsymbol{s}_d \in \mathbb{R}^{d_d} \,\middle|\, \bar{\boldsymbol{s}}_d^{(l)} - 2\boldsymbol{\sigma}_{s_d} \leq \boldsymbol{s}_d \leq \bar{\boldsymbol{s}}_d^{(u)} + 2\boldsymbol{\sigma}_{s_d}\right\}$ where $\bar{\boldsymbol{s}}_d^{(l)}$, $\bar{\boldsymbol{s}}_d^{(u)}$

Algorithm 1 Pseudocode for training the RDVGP surrogate

Data: $\mathcal{D} = \{(\boldsymbol{s}_i, y_i) \,|\, i \in \mathbb{N}, i \leq n\}$ ▷ Sampled using LHS
Input: restarts n_r, ADAM iterations n_t, and MC samples n_l
for $r \in \{1, 2, ..., n_r\}$ **do** ▷ Multiple restarts
  Initialise model hyperparameters $\boldsymbol{\Theta}_0, \boldsymbol{\Psi}_0$
  for $t \in \{1, 2, ..., n_t\}$ **do** ▷ Gradient descent
    for $l \in \{1, 2, ..., n_l\}$ **do** ▷ MC sampling
      $\boldsymbol{Z}_l \sim q_\psi(\boldsymbol{Z})$ ▷ Reparameterisation (44)
    Compute $\mathcal{F}_t(\boldsymbol{y})$ and its gradient using $\{\boldsymbol{Z}_1, \boldsymbol{Z}_2, ..., \boldsymbol{Z}_{n_l}\}$, $\boldsymbol{\Theta}_{t-1}$, and $\boldsymbol{\Psi}_{t-1}$ ▷ ELBO (36)
    $\boldsymbol{\Theta}_t \backslash \{\boldsymbol{W}_t\}, \boldsymbol{\Psi}_t \leftarrow \text{ADAM}(\mathcal{F}_t(\boldsymbol{y}))$ ▷ Stochastic gradient method
    $\boldsymbol{W}_t \leftarrow \text{CAYLEYADAM}(\mathcal{F}_t(\boldsymbol{y}))$ ▷ Stiefel manifold optimisation
Result: optimum model hyperparameters $\boldsymbol{\Theta}^*, \boldsymbol{\Psi}^*$

and $\boldsymbol{\sigma}_{s_d}$ are vectors of lower limits, upper limits, and standard deviations corresponding to each dimension.

Slice sampling is used to infer the test output variables $\boldsymbol{f}_*$ corresponding to the design variables $\bar{\boldsymbol{s}}_{d*}$. The approximate posterior probability density $q_{\theta,\omega}(\boldsymbol{f}_*|\boldsymbol{z}_*)$ given by (38), is evaluated at the projected mean input variables $\bar{\boldsymbol{z}}_* = \boldsymbol{W}^\mathsf{T}\bar{\boldsymbol{s}}_*$ where $\bar{\boldsymbol{s}}_* = (\bar{\boldsymbol{s}}_{d*}^\mathsf{T}\ \bar{\boldsymbol{s}}_f^\mathsf{T})^\mathsf{T}$ and $\bar{\boldsymbol{s}}_{d*} \in \left\{\bar{\boldsymbol{s}}_{d*} \in \mathbb{R}^{d_d} \,\middle|\, \bar{\boldsymbol{s}}_d^{(l)} \leq \bar{\boldsymbol{s}}_{d*} \leq \bar{\boldsymbol{s}}_d^{(u)}\right\}$ (Figure 5).

3.5. *Surrogate comparison*

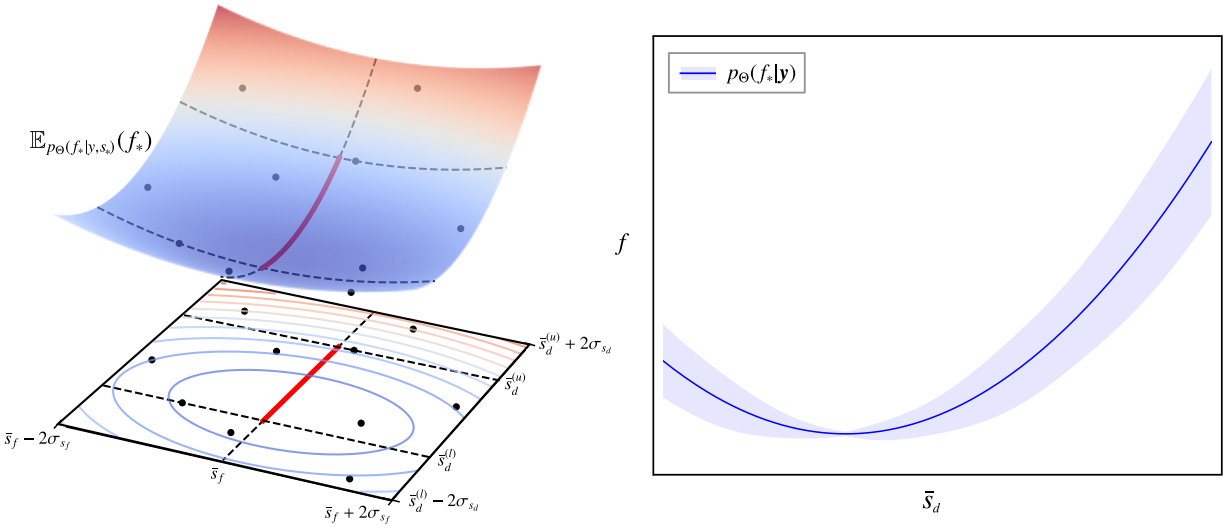
We verify the predictive accuracy of the RDVGP surrogate using two performance metrics; the coefficient of determination (COD) [62], and the maximum mean discrepancy (MMD) [63]. For any validation mean input variables $\bar{\boldsymbol{S}}_v$, corresponding true observations $\boldsymbol{y}_v$ of the objective or constraint function are obtained by sampling the input variable density $p(\boldsymbol{S}_v)$. Summary statistics of the true observations (such as mean and variance) can be compared to the inferred target output variables $\boldsymbol{f}_v$. The COD values for the mean and variance

$$R_\mu^2 = 1 - \frac{\sum_{v=1}^{n_v} \left|\mathbb{E}(\boldsymbol{y}_v) - \mathbb{E}(\boldsymbol{f}_v)\right|^2}{\sum_{v=1}^{n_v} \left|\mathbb{E}(\boldsymbol{y}_v) - \mu_{y_v}\right|^2}, \tag{45a}$$

$$R_\sigma^2 = 1 - \frac{\sum_{v=1}^{n_v} \left|\mathrm{var}(\boldsymbol{y}_v) - \mathrm{var}(\boldsymbol{f}_v)\right|^2}{\sum_{v=1}^{n_v} \left|\mathrm{var}(\boldsymbol{y}_v) - \sigma_{y_v}^2\right|^2}, \tag{45b}$$

of the marginal posterior probability density $q_{\theta,\omega,W}(\boldsymbol{f}_v)$ in (39) are computed, where $\mu_{y_v} = \frac{1}{N_v}\sum_{v=1}^{n_v}\mathbb{E}(\boldsymbol{y}_v)$ and $\sigma_{y_v}^2 = \frac{1}{N_v}\sum_{v=1}^{n_v}\mathrm{var}(\boldsymbol{y}_v)$.

While the COD provides a measure of the quality of fit over the entire model, the distance between probability densities at points of interest such as local minima is important for determining the reliability of the predicted robust optimum design variables. The MMD $D_{MMD} \in \mathbb{R}^+$ provides this measure in terms of



(a) Slice sampling of $\mathbb{E}_{p_\Theta(f_*|\boldsymbol{y},\boldsymbol{s}_*)}(f_*)$

(b) Marginal posterior density $p_\Theta(f_*|\boldsymbol{y})$ along slice $\bar{\boldsymbol{s}}_*$

Figure 5: Schematic of slice sampling showing (a) the posterior probability density expected value $\mathbb{E}_{p_\Theta(f_*|\boldsymbol{y},\boldsymbol{s}_*)}(f_*)$ along slice $\boldsymbol{s}_* = (s_{d*}\ \bar{s}_f)^\mathsf{T}$, and (b) the marginal posterior probability density $p_\Theta(f_*|\boldsymbol{y})$ over mean test input variable $\bar{\boldsymbol{s}}_* = (\bar{s}_{d*}\ \bar{s}_f)^\mathsf{T}$.

a mean embedding in reproducing kernel Hilbert space with associated kernel $k : \mathbb{R}^{d_y} \times \mathbb{R}^{d_y} \to \mathbb{R}$. For two probability densities $p(\boldsymbol{w})$ and $q(\boldsymbol{v})$, the MMD is

$$D_{MMD} = \mathbb{E}_{p(\boldsymbol{w})}\left(k(\boldsymbol{w}, \boldsymbol{w}')\right) - 2\,\mathbb{E}_{p(\boldsymbol{w}),q(\boldsymbol{v})}\left(k(\boldsymbol{w}, \boldsymbol{v})\right) + \mathbb{E}_{q(\boldsymbol{v})}\left(k(\boldsymbol{v}, \boldsymbol{v}')\right). \tag{46}$$

In all examples in this paper, we use a Gaussian kernel with unit amplitude and length scale to compute the MMD.

4. Examples

We demonstrate the accuracy of the proposed RDVGP surrogate using four examples of increasing complexity. In all examples, we use the sparse RDVGP surrogate, but refer to it as the RDVGP surrogate for brevity. Illustrative examples are used to compare the RDVGP surrogate to standard GP regression and examines its versatility to different levels of input variance. We also demonstrate the RDVGP surrogate on RDO examples.

The robust minimum as predicted by the RDVGP surrogate, is determined with genetic algorithms (NSGA-II), which use the emulated solutions inferred by the surrogate. To verify its accuracy, we directly MC sample the input variable and solve the corresponding FE model to estimate the true posterior probability density over the domain of the design variables. This is also optimised using genetic algorithms to yield an estimate of the true solution, for comparison.

4.1. One-dimensional illustrative example

The approximation of marginal posterior probability densities using the RDVGP surrogate is compared to the conventional approach of MC sampling from a standard GP surrogate as in (14) [64]. For both surrogates, the inferred variance depends on the random input variables $\boldsymbol{s}$, allowing variation over the input variable domain $\boldsymbol{s} \in D_s$. Consider the non-linear multi-modal objective function

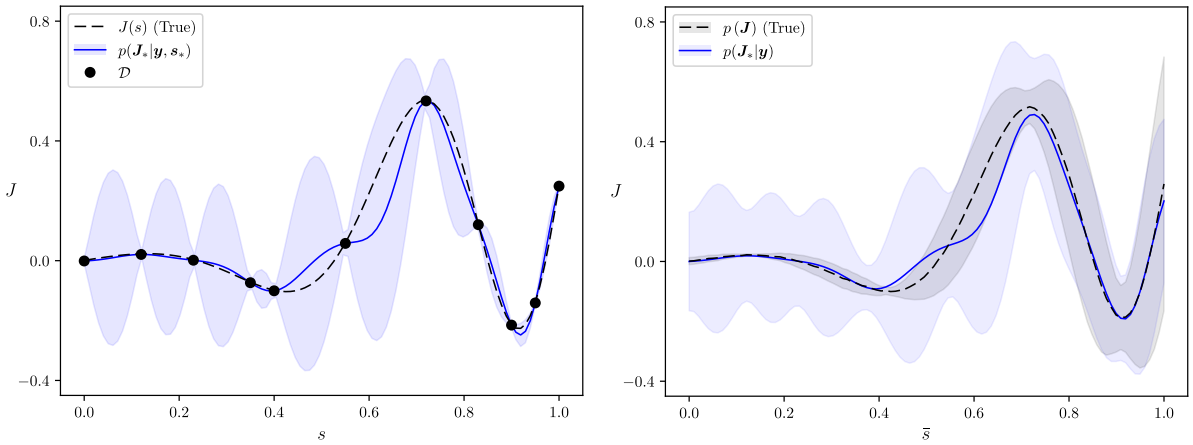
$$J(s) = -0.5s\sin(3\pi s^2) + 0.25s, \tag{47}$$

where $s \sim \mathcal{N}(\bar{s}, \sigma^2)$, $\bar{s} \in \{\bar{s} \in \mathbb{R} \,|\, 0 \leq \bar{s} \leq 1\}$, and $\sigma = 0.025$. The training data set $\mathcal{D}$ is initialised using LHS with $n = 11$ random samples (as shown in Figures 6a and 6c) and the same number of inducing points (with $m = 11$). This example considers only a single input variable, hence no dimensionality reduction is required. The standard GP and RDVGP surrogates are trained by maximising the log marginal likelihood $\ln p_\Theta(\boldsymbol{y})$ from (10) and ELBO $\mathcal{F}(\boldsymbol{y})$ given by (36) respectively.

