



A unified framework for closed-form nonparametric regression, classification, preference and mixed problems with Skew Gaussian Processes

Alessio Benavoli¹ · Dario Azzimonti² · Dario Piga²

Received: 27 January 2021 / Revised: 14 June 2021 / Accepted: 8 July 2021
© The Author(s) 2021

Abstract

Skew-Gaussian Processes (SkewGPs) extend the multivariate Unified Skew-Normal distributions over finite dimensional vectors to distribution over functions. SkewGPs are more general and flexible than Gaussian processes, as SkewGPs may also represent asymmetric distributions. In a recent contribution, we showed that SkewGP and probit likelihood are conjugate, which allows us to compute the exact posterior for non-parametric binary classification and preference learning. In this paper, we generalize previous results and we prove that SkewGP is conjugate with both the normal and affine probit likelihood, and more in general, with their product. This allows us to (i) handle classification, preference, numeric and ordinal regression, and mixed problems in a unified framework; (ii) derive closed-form expression for the corresponding posterior distributions. We show empirically that the proposed framework based on SkewGP provides better performance than Gaussian processes in active learning and Bayesian (constrained) optimization. These two tasks are fundamental for design of experiments and in Data Science.

Keywords Skew Gaussian process · Regression · Classification · Preference · Closed-form

1 Introduction

Gaussian Processes (GPs) are powerful nonparametric distributions over functions. For real-valued outputs, we can combine the GP prior with a Gaussian likelihood and perform exact posterior inference in closed form. However, in other cases, such as classification, preference learning, ordinal regression and mixed problems, the likelihood is no longer conjugate to the GP prior and a closed-form expression for the posterior is not available.

Editors: João Gama, Alípio Jorge, Salvador García.

Alessio Benavoli
alessio.benavoli@tcd.ie

¹ School of Computer Science and Statistics, Trinity College, Dublin, Ireland

² Dalle Molle Institute for Artificial Intelligence Research (IDSIA) - USI/SUPSI, Manno, Switzerland

In this paper, we show that it is actually possible to derive closed-form expression for the posterior process in all the above cases (not only for regression), and that it is a Skew Gaussian Process (SkewGP), a stochastic process whose finite dimensional marginals follow a multivariate Skew-Normal distribution.

Consider, for example, a classification task with a probit likelihood and a GP prior on the latent function. We can show that the posterior is a SkewGP and that the posterior latent function can be computed analytically on the training points. By exploiting the closure properties of SkewGPs, we can compute the distribution of the latent function at any new test point. We can easily obtain posterior samples at any test point by exploiting an additive representation for Skew-Normal vectors which decomposes a Skew-Normal vector into a linear combination of a normal vector and a truncated-normal vector. This decomposition has two main advantages. First of all, it just requires normal and truncated-normal samples. Truncated-normal samples can be obtained with rejection-free Monte Carlo sampling by using the linear elliptical slice sampling (*lin-ess*) method (Gessner et al. 2020); this avoids the need of expensive Markov Chain Monte Carlo methods. Secondly, the closure of the SkewGP family requires only sampling of the posterior at the training points. These samples can then be reused for any new test point, thus greatly reducing the computational cost for predicting many test points. Such scenario comes up, for instance, in Bayesian optimization tasks as we explain in Sect. 5.

As a prior class, SkewGPs are more general and more flexible nonparametric distributions than GPs, since SkewGPs may also represent asymmetric distributions. Moreover, SkewGPs include GPs as a particular case. By exploiting the closed-form expression for the posterior and predictive distribution, we compute inferences for regression, classification, preference and mixed problems with computation complexity of $O(n^3)$ and storage demands of $O(n^2)$, i.e. identical to GP regression.

This allows us to provide a unified framework for nonparametric inference for a large class of likelihoods and, consequently, supervised learning problems, as illustrated in Table 1.