The standard GP overfits the training data (Figure 6a) due to the small length scale (hyperparameter of the covariance function), resulting in a posterior probability density $p(\boldsymbol{J}_*|\boldsymbol{y}, \boldsymbol{s}_*)$ with large epistemic uncertainty between sampled input variables. As mentioned in the introduction, epistemic uncertainty is caused by a lack of information and can in principle be reduced with additional training data. Conversely, aleatoric uncertainty is caused by the random input variable $\boldsymbol{s}$ itself [20]. When using MC sampling to approximate the true marginal posterior probability density $p(\boldsymbol{J})$ using the standard GP posterior probability density $p(\boldsymbol{J}_*|\boldsymbol{y}, \boldsymbol{s}_*)$ given by (14), the aleatoric and epistemic uncertainties are confounded and the total uncertainty is significantly overestimated (Figure 6b).

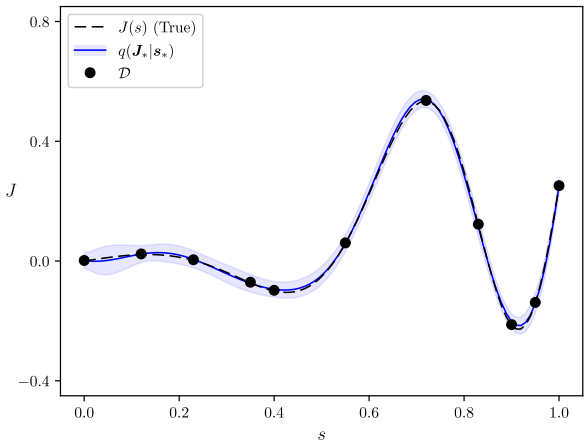
In contrast, the RDVGP surrogate does not overfit the training data (Figure 6c); due to the inclusion of the prescribed input variable probability density $p(\boldsymbol{s})$ in the ELBO $\mathcal{F}(\boldsymbol{y})$, adding a data-informed regularisation. The expectation in the ELBO $\mathcal{F}(\boldsymbol{y})$ formulation promotes smoothness near the sampled training data, since each MC sample $\boldsymbol{Z} \sim q_\Psi(\boldsymbol{Z})$ (following from (27)) is verified for a closeness of fit between the

true observations $\boldsymbol{y}$ and the inferred target output variables $\boldsymbol{f}$ using a likelihood formulation $\hat{\mathcal{F}}(\boldsymbol{y}, \boldsymbol{Z})$ given by (37), which promotes length scale hyperparameters that increase smoothness. This adds regularisation to the model and the use of pseudo target output variables $\tilde{\boldsymbol{f}}$ adds further flexibility when fitting the model (to prevent over-fitting). In comparison to the standard GP surrogate, this reduces epistemic uncertainty and more accurately estimates the total uncertainty, for improved selection of robust optimum design variables $\bar{\boldsymbol{s}}^*$ (Figure 6d). In this example, the standard GP surrogate would require more training data than the RDVGP surrogate to reduce overfitting before accurately emulating the true marginal posterior probability density $p(\boldsymbol{J})$.

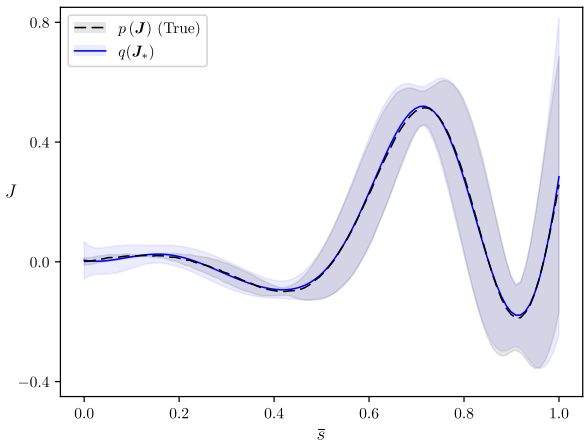


(a) Standard GP surrogate $p(\boldsymbol{J}_*|\boldsymbol{y}, \boldsymbol{s}_*)$

(b) Standard GP marginal posterior density $p(\boldsymbol{J}_*|\boldsymbol{y})$



(c) RDVGP surrogate $q(\boldsymbol{J}_*|\boldsymbol{s}_*)$



(d) RDVGP marginal posterior density $q(\boldsymbol{J}_*)$

Figure 6: One-dimensional illustrative example. Comparing the (a) standard GP surrogate $p(\boldsymbol{J}_*|\boldsymbol{y}, \boldsymbol{s}_*)$ given by (14) and its (b) marginal posterior probability density $p(\boldsymbol{J}_*|\boldsymbol{y}) = \mathbb{E}_{p(\boldsymbol{s}_*)}(p(\boldsymbol{J}_*|\boldsymbol{y}, \boldsymbol{s}_*))$, with the (c) RDVGP surrogate $q(\boldsymbol{J}_*|\boldsymbol{s}_*)$ given by (38), and (d) RDVGP marginal posterior probability density $q(\boldsymbol{J}_*) = \mathbb{E}_{p(\boldsymbol{s}_*)}(q(\boldsymbol{J}_*|\boldsymbol{s}_*))$ from (39). The training data set $\mathcal{D}$ consists of $n = 11$ random input variable s and observation $\boldsymbol{y}$ pairs sampled from the true objective function $J(s)$, and $p(\boldsymbol{J})$ is the true marginal posterior probability density.

4.2. Three-dimensional illustrative example

This example investigates the accuracy and robustness of the RDVGP surrogate to different levels of input variable uncertainty and choices of parameters used to train the surrogate. Consider the minimisation of a non-linear multi-modal objective function

$$J(\boldsymbol{s}) = (s_2 s_1 - 2)^2 \sin(12 s_1 - 4) + 8 s_1 + s_3, \tag{48}$$

of three dimensional input variables $\boldsymbol{s} = (s_1 \;\; s_2 \;\; s_3)^{\mathsf{T}}$, subject to the constraint function

$$H(\boldsymbol{s}) = -\cos(2\pi s_1) - s_3 - 0.7, \tag{49}$$

where $s_1 \sim \mathcal{N}(\bar{s}_1, \sigma_1^2)$ is a design variable, and $s_2 \sim \mathcal{N}(6.0, \sigma_2^2)$ and $s_3 \sim \mathcal{N}(0, 0.1^2)$ are immutable variables. The mean of the design variable is sampled from the domain $\bar{s}_1 \in \{\bar{s}_1 \in \mathbb{R} \,|\, 0 \leq \bar{s}_1 \leq 1\}$. We perform RDO at three different choices for $\sigma_1 \in \{0.025, 0.05, 0.075\}$ and $\sigma_2 \in \{0.25, 0.5, 0.75\}$. According to the RDO problem stated in (1), a weighting factor of $\alpha = 0.25$ and feasibility index of $\beta = 2$ is chosen. No limit is imposed on the variance of the constraint function $H(\boldsymbol{s})$.

The training data set $\mathcal{D}$ is initialised using LHS with $n = 100$ samples and choosing $m = 25$ inducing points (reducing time complexity by a factor of 16 from $O(n^3)$ to $O(nm^2)$). The robust minimum variable $\bar{s}_1^*$ computed using the RDVGP surrogate is compared to the true minimum estimated using MC sampling $\langle \bar{s}_1^* \rangle$ with 10^4 samples. A total of $n_v = 30$ validation samples are used to compute COD values (given by (45)) of the inferred objective and constraint marginal posterior probability densities $q(\boldsymbol{J}_*)$ and $q(\boldsymbol{H}_*)$ respectively.

The immutable variable s_3 is substantially less influential than s_1 and s_2, giving a subspace dimension of $d_z = 2$. An accurate robust minimum variable $\bar{s}_1^*$ is computed for all standard deviation combinations σ_1 and σ_2, when compared to the true solution (Table 1). The largest computed percentage difference between the RDVGP $\bar{s}_1^*$ and true $\langle \bar{s}_1^* \rangle$ robust optimum design variables is 4%. Although the objective function $J(\boldsymbol{s})$ has two local minima, one is not robust since the objective function varies significantly in close proximity to that local minima (Figure 7). The approximate marginal posterior probability density $q(\boldsymbol{J}_*)$ computed using the RDVGP surrogate accurately emulates the true marginal posterior probability density $p(\boldsymbol{J})$, therefore the robust minimum $\bar{s}_1^*$ is well estimated. If the uncertainty in the random input variables becomes negligible, the global minimum would be $\bar{s}_1 = 0.749$, which does not coincide with the robust optimum design variable for any of the sampled statistics (Table 1).

Large COD values R_μ^2 are obtained from (45a) for the mean of both the objective $J(\boldsymbol{s})$ and constraint $H(\boldsymbol{s})$ functions for all combinations of σ_1 and σ_2, indicating the mean is well predicted. However, the COD value R_σ^2 obtained from (45b) for the variance of the objective and constraint functions varies for different combinations of σ_1 and σ_2, and is most affected by the choice of σ_1. Furthermore, the MMD D_{MMD}, which compares the true $p(J(\langle \bar{s}_1^* \rangle))$ and approximate (RDVGP) $q(J(\langle \bar{s}_1^* \rangle))$ marginal posterior probability densities of the objective function at the true robust optimum design variable $\langle \bar{s}_1^* \rangle$ as given by (46), is affected by both σ_1 and σ_2. At smaller values for σ_1, the epistemic uncertainty makes up a greater proportion of the total uncertainty predicted by the RDVGP surrogate, increasing its discrepancy with the true solution. Since a slice of the full surrogate is taken at $s_2 = 6.0$, it becomes more difficult to approximate the aleatoric uncertainty at larger values of σ_2.

As previously mentioned, the epistemic uncertainty can be reduced by increasing the size of training data set $\mathcal{D}$ (Figure 8). By including supplementary data sampled using LHS centred on the true robust optimum design variable $\langle \bar{s}_1^* \rangle$ and box bounded within two standard deviations of each random input variable, the approximate (RDVGP) marginal posterior probability density $q(J(\langle \bar{s}_1^* \rangle))$ converges towards the true marginal posterior probability density $p(J(\langle \bar{s}_1^* \rangle))$ with an increasing number of supplementary samples n_s.

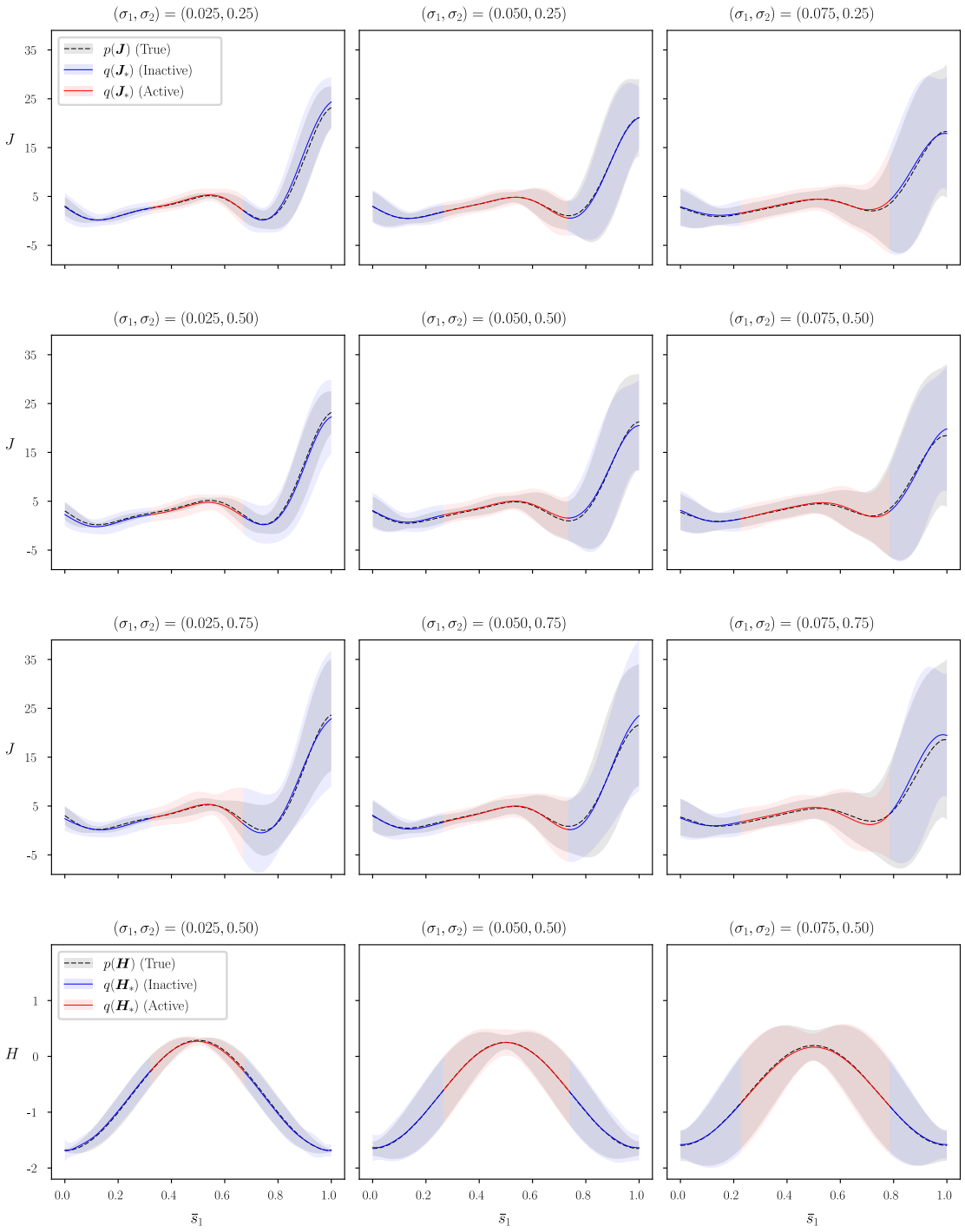
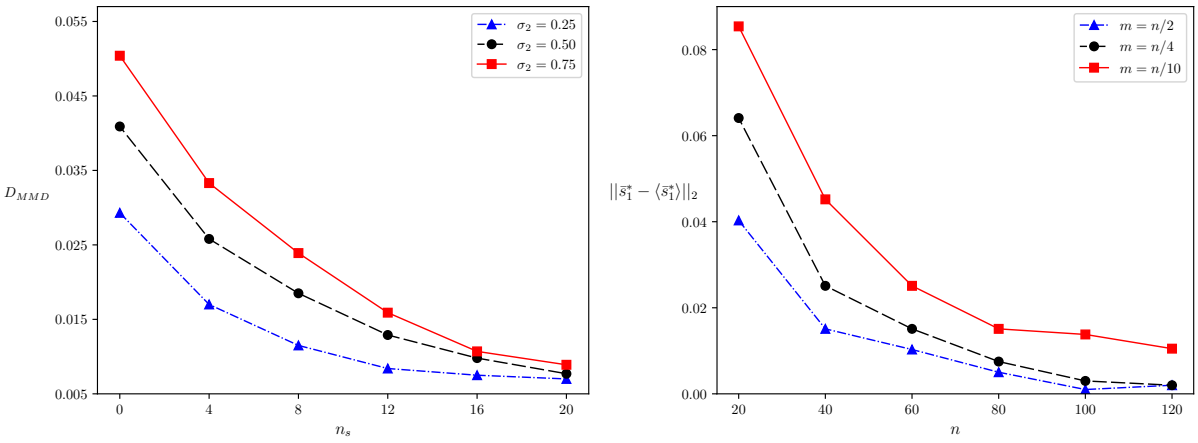


Figure 7: Three-dimensional illustrative example. Comparison of the true $p(\cdot)$ and approximate (RDVGP) $q(\cdot)$ marginal posterior probability densities of the objective and constraint functions $J(\boldsymbol{s})$ and $H(\boldsymbol{s})$ over the domain of the design variable s_1. The design variable has density $\bar{s}_1 \sim \mathcal{N}(\bar{s}_1, \sigma_1^2)$ and the immutable variables have densities $\bar{s}_2 \sim \mathcal{N}(\bar{s}_2, \sigma_2^2)$ and $\bar{s}_3 \sim \mathcal{N}(\bar{s}_3, 0.1^2)$. The constraint is active only in the red region.

Table 1: Three-dimensional illustrative example. Robust minimum design variable as predicted by the RDVGP surrogate $\bar{s}_1^*$ and MC sampling $\langle \bar{s}_1^* \rangle$, with COD performance metrics computed using (45) for the objective $\left(\left(R_\mu^J\right)^2, \left(R_\sigma^J\right)^2 \right)$ and constraint $\left(\left(R_\mu^H\right)^2, \left(R_\sigma^H\right)^2 \right)$ functions, and MMD D_{MMD} (computed using (46)) between the approximate (RDVGP) $q(J(\langle \bar{s}_1^* \rangle))$ and true $p(J(\langle \bar{s}_1^* \rangle))$ marginal posterior probability densities evaluated at $\langle \bar{s}_1^* \rangle$.

σ_1	σ_2	$\left(R_\mu^J\right)^2$	$\left(R_\sigma^J\right)^2$	$\left(R_\mu^H\right)^2$	$\left(R_\sigma^H\right)^2$	D_{MMD}	$\bar{s}_1^*$	$\langle \bar{s}_1^* \rangle$
0.025	0.25	0.997	0.811	0.999	0.790	0.0293	0.128	0.123
	0.50	0.993	0.828	0.999	0.789	0.0409	0.127	0.124
	0.75	0.994	0.828	0.999	0.793	0.0504	0.127	0.124
0.050	0.25	0.991	0.923	0.998	0.883	0.0175	0.137	0.138
	0.50	0.985	0.924	0.999	0.888	0.0263	0.137	0.135
	0.75	0.988	0.929	0.998	0.887	0.0364	0.134	0.136
0.075	0.25	0.992	0.959	0.995	0.949	0.0039	0.152	0.148
	0.50	0.974	0.965	0.995	0.959	0.0076	0.151	0.155
	0.75	0.969	0.944	0.994	0.932	0.0108	0.151	0.150



(a) MMD D_{MMD} convergence

(b) Optimum solution $\bar{s}_1^*$ convergence

Figure 8: Three-dimensional illustrative example. Convergence plots showing the (a) effect of the number of supplementary training data points n_s sampled within the vicinity of the true robust optimum design variable $\langle \bar{s}_1^* \rangle$, on the MMD D_{MMD} computed between the true marginal posterior probability density $p(J(\langle \bar{s}_1^* \rangle))$ and the approximate (RDVGP) marginal posterior probability density $q(J(\langle \bar{s}_1^* \rangle))$ at $\sigma_1 = 0.025$ and varying values of σ_2. In (b) the convergence of the predicted solution $\bar{s}_1^*$ to the true robust optimum design variable $\langle \bar{s}_1^* \rangle$ for different training data sample sizes n and choices of number of inducing points m, for $(\sigma_1, \sigma_2) = (0.050, 0.50)$ is shown.

This is demonstrated by the MMD D_{MMD} decreasing for an increasing number of supplementary samples, a trend seen across all values of σ_2. Consequently, sequential or adaptive sampling strategies [65] would be effective for reducing the epistemic uncertainty when using the RDVGP surrogate.

4.3. Plate with elliptical hole

We study the use of the RDVGP surrogate on the RDO of a Poisson problem on a plate with an elliptical hole (Figure 9). The boundary value problem is given by

$$-\kappa\nabla^2 u(\boldsymbol{x},\boldsymbol{s}) = b(\boldsymbol{x},\boldsymbol{s}), \qquad \boldsymbol{x}\in\Omega, \tag{50a}$$

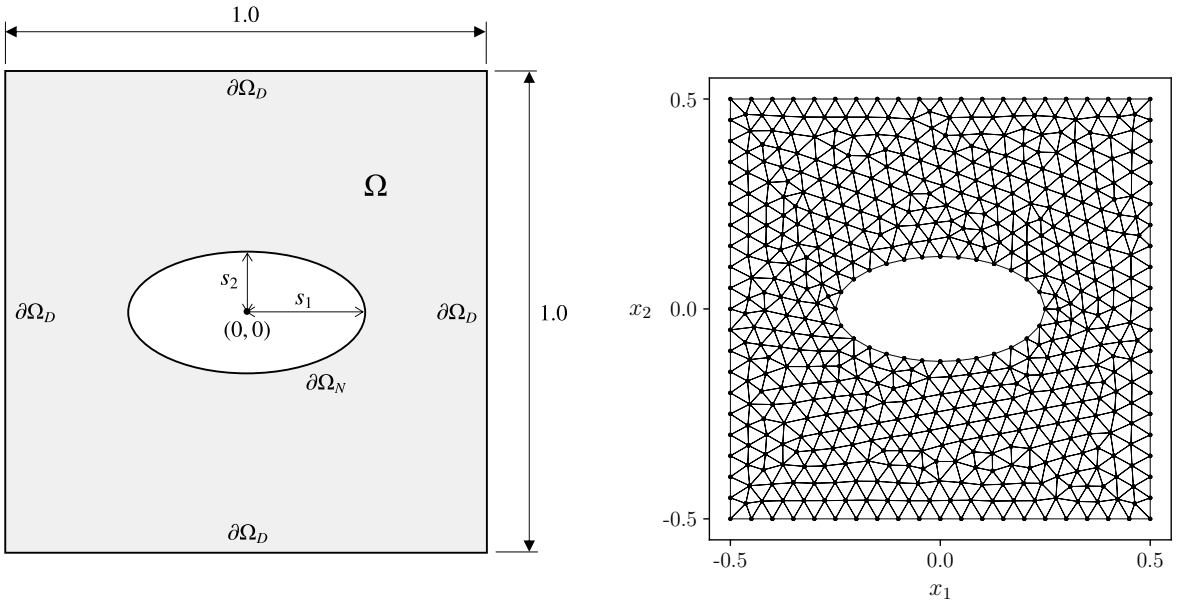
$$u(\boldsymbol{x},\boldsymbol{s}) = 0, \qquad \boldsymbol{x}\in\partial\Omega_D, \tag{50b}$$

$$\nabla u(\boldsymbol{x},\boldsymbol{s})\cdot\boldsymbol{n} = 0, \qquad \boldsymbol{x}\in\partial\Omega_N, \tag{50c}$$

where $\kappa\in\mathbb{R}^+$ is the diffusion coefficient, $b(\boldsymbol{x},\boldsymbol{s})\in\mathbb{R}$ is the source, $u(\boldsymbol{x},\boldsymbol{s})\in\mathbb{R}$ is the solution, $\partial\Omega_D$ is the Dirichlet boundary, $\partial\Omega_N$ is the Neumann boundary, and $\boldsymbol{n}$ is the normal to the Neumann boundary $\partial\Omega_N$. The domain Ω is discretised over spatial coordinates $\boldsymbol{x} = (x_1\ x_2)^\mathsf{T}$, and consists of a unit square $\Omega_1 = \{x_1, x_2\in\mathbb{R}\,|\,-0.5\le x_1\le 0.5, -0.5\le x_2\le 0.5\}$ with an elliptical void centred on the origin $\Omega_2 = \{x_1, x_2\in\mathbb{R}\,|\,(x_1/s_1)^2 + (x_2/s_2)^2 < 1\}$, where $\boldsymbol{s}_d = (s_1\ s_2)^\mathsf{T}$ are design variables and $\Omega = \Omega_1\setminus\Omega_2$. The random source $b(\boldsymbol{x},\boldsymbol{s})$ is a function of the fifth-order Chebyshev polynomial $T_5:\mathbb{R}\to\mathbb{R}$, given by

$$b(\boldsymbol{x},\boldsymbol{s}) = s_3\,(T_5(x_1 - s_4))^2 + s_5\,(T_5(x_2 - s_6))^2 + s_7, \tag{51}$$

where $\boldsymbol{s}_f = (s_3\ s_4\ \ldots\ s_7)^\mathsf{T}$ are immutable variables and $\boldsymbol{s} = (\boldsymbol{s}_d^\mathsf{T}\ \boldsymbol{s}_f^\mathsf{T})^\mathsf{T}$. The diffusion coefficient $\kappa = 1$ is constant throughout the domain Ω. All the random input variables s_i are Gaussian, where $s_i\sim\mathcal{N}(\bar{s}_i,\sigma_i^2)$ and $\bar{s}_i\in\mathbb{R}$ is a fixed scalar for $i\in\{3,4,\ldots,7\}$, and box-constrained $\bar{s}_i\in\{\bar{s}_i\in\mathbb{R}\,|\,0\le\bar{s}_i\le 0.4\}$ for $i\in\{1,2\}$. The goal is to compute the robust optimum design with $\bar{\boldsymbol{s}}_d^* = (\bar{s}_1^*\ \bar{s}_2^*)^\mathsf{T}$ that minimises the energy



(a) Domain with boundary conditions

(b) Example geometry mesh

Figure 9: Plate with elliptical hole. Schematic showing (a) the specification of the domain Ω with the elliptical void parameterised by design variables s_1 and s_2 centred at the origin with boundaries $\partial\Omega_D$ and $\partial\Omega_N$ where the Dirichlet and Neumann boundary conditions are applied respectively, and (b) mesh of an example geometry containing 982 linear triangle elements.

$$J(\boldsymbol{s}) = \frac{1}{2}\kappa \int_{\Omega} \|\nabla u(\boldsymbol{x}, \boldsymbol{s})\|_2^2 \, d\boldsymbol{x}, \tag{52}$$

subject to the area constraint

$$H(\boldsymbol{s}) = A_0 - \int_{\Omega} d\boldsymbol{x}, \tag{53}$$

where $A_0 = 0.8$ is the limit on the domain area. The constraint function $H(\boldsymbol{s})$ is introduced to eliminate trivial solutions with a zero hole size. According to the RDO formulation given by (1), a weighting factor of $\alpha = 0.5$ and a feasibility index of $\beta = 2$ is used. No condition is applied to the variance of the constraint function $H(\boldsymbol{s})$.

The training data set $\mathcal{D}$ is initialised using LHS with $n = 200$ samples and $m = 100$ inducing points. The objective $J(\boldsymbol{s})$ and constraint $H(\boldsymbol{s})$ functions are sampled from a deterministic FE model (Figure 9). To test the reliability of the RDVGP surrogate, five source terms $b(\boldsymbol{x}, \boldsymbol{s})$ are considered, each with different immutable variables $\boldsymbol{s}_f$ (Table 2). Design variables s_1 and s_2 have standard deviations of $\sigma_1 = 0.010$ and $\sigma_2 = 0.025$ respectively. RDO normalisation constants $\bar{\mu}$ and $\bar{\sigma}$ are equal to the mean and standard deviation of the objective function $J(\boldsymbol{s})$ observations. RDVGP robust optimum design variables $\bar{\boldsymbol{s}}_d^*$ are compared to the true solution $\langle \bar{\boldsymbol{s}}_d^* \rangle$ evaluated using MC sampling with 10^5 evaluations of the FE model. A total of $n_v = 30$ validation samples $\mathcal{D}_v$ are used for computing COD values (given by (45)) of the objective and constraint functions.

The computed robust optimum design variables $\bar{\boldsymbol{s}}_d^*$ maintain similarity to the true solution $\langle \bar{\boldsymbol{s}}_d^* \rangle$ for all source terms, with the most and least similarity corresponding to the third $b^{(3)}(\boldsymbol{x}, \boldsymbol{s})$ and fourth $b^{(4)}(\boldsymbol{x}, \boldsymbol{s})$ source terms respectively (Table 3). In both cases, the source term varies uniaxially and is aligned perpendicular and parallel to the direction of greatest variance in the shape of the elliptical hole respectively (since σ_2 is greater than σ_1) (Figure 10). This effect propagates through to the solution $u(\boldsymbol{x}, \boldsymbol{s})$ (Figure 11).

If the random input variables $\boldsymbol{s}$ are treated as deterministic, the optimum design variables $\hat{\boldsymbol{s}}_d^*$ would differ significantly from the robust optimum design variables $\bar{\boldsymbol{s}}_d^*$ computed using the RDVGP surrogate (Table 3). For each source term, a significant difference between the deterministic $\hat{\boldsymbol{s}}_d^*$ and robust $\bar{\boldsymbol{s}}_d^*$ optimum design variables is observed. Since this example considers the uncertainty as equally important as the mean (since $\alpha = 0.5$), variance in the objective $J(\boldsymbol{s})$ and constraint $H(\boldsymbol{s})$ functions cannot be ignored. In the case of all source terms $b(\boldsymbol{x}, \boldsymbol{s})$, the location of the minimum mean and variance in the objective function $J(\boldsymbol{s})$ do not coincide, although the difference is more prominent with the fourth $b^{(4)}(\boldsymbol{x}, \boldsymbol{s})$ and fifth $b^{(5)}(\boldsymbol{x}, \boldsymbol{s})$ source terms (Figure 12). Even though the variance in the second source term $b^{(2)}(\boldsymbol{x}, \boldsymbol{s})$ is small, a relatively large variance in the objective $J(\boldsymbol{s})$ is obtained, since the variance in the design variables s_1 and s_2 contribute most to variance in the objective (as supported by ARD).

Large COD values are indicative of similarity between the true $p(\boldsymbol{J})$ and approximate (RDVGP) $q(\boldsymbol{J}_*)$ marginal posterior probability densities. In comparison to the COD values for the mean R_μ^2, smaller COD values R_σ^2 are observed for the variance of both the objective $J(\boldsymbol{s})$ and constraint $H(\boldsymbol{s})$ functions. This is due to the presence of epistemic uncertainty in the approximate (RDVGP) marginal posterior probability density $q(\boldsymbol{J}_*)$, which increases the total uncertainty when compared to the true density $p(\boldsymbol{J})$. However, COD values greater than 0.7 indicate the RDVGP surrogate emulates the true objective $J(\boldsymbol{s})$ and constraint $H(\boldsymbol{s})$ functions well. Larger COD values for the standard deviation are observed where variance in the source $b(\boldsymbol{x}, \boldsymbol{s})$ is low. The relatively low MMD D_{MMD} values (as computed using (46)) show the similarity between the true $p(J(\langle \bar{\boldsymbol{s}}_d^* \rangle))$ and approximate (RDVGP) $q(J(\langle \bar{\boldsymbol{s}}_d^* \rangle))$ marginal posterior probability densities.

Next, we investigate the effect of the weighting factor α prescribed for the RDO problem (1). Depending on the choice of weighting factor, the robust optimum design variables $\bar{\boldsymbol{s}}_d^*$ predicted from the RDVGP

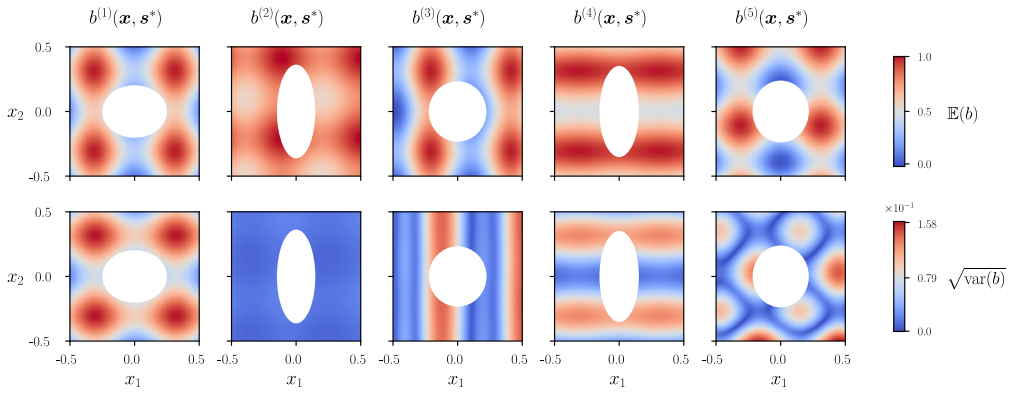


Figure 10: Plate with elliptical hole. Optimised geometries for different source terms $b^{(j)}(\boldsymbol{x}, \boldsymbol{s}^*)$ (with $j \in \{1, 2, ..., 5\}$), and the respective empirical expectation $\mathbb{E}(b(\boldsymbol{x}, \boldsymbol{s}^*))$ and standard deviation $\sqrt{\text{var}(b(\boldsymbol{x}, \boldsymbol{s}^*))}$.

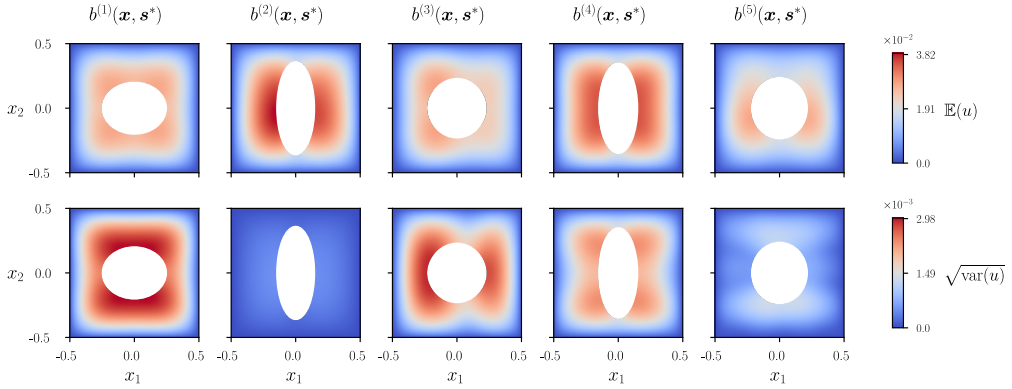


Figure 11: Plate with elliptical hole. Optimised geometries with solution fields $u(\boldsymbol{x}, \boldsymbol{s}^*)$ for each source term $b^{(j)}(\boldsymbol{x}, \boldsymbol{s})$ (with $j \in \{1, 2, ..., 5\}$), and the respective empirical expectation $\mathbb{E}(u(\boldsymbol{x}, \boldsymbol{s}^*))$ and standard deviation $\sqrt{\text{var}(u(\boldsymbol{x}, \boldsymbol{s}^*))}$.

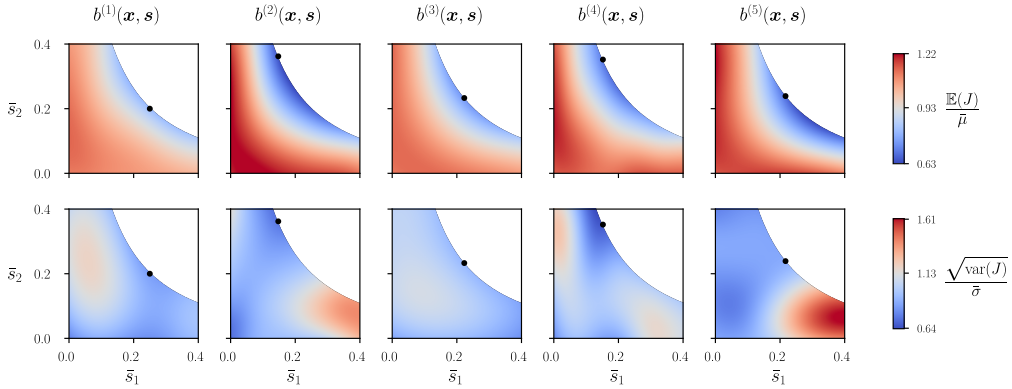


Figure 12: Plate with elliptical hole. Marginal posterior probability density $q(\boldsymbol{J}_*)$ showing the empirical normalised expectation $\mathbb{E}(J(\bar{\boldsymbol{s}}_d))/\bar{\mu}$ and standard deviation $\sqrt{\text{var}(J(\bar{\boldsymbol{s}}_d))}/\bar{\sigma}$ for each of the source terms $b^{(j)}(\boldsymbol{x}, \boldsymbol{s})$ (with $j \in \{1, 2, ..., 5\}$), showing the approximate (RDVGP) robust optimum design variables $\bar{\boldsymbol{s}}_d^*$ (black dot).

Table 2: Plate with elliptical hole. Prescribed mean $\bar{s}_i$ and standard deviation σ_i of the random immutable variables s_i (with $i \in \{3, 4, ..., 7\}$) of the j^{th} source term $b^{(j)}(\boldsymbol{x}, \boldsymbol{s})$ given by (51).

$b^{(j)}(\boldsymbol{x}, \boldsymbol{s})$	$\bar{s}_3$	σ_3	$\bar{s}_4$	σ_4	$\bar{s}_5$	σ_5	$\bar{s}_6$	σ_6	$\bar{s}_7$	σ_7
$b^{(1)}(\boldsymbol{x}, \boldsymbol{s})$	0.55	0.060	0.00	0.001	0.35	0.080	0.00	0.001	0.10	0.010
$b^{(2)}(\boldsymbol{x}, \boldsymbol{s})$	0.20	0.005	0.20	0.001	0.20	0.005	0.10	0.001	0.60	0.010
$b^{(3)}(\boldsymbol{x}, \boldsymbol{s})$	0.65	0.100	0.10	0.020	0.25	0.001	0.00	0.001	0.10	0.001
$b^{(4)}(\boldsymbol{x}, \boldsymbol{s})$	0.05	0.010	0.00	0.001	0.45	0.100	0.00	0.001	0.50	0.010
$b^{(5)}(\boldsymbol{x}, \boldsymbol{s})$	0.45	0.010	0.00	0.025	0.45	0.010	0.20	0.025	0.10	0.010

Table 3: Plate with elliptical hole. Robust optimum design variables obtained using the RDVGP surrogate $(\bar{\boldsymbol{s}}_d^*)^\mathsf{T} = (\bar{s}_1^* \ \bar{s}_2^*)$ and MC sampling $\langle \bar{\boldsymbol{s}}_d^* \rangle^\mathsf{T} = (\langle \bar{s}_1^* \rangle, \langle \bar{s}_2^* \rangle)$, compared to the deterministic prediction $(\hat{\boldsymbol{s}}_d^*)^\mathsf{T} = (\hat{s}_1^* \ \hat{s}_2^*)$, with COD performance metrics (45) for the objective $J(\boldsymbol{s})$ $\left(\left(R_\mu^J\right)^2, \left(R_\sigma^J\right)^2\right)$ and constraint $H(\boldsymbol{s})$ $\left(\left(R_\mu^H\right)^2, \left(R_\sigma^H\right)^2\right)$ functions, and MMD D_{MMD} between the true $p(J(\langle \bar{\boldsymbol{s}}_d^* \rangle))$ and approximate (RDVGP) $q(J(\langle \bar{\boldsymbol{s}}_d^* \rangle))$ marginal posterior probability densities.

$b^{(j)}(\boldsymbol{x}, \boldsymbol{s})$	d_z	$\left(R_\mu^J\right)^2$	$\left(R_\sigma^J\right)^2$	$\left(R_\mu^H\right)^2$	$\left(R_\sigma^H\right)^2$	D_{MMD}	$(\hat{\boldsymbol{s}}_d^*)^\mathsf{T}$	$(\bar{\boldsymbol{s}}_d^*)^\mathsf{T}$	$\langle \bar{\boldsymbol{s}}_d^* \rangle^\mathsf{T}$
$b^{(1)}(\boldsymbol{x}, \boldsymbol{s})$	4	0.957	0.821	0.997	0.924	0.000307	(0.326 0.196)	(0.250 0.203)	(0.270 0.183)
$b^{(2)}(\boldsymbol{x}, \boldsymbol{s})$	3	0.996	0.925	0.998	0.941	0.000434	(0.223 0.286)	(0.148 0.362)	(0.159 0.336)
$b^{(3)}(\boldsymbol{x}, \boldsymbol{s})$	4	0.959	0.801	0.995	0.915	0.000345	(0.269 0.237)	(0.223 0.233)	(0.228 0.227)
$b^{(4)}(\boldsymbol{x}, \boldsymbol{s})$	3	0.967	0.885	0.987	0.931	0.000314	(0.203 0.316)	(0.152 0.352)	(0.153 0.350)
$b^{(5)}(\boldsymbol{x}, \boldsymbol{s})$	4	0.997	0.744	0.992	0.902	0.000074	(0.397 0.161)	(0.217 0.239)	(0.196 0.268)

surrogate may not converge to the true solution $\langle \bar{\boldsymbol{s}}_d^* \rangle$. Robust optimum design variables $\bar{\boldsymbol{s}}_d^*$ predicted using large weighting factors α are unreliable (Figure 13), since the weighted sum being minimised is predominantly influenced by the variance, exacerbating the effect of the large epistemic uncertainty. The epistemic uncertainty is large when the sample size n in the training data set $\mathcal{D}$ is small, with confounding between the aleatoric and epistemic uncertainty causing discrepancy between the true $p(\boldsymbol{J})$ and approximate (RDVGP) $q(\boldsymbol{J}_*)$ marginal posterior probability densities. As expected, increasing the size n of the training data set $\mathcal{D}$ reduces the epistemic uncertainty and increases the reliability of the approximate (RDVGP) robust optimum design variables $\bar{\boldsymbol{s}}_d^*$ (Figure 13). This emphasises the importance of using a training data set $\mathcal{D}$ with a sufficiently large sample size n when using surrogates to emulate the objective $J(\boldsymbol{s})$ and constraint $H^{(j)}(\boldsymbol{s})$ functions in RDO. Overestimating the uncertainty may result in a sub-optimal choice of design variables for complex engineering systems.

4.4. Bracket

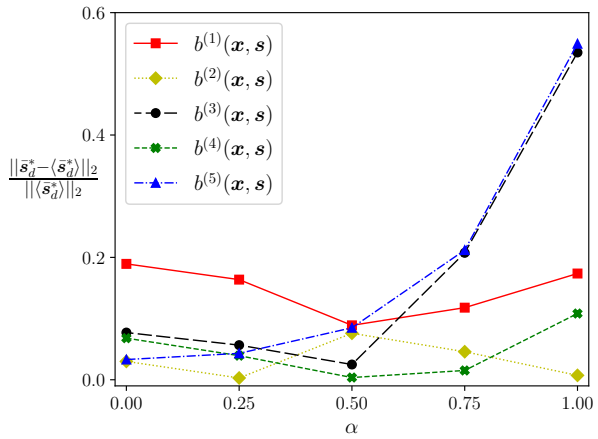
Next, we demonstrate the use of the RDVGP surrogate on the RDO of a bracket. The three-dimensional equations of elasticity are given by

$$-\nabla \cdot \boldsymbol{\sigma}_s(\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{s})) = \boldsymbol{b}(\boldsymbol{x}, \boldsymbol{s}), \qquad \boldsymbol{x} \in \Omega, \tag{54a}$$

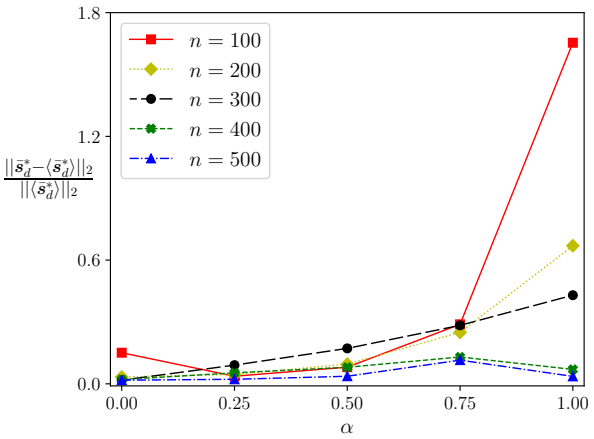
$$\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{s}) = \boldsymbol{0}, \qquad \boldsymbol{x} \in \partial\Omega_D, \tag{54b}$$

$$\boldsymbol{\sigma}_s(\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{s})) \cdot \boldsymbol{n} = \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{s}), \qquad \boldsymbol{x} \in \partial\Omega_N, \tag{54c}$$

where $\boldsymbol{\sigma}_s(\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{s})) \in \mathbb{R}^{3\times3}$ is the stress tensor, $\boldsymbol{b}(\boldsymbol{x}, \boldsymbol{s}) \in \mathbb{R}^3$ is the body force vector, $\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{s}) \in \mathbb{R}^3$ is the displacement vector, and $\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{s}) \in \mathbb{R}^3$ is the applied surface traction at the Neumann boundary $\partial\Omega_N$. For plane stress deformation, the stress tensor $\boldsymbol{\sigma}_s(\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{s}))$ consists of out-of-plane components $\sigma_{13} = \sigma_{23} = \sigma_{33} = 0$,



(a) Optimum solution $\bar{\boldsymbol{s}}_s^*$ convergence on $b^{(j)}(\boldsymbol{x}, \boldsymbol{s})$



(b) Optimum solution $\bar{\boldsymbol{s}}_s^*$ convergence on n

Figure 13: Plate with elliptical hole. Effect of the (a) weighting factor α for different source terms $b^{(j)}(\boldsymbol{x}, \boldsymbol{s})$ (with $j \in \{1, 2, ..., 5\}$) on the convergence (as measured by the relative error) between the approximate (RDVGP) $\bar{\boldsymbol{s}}_d^*$ and true $\langle \bar{\boldsymbol{s}}_d^* \rangle$ robust optimum design variables. The effect of (b) the weighting factor α and size n of the training data set (assuming $m = n/2$) for the fifth source term $b^{(5)}(\boldsymbol{x}, \boldsymbol{s})$ on convergence.

and in-plane components given by the constitutive relation (Hooke's law)

$$\sigma_{ij} = \frac{E}{1+\upsilon}\left(\varepsilon_{ij} + \frac{\upsilon}{1-\upsilon}\,\mathrm{Tr}\left(\boldsymbol{\varepsilon}\left(\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{s})\right)\right)\delta_{ij}\right), \qquad i, j \in \{1, 2\}, \tag{55}$$

with the strain tensor given by

$$\boldsymbol{\varepsilon}\left(\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{s})\right) = \frac{1}{2}\left(\nabla \boldsymbol{u}(\boldsymbol{x}, \boldsymbol{s}) + \left(\nabla \boldsymbol{u}(\boldsymbol{x}, \boldsymbol{s})\right)^{\mathsf{T}} + \left(\nabla \boldsymbol{u}(\boldsymbol{x}, \boldsymbol{s})\right)^{\mathsf{T}} \nabla \boldsymbol{u}(\boldsymbol{x}, \boldsymbol{s})\right), \tag{56}$$

where $E, \upsilon \in \mathbb{R}^+$ are the Young's modulus and Poisson's ratio respectively, δ_{ij} is the Kronecker delta function, and $\boldsymbol{\varepsilon}(\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{s})) \in \mathbb{R}^{3\times 3}$ is the strain tensor with components ε_{ij} (with $i, j \in \{1, 2, 3\}$). As shown in Figure 14, the bracket has an outer boundary $\partial\Omega_o$, a polygonal void with boundary $\partial\Omega_v$, and a circular hole with boundary $\partial\Omega_c$. The Dirichlet boundary condition is applied along the left edge $\partial\Omega_D$, with the Neumann boundary condition applied along the remaining boundary $\partial\Omega_N$ where $\partial\Omega_N = \partial\Omega_o \cup \partial\Omega_v \cup \partial\Omega_c$.

The domain Ω is discretised over spatial coordinates $\boldsymbol{x} = (x_1\ x_2\ x_3)^{\mathsf{T}}$, and is parameterised by five design variables $\boldsymbol{s}_d = (s_1\ s_2\ ...\ s_5)^{\mathsf{T}}$ and ten immutable variables $\boldsymbol{s}_f = (s_6\ s_7\ ...\ s_{15})^{\mathsf{T}}$ controlling the shape of the domain Ω, and the magnitude s_{14} and angle s_{15} of the resultant applied load. The body force $\boldsymbol{b}(\boldsymbol{x}, \boldsymbol{s})$ is constant over the geometry domain, where $\boldsymbol{b}(\boldsymbol{x}, \boldsymbol{s}) = (0\ \ -7.7 \times 10^{-5}\ \ 0)^{\mathsf{T}}$. The surface traction $\boldsymbol{g}(\boldsymbol{x}, \boldsymbol{s})$ yields the resultant applied load (applied as uniformly distributed) along the circular hole boundary $\partial\Omega_c$ where $\int_{\partial\Omega_c} \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{s})\, d\boldsymbol{x} = (s_{14}\cos(s_{15})\ \ s_{14}\sin(s_{15})\ \ 0)^{\mathsf{T}}$. All random input variables $s_i \in \boldsymbol{s}$ are Gaussian, where $s_i \sim \mathcal{N}(\bar{s}_i, \sigma_i^2)$ and $\bar{s}_i \in \mathbb{R}$ is a fixed constant for $i \in \{6, 7, ..., 15\}$, and box-constrained on the interval $\bar{s}_i \in \{\bar{s}_i \in \mathbb{R} \mid \bar{s}_i^{(l)} \leq \bar{s}_i \leq \bar{s}_i^{(u)}\}$ for $i \in \{1, 2, ..., 5\}$ where $\bar{s}_i^{(l)}$ and $\bar{s}_i^{(u)}$ are the lower and upper limits respectively.

The goal is to compute the robust optimum design variables $\bar{\boldsymbol{s}}_d^* = (\bar{s}_1^*\ \bar{s}_2^*\ ...\ \bar{s}_5^*)^{\mathsf{T}}$ (1) that minimise structural compliance

$$J(\boldsymbol{s}) = \int_\Omega \boldsymbol{\sigma}_s\left(\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{s})\right) : \boldsymbol{\varepsilon}\left(\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{s})\right) d\boldsymbol{x}, \tag{57}$$

subject to the volume constraint function

$$H^{(1)}(\boldsymbol{s}) = \int_{\Omega} d\boldsymbol{x} - V_0, \tag{58}$$

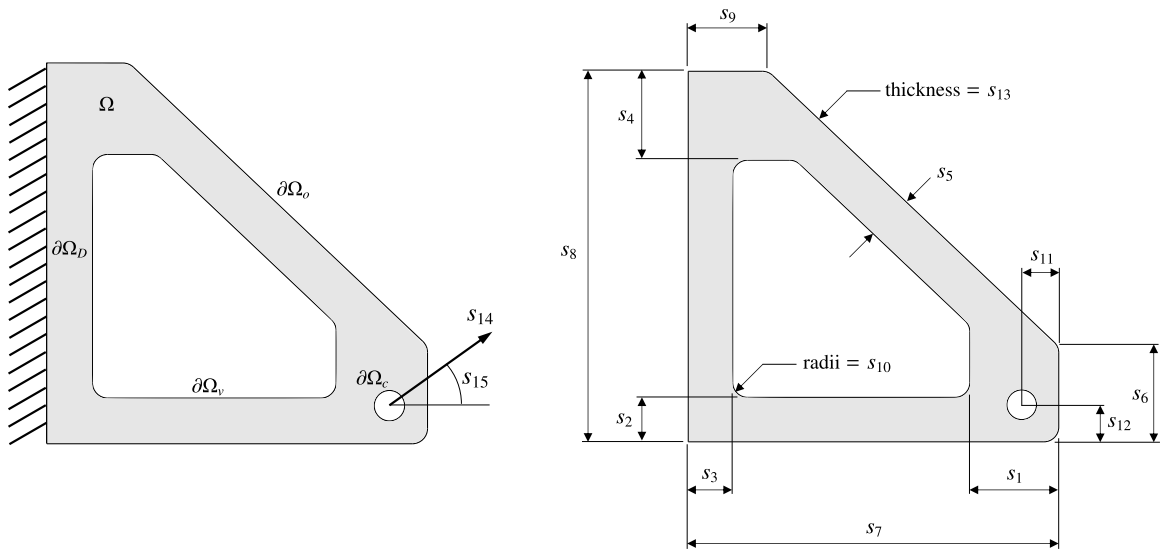
where $V_0 = 1.25 \times 10^5$ is the chosen limit on the volume. A constraint is also applied on the stress (von Mises), where

$$H^{(2)}(\boldsymbol{s}) = \max_{x \in \Omega} \sigma_v(\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{s})) - \sigma_{v0}, \tag{59}$$

$\sigma_v(\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{s}))$ is the von Mises stress, and $\sigma_{v0} = 125$ is the upper limit on the stress. According to the RDO formulation (1), a weighting factor of $\alpha = 0.5$ and feasibility indices of $\beta^{(1)} = \beta^{(2)} = 2$ are used. The standard deviation in the maximum von Mises stress $\sigma_v(\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{s}))$ must also be less than 35.

The training data set $\mathcal{D}$ is initialised using LHS with $n = 500$ samples and $m = 250$ inducing points. The objective $J(\boldsymbol{s})$ and stress constraint $H^{(2)}(\boldsymbol{s})$ are sampled from a deterministic FE model (Figure 15). To avoid sampling stress singularities, the maximum von Mises stress $\max \sigma_v(\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{s}))$ is not sampled within a distance of 5 from the Dirichlet $\partial\Omega_D$ and Neumann $\partial\Omega_N$ boundary condition surfaces. The bracket is modelled as a thin plate, using the plane stress assumption. A Young's modulus of $E = 2.1 \times 10^5$ and Poisson's ratio of $\upsilon = 0.3$ are used. The mean and standard deviation of the immutable variables $\boldsymbol{s}_f$ are prescribed. Similarly, the lower and upper bounds on the mean, and standard deviation of the design variables $\boldsymbol{s}_d$ are also prescribed (Table 4). Normalisation constants $\bar{\mu}$ and $\bar{\sigma}$ in the RDO weighted sum (1) are set equal to the mean and standard deviation of the objective function $J(\boldsymbol{s})$ observations.

The robust optimum design variables $\bar{\boldsymbol{s}}_d^*$ computed using the RDVGP surrogate are compared to the true



(a) Domain with boundary conditions

(b) Bracket geometry

Figure 14: Bracket. Schematic showing (a) the domain of the bracket with a Dirichlet boundary condition applied on the left edge $\partial\Omega_D$, the Neumann boundary condition applied along the union of the outer boundary $\partial\Omega_o$, void boundary $\partial\Omega_v$, and circular hole boundary $\partial\Omega_c$ and (b) bracket geometry with random design variables $\boldsymbol{s}_d = (s_1\ s_2\ \ldots\ s_5)^\mathsf{T}$ and random immutable variables $\boldsymbol{s}_f = (s_6\ s_7\ \ldots\ s_{15})^\mathsf{T}$.

optimum $\langle \bar{\boldsymbol{s}}_d^* \rangle$ evaluated using MC sampling with 10^5 samples. A total of $n_v = 30$ validation samples are used for computing COD values given by (45).

The immutable variables $\boldsymbol{s}_f$ with negligible variance when propagated through the objective $J(\boldsymbol{s})$ and constraint $H^{(j)}(\boldsymbol{s})$ functions, contribute insignificantly to variance in the observations. This is a reasonable representation of many complex engineering systems manufactured with high-precision equipment. As expected, the subspace is primarily composed of a linear combination of the design variables $\boldsymbol{s}_d$ and the magnitude s_{14} and angle s_{15} of the applied load, where the projection matrix $\boldsymbol{W}$ has dimension $d_z = 7$.

The true robust optimum design variables $\langle \bar{\boldsymbol{s}}_d^* \rangle = (88.6\ 41.1\ 10.5\ 69.6\ 29.2)^\mathsf{T}$(as estimated with MC sampling) are well approximated by the robust optimum design variables $\bar{\boldsymbol{s}}_d^* = (84.8\ 40.2\ 10.8\ 75.2\ 28.5)^\mathsf{T}$ proposed by performing RDO (1) with the RDVGP surrogate. Due to the uncertainty in the magnitude s_{14} and angle s_{15} of the applied load (Figure 14), the marginal posterior probability densities for the compliance $q(\boldsymbol{J}_*)$ and the stress constraint $q(\boldsymbol{H}_*^{(2)})$ yield a large aleatoric uncertainty prediction (Figure 16). Consequently, permissible COD values are obtained for the mean R_μ^2 (0.801,0.907) and variance R_σ^2 (0.757,0.716) of the objective $J(\boldsymbol{s})$ and stress constraint $H^{(2)}(\boldsymbol{s})$ functions respectively. Since the aleatoric uncertainty in the volume constraint $H^{(1)}(\boldsymbol{s})$ is negligible due to the small standard deviation in the random input variables $\boldsymbol{s}$ (Table 4), the epistemic uncertainty constitutes most of the total variance, and the RDVGP surrogate be-

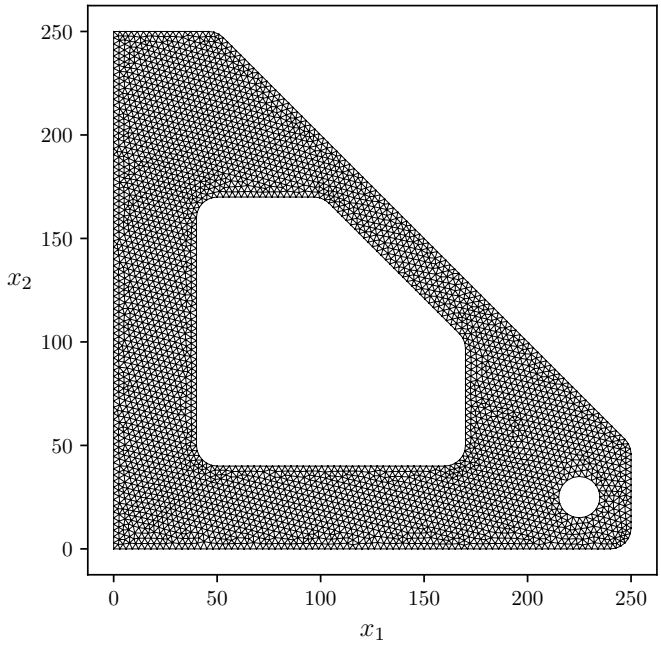


Figure 15: Bracket. FE mesh of an example geometry containing 7668 linear triangle elements, discretised over spatial coordinates $\boldsymbol{x} = (x_1\ x_2\ x_3)^\mathsf{T}$ according to the plane stress assumption.

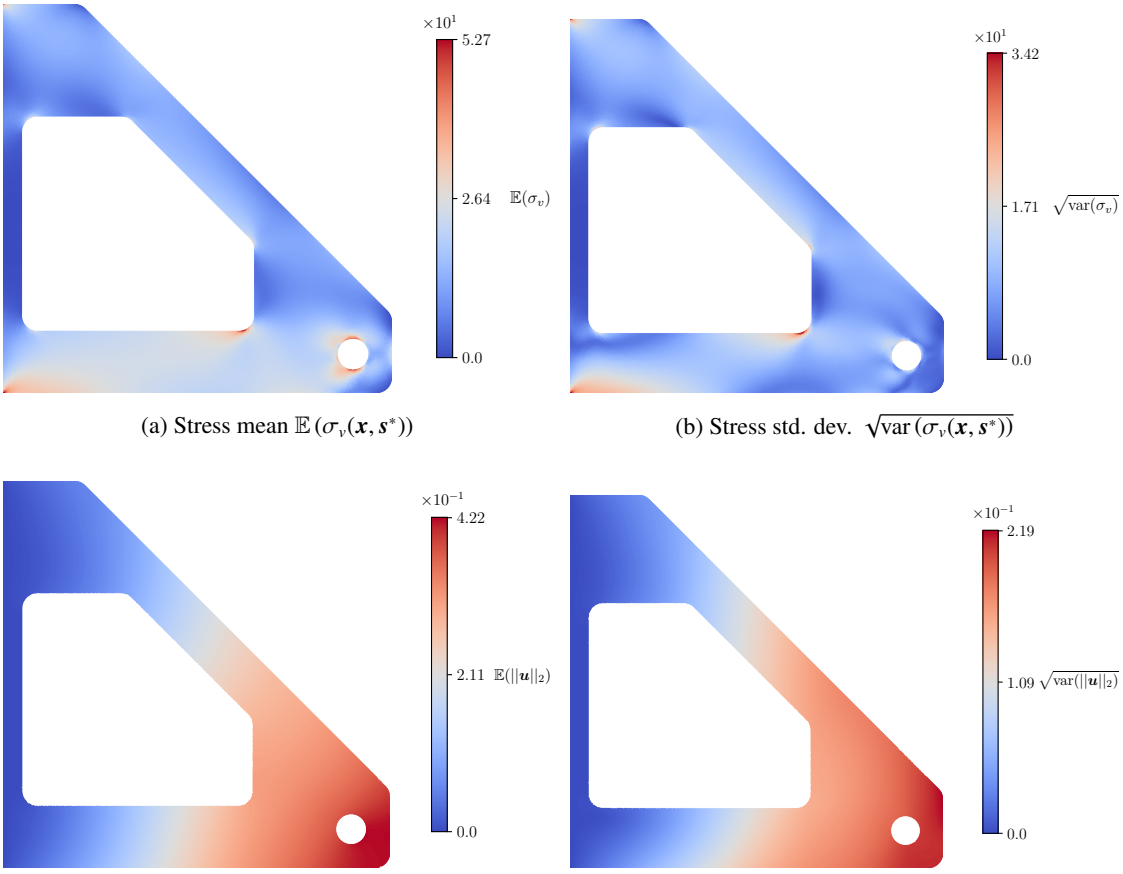
Table 4: Bracket. Mean $\bar{s}_i$ and standard deviation σ_i of the random input variables $\boldsymbol{s} = (s_1\ s_2\ \ldots\ s_{15})^\mathsf{T}$, where box constraints $\bar{s}_i \in [\bar{s}_i^{(l)}, \bar{s}_i^{(u)}]$ are given for the design variables $\boldsymbol{s}_d = (s_1\ s_2\ \ldots\ s_5)^\mathsf{T}$.

	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}	s_{11}	s_{12}	s_{13}	s_{14}	s_{15}
$\bar{s}_i$	[60,100]	[10,50]	[10,50]	[60,100]	[0,30]	50	250	250	50	10	25	25	5	5000	0
σ_i	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	1000	$\pi/6$

comes a poor predictor of the aleatoric uncertainty in the volume constraint $H^{(1)}(\boldsymbol{s})$. This is demonstrated by COD values of 0.864 and 0.331 for the mean and variance respectively. The probability density near the robust optimum design variables is accurate, as demonstrated by an MMD D_{MMD} (computed using (46)) of 0.000983 between the true $p(J(\langle\bar{\boldsymbol{s}}_d^*\rangle))$ and approximate (RDVGP) $q(J(\langle\bar{\boldsymbol{s}}_d^*\rangle))$ marginal posterior probability densities. This example is a good demonstration of the RDVGP surrogate accuracy in uncertainty quantification with application to RDO.

5. Conclusion

We introduced the RDVGP surrogate, which presents a statistical surrogate that can be used to emulate the quantities of interest in complex engineering systems, for application in solving RDO problems. Due to the availability of only a finite number of deterministic PDE observations that may depend on high-dimensional random input variables, we generalised a standard GP by postulating a graphical model with intrinsic assumptions relating the random input variables and observations to latent variables sampled from a low-dimensional subspace. Both the surrogate hyperparameters and the linear projection matrix describ-



(a) Stress mean $\mathbb{E}(\sigma_v(\boldsymbol{x}, \boldsymbol{s}^*))$

(b) Stress std. dev. $\sqrt{\text{var}(\sigma_v(\boldsymbol{x}, \boldsymbol{s}^*))}$

(c) Displacement mean $\mathbb{E}(\|\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{s}^*)\|_2)$

(d) Displacement std. dev. $\sqrt{\text{var}(\|\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{s}^*)\|_2)}$

Figure 16: Bracket. Deformation showing for the optimum geometry $\boldsymbol{s}^* = ((\bar{\boldsymbol{s}}_d^*)^\mathsf{T}\ \boldsymbol{s}_f^\mathsf{T})^\mathsf{T}$ (a) the von Mises stress mean $\mathbb{E}(\sigma_v(\boldsymbol{x}, \boldsymbol{s}^*))$, (b) the von Mises stress standard deviation $\sqrt{\text{var}(\sigma_v(\boldsymbol{x}, \boldsymbol{s}^*))}$, (c) the displacement mean $\mathbb{E}(\|\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{s}^*)\|_2)$, and (d) the displacement standard deviation $\sqrt{\text{var}(\|\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{s}^*)\|_2)}$.

ing the subspace are learned jointly by optimising a single log marginal likelihood formulation, which is computed using VB. The computational simplicity of the RDVGP surrogate provides a favourable alternative to other algorithmic approaches, which do not consider input variable uncertainty when training a standard GP, leading to over-fitting.

Compared to the conventional approach of fitting a standard GP and computing the marginal posterior probability density by MC sampling over the prescribed input variable probability density, the RDVGP surrogate is more accurate while remaining efficient, since the induced sparsity reduces the computational expense of sampling over the random input variables. The approximate marginal posterior probability density predicted using the RDVGP surrogate sufficiently emulates the true marginal posterior probability density, yielding accurate robust optimum design variables in RDO applications.

In closing, we mention promising research extensions to the proposed approach. The statistical surrogate model is implemented using a GP; however, since our approach uses a graphical model, this can be further generalised to consider alternative surrogates such as Bayesian physics-informed deep neural networks [66, 67], extending to grey-box applications [68]. The linear subspace could also be discovered using Riemmannian manifolds that retain knowledge of intrinsic geometric constraints, for use in geometry-aware Bayesian optimisation [69]. The efficiency, accuracy, and scalability of the proposed model makes it promising for use in the RDO of complex engineering systems, such as construction-aware optimisation of concrete-based structures [70]. Although we assumed a linear dimensionality reduction in the statistical observation model, non-linear dimensionality reduction [71] may further improve efficiency and accuracy for some types of problems. The use of the RDVGP surrogate when used with adaptive sampling algorithms such as Bayesian optimisation, is also a promising direction for further research.

Appendix A : Derivation of the KL divergence terms

The KL divergence terms in the ELBO $\mathcal{F}(\boldsymbol{y})$ given by (36) are analytically tractable and can be computed in exact form. For example, consider the second KL divergence term

$$
\begin{aligned}
D_{KL}\left(q_\psi(\mathbf{Z}) \,\|\, p_W(\mathbf{Z})\right) &= \int q_\psi(\mathbf{Z}) \ln\left(\frac{q_\psi(\mathbf{Z})}{p_W(\mathbf{Z})}\right) d\mathbf{Z} \\
&= \sum_{i=1}^{n} \mathbb{E}_{q_\psi(z_i)}\left(\ln q_\psi(\boldsymbol{z}_i) - \ln p_W(\boldsymbol{z}_i)\right) \\
&= \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}_{q_\psi(z_i)}\left(\ln\left(\frac{|\boldsymbol{W}^\mathsf{T}\boldsymbol{\Sigma}_s\boldsymbol{W}|}{|\tilde{\boldsymbol{\Sigma}}_{z,i}|}\right) + \left(\boldsymbol{z}_i - \boldsymbol{W}^\mathsf{T}\bar{\boldsymbol{s}}_i\right)^\mathsf{T}\left(\boldsymbol{W}^\mathsf{T}\boldsymbol{\Sigma}_s\boldsymbol{W}\right)^{-1}\left(\boldsymbol{z}_i - \boldsymbol{W}^\mathsf{T}\bar{\boldsymbol{s}}_i\right) - \left(\boldsymbol{z}_i - \tilde{\boldsymbol{\mu}}_{z,i}\right)^\mathsf{T}\left(\tilde{\boldsymbol{\Sigma}}_{z,i}\right)^{-1}\left(\boldsymbol{z}_i - \tilde{\boldsymbol{\mu}}_{z,i}\right)\right) \\
&= \frac{1}{2}\sum_{i=1}^{n}\left(\ln\left(\frac{|\boldsymbol{W}^\mathsf{T}\boldsymbol{\Sigma}_s\boldsymbol{W}|}{|\tilde{\boldsymbol{\Sigma}}_{z,i}|}\right) + \mathbb{E}_{q_\psi(z_i)}\left(\mathrm{Tr}\left(\left(\boldsymbol{W}^\mathsf{T}\boldsymbol{\Sigma}_s\boldsymbol{W}\right)^{-1}\left(\boldsymbol{z}_i - \boldsymbol{W}^\mathsf{T}\bar{\boldsymbol{s}}_i\right)\left(\boldsymbol{z}_i - \boldsymbol{W}^\mathsf{T}\bar{\boldsymbol{s}}_i\right)^\mathsf{T}\right) - \mathrm{Tr}\left(\left(\tilde{\boldsymbol{\Sigma}}_{z,i}\right)^{-1}\left(\boldsymbol{z}_i - \tilde{\boldsymbol{\mu}}_{z,i}\right)\left(\boldsymbol{z}_i - \tilde{\boldsymbol{\mu}}_{z,i}\right)^\mathsf{T}\right)\right)\right) \\
&= \frac{1}{2}\sum_{i=1}^{n}\left(\ln\left(\frac{|\boldsymbol{W}^\mathsf{T}\boldsymbol{\Sigma}_s\boldsymbol{W}|}{|\tilde{\boldsymbol{\Sigma}}_{z,i}|}\right) + \mathbb{E}_{q_\psi(z_i)}\left(\mathrm{Tr}\left(\left(\boldsymbol{W}^\mathsf{T}\boldsymbol{\Sigma}_s\boldsymbol{W}\right)^{-1}\left(\boldsymbol{z}_i\boldsymbol{z}_i^\mathsf{T} - 2\boldsymbol{z}_i\bar{\boldsymbol{s}}_i^\mathsf{T}\boldsymbol{W} + \boldsymbol{W}^\mathsf{T}\bar{\boldsymbol{s}}_i\bar{\boldsymbol{s}}_i^\mathsf{T}\boldsymbol{W}\right)\right) - \mathrm{Tr}\left(\left(\tilde{\boldsymbol{\Sigma}}_{z,i}\right)^{-1}\tilde{\boldsymbol{\Sigma}}_{z,i}\right)\right)\right) \\
&= \frac{1}{2}\sum_{i=1}^{n}\left(\ln\left(\frac{|\boldsymbol{W}^\mathsf{T}\boldsymbol{\Sigma}_s\boldsymbol{W}|}{|\tilde{\boldsymbol{\Sigma}}_{z,i}|}\right) - d_z + \mathrm{Tr}\left(\left(\boldsymbol{W}^\mathsf{T}\boldsymbol{\Sigma}_s\boldsymbol{W}\right)^{-1}\left(\tilde{\boldsymbol{\Sigma}}_{z,i} + \tilde{\boldsymbol{\mu}}_{z,i}\left(\tilde{\boldsymbol{\mu}}_{z,i}\right)^\mathsf{T} - 2\boldsymbol{W}^\mathsf{T}\bar{\boldsymbol{s}}_i\left(\tilde{\boldsymbol{\mu}}_{z,i}\right)^\mathsf{T} + \boldsymbol{W}^\mathsf{T}\bar{\boldsymbol{s}}_i\bar{\boldsymbol{s}}_i^\mathsf{T}\boldsymbol{W}\right)\right)\right) \\
&= \frac{1}{2}\sum_{i=1}^{n}\left(\ln\left(\frac{|\boldsymbol{W}^\mathsf{T}\boldsymbol{\Sigma}_s\boldsymbol{W}|}{|\tilde{\boldsymbol{\Sigma}}_{z,i}|}\right) - d_z + \mathrm{Tr}\left(\left(\boldsymbol{W}^\mathsf{T}\boldsymbol{\Sigma}_s\boldsymbol{W}\right)^{-1}\tilde{\boldsymbol{\Sigma}}_{z,i}\right) + \left(\boldsymbol{W}^\mathsf{T}\bar{\boldsymbol{s}}_i - \tilde{\boldsymbol{\mu}}_{z,i}\right)^\mathsf{T}\left(\boldsymbol{W}^\mathsf{T}\boldsymbol{\Sigma}_s\boldsymbol{W}\right)^{-1}\left(\boldsymbol{W}^\mathsf{T}\bar{\boldsymbol{s}}_i - \tilde{\boldsymbol{\mu}}_{z,i}\right)\right).
\end{aligned}
\tag{A.1}
$$

and after following a similar derivation for the first KL divergence term

$$
D_{KL}\left(q_\omega(\tilde{\boldsymbol{f}}) \,\|\, p_\theta(\tilde{\boldsymbol{f}})\right) = \frac{1}{2}\left(\ln\left(\frac{|\boldsymbol{C}_{\tilde{Z}\tilde{Z}}|}{|\tilde{\boldsymbol{\Sigma}}_{\tilde{f}}|}\right) - m + \mathrm{Tr}\left(\boldsymbol{C}_{\tilde{Z}\tilde{Z}}^{-1}\tilde{\boldsymbol{\Sigma}}_{\tilde{f}}\right) + \left(\tilde{\boldsymbol{\mu}}_{\tilde{f}}\right)^\mathsf{T}\boldsymbol{C}_{\tilde{Z}\tilde{Z}}^{-1}\tilde{\boldsymbol{\mu}}_{\tilde{f}}\right). \tag{A.2}
$$

Appendix B: Multiple observation vectors

The d_y observation vectors and corresponding target output variables can be collected into the matrices $\boldsymbol{Y} \in \mathbb{R}^{n\times d_y}$ and $\boldsymbol{F} \in \mathbb{R}^{n\times d_y}$ respectively, where $\boldsymbol{Y} = \left(\boldsymbol{y}^{(1)}\ \boldsymbol{y}^{(2)}\ \ldots\ \boldsymbol{y}^{(d_y)}\right)$ and $\boldsymbol{F} = \left(\boldsymbol{f}^{(1)}\ \boldsymbol{f}^{(2)}\ \ldots\ \boldsymbol{f}^{(d_y)}\right)$. The likelihood

$$p_{\sigma_y}(\boldsymbol{Y}|\boldsymbol{F}) = \prod_{i=1}^{d_y} p_{\sigma_y}\left(\boldsymbol{y}^{(i)}\middle|\boldsymbol{f}^{(i)}\right) = \prod_{i=1}^{d_y} \mathcal{N}\left(\boldsymbol{f}^{(i)}, \sigma_y^{(i)}\boldsymbol{I}\right), \tag{B.1}$$

and GP prior probability density

$$p_\theta(\boldsymbol{F}|\boldsymbol{Z}) = \prod_{i=1}^{d_y} p_\theta\left(\boldsymbol{f}^{(i)}\middle|\boldsymbol{Z}\right) = \prod_{i=1}^{d_y} \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{C}_{ZZ}^{(i)}\right), \tag{B.2}$$

can be expanded using the mean field approximation. Similarly, the ELBO $\mathcal{F}(\boldsymbol{y})$ in (28) for the statistical latent variable model may be expanded as the summation of the ELBO over each observation vector

$$\mathcal{F}(\boldsymbol{y}) = \sum_{i=1}^{d_y} \mathbb{E}_{q_\psi(\boldsymbol{Z})}\left(\mathbb{E}_{p_\theta(\boldsymbol{f}^{(i)}|\boldsymbol{Z})}\left(\ln p_{\sigma_y}\left(\boldsymbol{y}^{(i)}\middle|\boldsymbol{f}^{(i)}\right)\right)\right) - D_{KL}\left(q_\psi(\boldsymbol{Z})\|p_W(\boldsymbol{Z})\right). \tag{B.3}$$

Similarly for the sparse statistical latent variable model, the sparse GP prior

$$p_\theta\left(\boldsymbol{F}|\tilde{\boldsymbol{F}}, \boldsymbol{Z}\right) = \prod_{i=1}^{d_y} p_\theta\left(\boldsymbol{f}^{(i)}|\tilde{\boldsymbol{f}}^{(i)}, \boldsymbol{Z}\right) = \prod_{i=1}^{d_y} \mathcal{N}\left(\boldsymbol{C}_{Z\tilde{Z}}^{(i)}\left(\boldsymbol{C}_{\tilde{Z}\tilde{Z}}^{(i)}\right)^{-1}\tilde{\boldsymbol{f}}^{(i)},\ \boldsymbol{C}_{ZZ}^{(i)} - \boldsymbol{C}_{Z\tilde{Z}}^{(i)}\left(\boldsymbol{C}_{\tilde{Z}\tilde{Z}}^{(i)}\right)^{-1}\boldsymbol{C}_{\tilde{Z}Z}^{(i)}\right), \tag{B.4}$$

and trial probability density

$$q_\omega\left(\tilde{\boldsymbol{F}}\right) := \prod_{i=1}^{d_y} q_\omega\left(\tilde{\boldsymbol{f}}^{(i)}\right) := \prod_{i=1}^{d_y} \mathcal{N}\left(\tilde{\boldsymbol{\mu}}_{\tilde{f}}^{(i)}, \tilde{\boldsymbol{\Sigma}}_{\tilde{f}}^{(i)}\right), \tag{B.5}$$

can be expanded, where $\tilde{\boldsymbol{F}} \in \mathbb{R}^{m\times d_y}$ is the matrix of pseudo target output variables.

References

[1] D. Xiu, G. E. Karniadakis, Modeling uncertainty in steady state diffusion problems via generalized polynomial chaos, Comput. Methods Appl. Mech. Engrg. 191 (43) (2002) 4927–4948. doi:10.1016/S0045-7825(02)00421-8.

[2] C. Soize, R. Ghanem, Physical systems with random uncertainties: chaos representations with arbitrary probability measure, SIAM J. Sci. Comput. 26 (2) (2004) 395–410. doi:10.1137/S1064827503424505.

[3] R. Zhang, H. Dai, Stochastic analysis of structures under limited observations using kernel density estimation and arbitrary polynomial chaos expansion, Comput. Methods Appl. Mech. Engrg. 403 (2023) 115689. doi:10.1016/j.cma.2022.115689.

[4] K. D. Kantarakias, G. Papadakis, Sensitivity-enhanced generalized polynomial chaos for efficient uncertainty quantification, J. Comput. Phys. 491 (2023) 112377. doi:10.1016/j.jcp.2023.112377.

[5] M. A. Nielsen, Neural networks and deep learning, Determination Press, 2015.

[6] M. Papadrakakis, N. D. Lagaros, Reliability-based structural optimization using neural networks and Monte Carlo simulation, Comput. Methods Appl. Mech. Engrg. 191 (32) (2002) 3491–3507. doi:10.1016/S0045-7825(02)00287-6.

[7] C. M. Bishop, H. Bishop, Deep learning: foundations and concepts, Springer, 2024.

[8] E. Haghighat, M. Raissi, A. Moure, H. Gomez, R. Juanes, A physics-informed deep learning framework for inversion and surrogate modeling in solid mechanics, Comput. Methods Appl. Mech. Engrg. 379 (2021) 113741. doi:10.1016/j.cma.2021.113741.
[9] C. K. I. Williams, C. E. Rasmussen, Gaussian processes for machine learning, MIT Press, 2006.
[10] T. J. Santner, B. J. Williams, W. I. Notz, B. J. Williams, The design and analysis of computer experiments, Springer, 2003.
[11] J. D. Martin, T. W. Simpson, Use of kriging models to approximate deterministic computer models, AIAA J. 43 (4) (2005) 853–863. doi:10.2514/1.8650.
[12] A. Forrester, A. Sobester, A. Keane, Engineering design via surrogate modelling: a practical guide, John Wiley & Sons, 2008.
[13] K. J. Koh, F. Cirak, Stochastic PDE representation of random fields for large-scale Gaussian process regression and statistical finite element analysis, Comput. Methods Appl. Mech. Engrg. 417 (2023) 116358. doi:10.1016/j.cma.2023.116358.
[14] G. I. Schuëller, H. A. Jensen, Computational methods in optimization considering uncertainties–an overview, Comput. Methods Appl. Mech. Engrg. 198 (1) (2008) 2–13. doi:10.1016/j.cma.2008.05.004.
[15] I. Doltsinis, Z. Kang, Robust design of structures using optimization methods, Comput. Methods Appl. Mech. Engrg. 193 (23) (2004) 2221–2237. doi:10.1016/j.cma.2003.12.055.
[16] A. Asadpoure, M. Tootkaboni, J. K. Guest, Robust topology optimization of structures with uncertainties in stiffness–application to truss structures, Comput. Struct. 89 (2011) 1131–1141. doi:10.1016/j.compstruc.2010.11.004.
[17] G. Da Silva, E. Cardoso, Stress-based topology optimization of continuum structures under uncertainties, Comput. Methods Appl. Mech. Engrg. 313 (2017) 647–672. doi:10.1016/j.cma.2016.09.049.
[18] I. Ben-Yelun, A. O. Yuksel, F. Cirak, Robust topology optimisation of lattice structures with spatially correlated uncertainties, Struct. Multidiscip. Optim. 67 (2) (2024) 16. doi:10.1007/s00158-023-03716-4.
[19] J. T. Oden, R. Moser, O. Ghattas, Computer predictions with quantified uncertainty, Part I, SIAM News 43 (2010) 1–3.
[20] C. J. Roy, W. L. Oberkampf, A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing, Comput. Methods Appl. Mech. Engrg. 200 (25-28) (2011) 2131–2144. doi:10.1016/j.cma.2011.03.016.
[21] A. O'Hagan, Bayesian analysis of computer code outputs: A tutorial, Reliab. Eng. Syst. Saf. 91 (10-11) (2006) 1290–1300. doi:10.1016/j.ress.2005.11.025.
[22] A. Der Kiureghian, O. Ditlevsen, Aleatory or epistemic? Does it matter?, Struct. Saf. 31 (2) (2009) 105–112. doi:10.1016/j.strusafe.2008.06.020.
[23] S. Sakata, F. Ashida, M. Zako, Structural optimization using kriging approximation, Comput. Methods Appl. Mech. Engrg. 192 (7-8) (2003) 923–939. doi:10.1016/S0045-7825(02)00617-5.
[24] J. Zhang, A. A. Taflanidis, J. Medina, Sequential approximate optimization for design under uncertainty problems utilizing kriging metamodeling in augmented input space, Comput. Methods Appl. Mech. Engrg. 315 (2017) 369–395. doi:10.1016/j.cma.2016.10.042.
[25] M. Tootkaboni, A. Asadpoure, J. K. Guest, Topology optimization of continuum structures under uncertainty–a polynomial chaos approach, Comput. Methods Appl. Mech. Engrg. 201 (2012) 263–275. doi:10.1016/j.cma.2011.09.009.
[26] M. A. Bessa, S. Pellegrino, Design of ultra-thin shell structures in the stochastic post-buckling range using Bayesian machine learning and optimization, Int. J. Solids Struct. 139 (2018) 174–188. doi:10.1016/j.ijsolstr.2018.01.035.
[27] T. Zafar, Y. Zhang, Z. Wang, An efficient kriging based method for time-dependent reliability based robust design optimization via evolutionary algorithm, Comput. Methods Appl. Mech. Engrg. 372 (2020) 113386. doi:10.1016/j.cma.2020.113386.
[28] V. Keshavarzzadeh, R. M. Kirby, A. Narayan, Stress-based topology optimization under uncertainty via simulation-based Gaussian process, Comput. Methods Appl. Mech. Engrg. 365 (2020) 112992. doi:10.1016/j.cma.2020.112992.
[29] K. Tian, Z. Li, L. Huang, K. Du, L. Jiang, B. Wang, Enhanced variable-fidelity surrogate-based optimization framework by Gaussian process regression and fuzzy clustering, Comput. Methods Appl. Mech. Engrg. 366 (2020) 113045. doi:10.1016/j.cma.2020.113045.
[30] M. C. Kennedy, A. O'Hagan, Bayesian calibration of computer models, J. R. Stat. Soc. Series B Stat. Methodol. 63 (2001) 425–464. doi:10.1111/1467-9868.00294.
[31] K. P. Murphy, Machine learning: A probabilistic perspective, MIT Press, 2012.
[32] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, D. B. Rubin, Bayesian data analysis, 3rd Edition, CRC Press, 2014.
[33] C. Grigo, P.-S. Koutsourelakis, Bayesian model and dimension reduction for uncertainty propagation: applications in random media, SIAM-ASA J. Uncertain. Quantif. 7 (2019) 292–323. doi:10.1137/17M1155867.
[34] M. Girolami, E. Febrianto, G. Yin, F. Cirak, The statistical finite element method (statFEM) for coherent synthesis of observation data and model predictions, Comput. Methods Appl. Mech. Engrg. 375 (2021) 113533. doi:10.1016/j.cma.2020.113533.
[35] P. G. Constantine, E. Dow, Q. Wang, Active subspace methods in theory and practice: applications to kriging surfaces, SIAM

J. Sci. Comput. 36 (4) (2014) A1500–A1524. doi:10.1137/130916138.
[36] M. A. Bouhlel, N. Bartoli, A. Otsmane, J. Morlier, Improving kriging surrogates of high-dimensional design models by partial least squares dimension reduction, Struct. Multidiscip. Optim. 53 (2016) 935–952. doi:10.1007/s00158-015-1395-9.
[37] D. Gaudrie, R. Le Riche, V. Picheny, B. Enaux, V. Herbert, Modeling and optimization with Gaussian processes in reduced eigenbases, Struct. Multidiscip. Optim. 61 (2020) 2343–2361. doi:10.1007/s00158-019-02458-6.
[38] R. R. Lam, O. Zahm, Y. M. Marzouk, K. E. Willcox, Multifidelity dimension reduction via active subspaces, SIAM J. Sci. Comput. 42 (2) (2020) A929–A956. doi:10.1137/18M1214123.
[39] F. Romor, M. Tezzele, M. Mrosek, C. Othmer, G. Rozza, Multi-fidelity data fusion through parameter space reduction with applications to automotive engineering, Int. J. Numer. Methods. Eng. (2021). doi:10.1002/nme.7349.
[40] R. Tripathy, I. Bilionis, M. Gonzalez, Gaussian processes with built-in dimensionality reduction: applications to high-dimensional uncertainty propagation, J. Comput. Phys. 321 (2016) 191–223. doi:10.1016/j.jcp.2016.05.039.
[41] M. Guo, J. S. Hesthaven, Reduced order modeling for nonlinear structural analysis using Gaussian process regression, Comput. Methods Appl. Mech. Engrg. 341 (2018) 807–826. doi:10.1016/j.cma.2018.07.017.
[42] P. Tsilifis, P. Pandita, S. Ghosh, V. Andreoli, T. Vandeputte, L. Wang, Bayesian learning of orthogonal embeddings for multi-fidelity Gaussian processes, Comput. Methods Appl. Mech. Engrg. 386 (2021) 114147. doi:10.1016/j.cma.2021.114147.
[43] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, L. K. Saul, An introduction to variational methods for graphical models, Mach. Learn. 37 (1999) 183–233. doi:10.1023/A:1007665907178.
[44] D. M. Blei, A. Kucukelbir, J. D. McAuliffe, Variational inference: A review for statisticians, J. Am. Stat. Assoc. 112 (518) (2017) 859–877. doi:10.1080/01621459.2017.1285773.
[45] J. Povala, I. Kazlauskaite, E. Febrianto, F. Cirak, M. Girolami, Variational Bayesian approximation of inverse problems using sparse precision matrices, Comput. Methods Appl. Mech. Engrg. 393 (2022) 114712. doi:10.1016/j.cma.2022.114712.
[46] A. Vadeboncoeur, Ö. D. Akyildiz, I. Kazlauskaite, M. Girolami, F. Cirak, Fully probabilistic deep models for forward and inverse problems in parametric PDEs, J. Comput. Phys. 491 (2023) 112369. doi:10.1016/j.jcp.2023.112369.
[47] M. D. Hoffman, D. M. Blei, C. Wang, J. Paisley, Stochastic variational inference, J. Mach. Learn. Res. 14 (2013) 1303–1347.
[48] J. Li, L. Fuxin, S. Todorovic, Efficient Riemannian optimization on the Stiefel manifold via the Cayley transform, arXiv preprint arXiv:2002.01113 (2020). doi:10.48550/arXiv.2002.01113.
[49] J. Quinonero-Candela, C. E. Rasmussen, A unifying view of sparse approximate Gaussian process regression, J. Mach. Learn. Res. 6 (2005) 1939–1959.
[50] M. Titsias, Variational learning of inducing variables in sparse Gaussian processes, in: Artificial Intelligence and Statistics, PMLR, 2009, pp. 567–574.
[51] N. D. Lawrence, Learning for larger datasets with the Gaussian process latent variable model, in: Artificial Intelligence and Statistics, PMLR, 2007, pp. 243–250.
[52] A. C. Damianou, M. K. Titsias, N. D. Lawrence, Variational inference for latent variables and uncertain inputs in Gaussian processes, J. Mach. Learn. Res. 17 (2016) 1–62.
[53] F. Leibfried, V. Dutordoir, S. John, N. Durrande, A tutorial on sparse Gaussian processes and variational inference, arXiv preprint arXiv:2012.13962 (2020). doi:10.48550/arXiv.2012.13962.
[54] J. Hensman, N. Fusi, N. D. Lawrence, Gaussian processes for big data, arXiv preprint arXiv:1309.6835 (2013). doi:10.48550/arXiv.1309.6835.
[55] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014). doi:10.48550/arXiv.1412.6980.
[56] D. P. Kingma, M. Welling, Auto-encoding variational Bayes, arXiv preprint arXiv:1312.6114 (2013). doi:10.48550/arXiv.1312.6114.
[57] D. P. Kingma, M. Welling, et al., An introduction to variational autoencoders, Trends Mach. Learn. 12 (4) (2019) 307–392.
[58] D. Wipf, S. Nagarajan, A new view of automatic relevance determination, Adv. Neural. Inf. Process. Syst. 20 (2007).
[59] R. Horst, P. M. Pardalos, Handbook of global optimization, Vol. 2, Springer, 2013.
[60] G. H. Golub, C. F. Van Loan, Matrix computations, JHU Press, 2013.
[61] M. D. McKay, R. J. Beckman, W. J. Conover, A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, Technometrics 42 (1) (2000) 55–61. doi:10.1080/00401706.2000.10485979.
[62] D. Zhang, A coefficient of determination for generalized linear models, Am. Stat. 71 (4) (2017) 310–316. doi:10.1080/00031305.2016.1256839.
[63] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, A. Smola, A kernel two-sample test, J. Mach. Learn. Res. 13 (1) (2012) 723–773.
[64] T. Zhou, Y. Peng, Structural reliability analysis via dimension reduction, adaptive sampling, and Monte Carlo simulation, Struct. Multidiscip. Optim. 62 (5) (2020) 2629–2651. doi:10.1007/s00158-020-02633-0.
[65] J. N. Fuhg, A. Fau, U. Nackenhorst, State-of-the-art and comparative review of adaptive sampling methods for kriging, Arch.

Comput. Methods Eng. 28 (2021) 2689–2747. doi:10.1007/s11831-020-09474-6.
[66] M. Rixner, P.-S. Koutsourelakis, A probabilistic generative model for semi-supervised training of coarse-grained surrogates and enforcing physical constraints through virtual observables, J. Comput. Phys. 434 (2021) 110218. doi:10.1016/j.jcp.2021.110218.
[67] K. Linka, A. Schäfer, X. Meng, Z. Zou, G. E. Karniadakis, E. Kuhl, Bayesian physics informed neural networks for real-world nonlinear dynamical systems, Comput. Methods Appl. Mech. Engrg. 402 (2022) 115346. doi:10.1016/j.cma.2022.115346.
[68] T. P. Bohlin, Practical grey-box process identification: theory and applications, Springer, 2006.
[69] N. Jaquier, L. Rozo, High-dimensional Bayesian optimization via nested Riemannian manifolds, Adv. Neural. Inf. Process. Syst. 33 (2020) 20939–20951.
[70] Y. Zelickman, J. K. Guest, Construction aware optimization of concrete plate thicknesses, Eng. Struct. 296 (2023) 116889. doi:10.1016/j.engstruct.2023.116889.
[71] J. A. Lee, M. Verleysen, et al., Nonlinear dimensionality reduction, Vol. 1, Springer, 2007.