1.1 Different types of observations and likelihood models

In supervised learning applications, we deal with the problem of learning input-output mappings from data. Consider a dataset consisting of n samples. Each of the samples is a pair of input vector $\mathbf{x}_i \in \mathbb{R}^d$ and output y_i . Depending on the type of the output variable, supervised learning problems can be divided into the following categories.

Continuous dependent: In the regression setting the outputs are real values $y_i \in \mathbb{R}$ and the input-output mapping is usually modelled as $y_i = f(\mathbf{x}_i) + v_i$, where v_i is an additive independent, identically distributed Gaussian noise with zero mean and variance σ_v^2 . The likelihood model is, for the i -th observation, given by:

$$p(y_i|f(\mathbf{x}_i)) = \phi\left(\frac{y_i - f(\mathbf{x}_i)}{\sigma_v}\right), \quad i = 1, \dots, n, \quad (1)$$

where $\phi(\cdot)$ is the PDF of the standard Normal distribution.

Binary: In the binary classification setting the outputs are categories that we can be labeled as $y_i = -1$ and $y_i = +1$, that is $y_i \in \{-1, 1\}$. The *probit* likelihood model is:¹

$$p(y_i|f(\mathbf{x}_i)) = \Phi(y_i f(\mathbf{x}_i)), \quad i = 1, \dots, n, \quad (2)$$

that is a Bernoulli distribution with probability $\Phi(f(\mathbf{x}_i))$, where $\Phi(\cdot)$ is the CDF of the standard Normal distribution.

Ordinal: In ordinal regression, $y_i \in \{1, 2, \dots, r\}$, where the integer $1, 2, \dots, r$ are used to denote order categories such as, for instance, movies' ratings. In ordinal regression, we map these ordered categories into a partition of \mathbb{R} , that is $j \rightarrow (b_{j-1}, b_j]$ with $b_0 = -\infty$ and $b_r = \infty$ so that $\mathbb{R} = (-\infty, b_1] \cup (b_1, b_2] \cup \dots \cup (b_{r-1}, \infty)$. The likelihood can be modelled as an indicator function $p(y_i|f(\mathbf{x}_i)) = I_{(b_{y_{i-1}}, b_{y_i}]}(f(\mathbf{x}_i))$. In case this observation model is contaminated by noise, we assume a Gaussian noise with zero mean and unknown variance σ_v^2 , the likelihood model becomes (Chu & Ghahramani, 2005a), for $i = 1, \dots, n$,

$$\begin{aligned} p(y_i|f(\mathbf{x}_i)) &= \int I_{(b_{y_{i-1}}, b_{y_i}]}(f(\mathbf{x}_i) + v) \phi\left(\frac{v}{\sigma_v}\right) dv \\ &= \Phi\left(\frac{b_{y_i} - f(\mathbf{x}_i)}{\sigma_v}\right) - \Phi\left(\frac{b_{y_{i-1}} - f(\mathbf{x}_i)}{\sigma_v}\right). \end{aligned} \quad (3)$$

The b_i defining the partition $(-\infty, b_1], (b_1, b_2], \dots, (b_{r-1}, \infty)$ are unknown, that is they are hyperparameters of the model. Note also that binary classification can be seen as a special case of ordinal regression with $r = 2$.

Preference: In preference learning, y_i is a preference relation over a set of predefined labels. For instance, assume we label each input \mathbf{x}_i with the label i , then we can define a preference relation over the inputs. The label preference $i > j$ means that the input \mathbf{x}_i is preferred to \mathbf{x}_j . This can be modelled by assuming that there is an unobservable latent function f associated with each input \mathbf{x}_k , and that the function values $\{f(\mathbf{x}_k)\}_k$ preserve the preference relations observed in the dataset. Then the likelihood can be modelled as an indicator function $p(\mathbf{x}_i > \mathbf{x}_j|f(\mathbf{x}_i), f(\mathbf{x}_j)) = I_{[f(\mathbf{x}_i), \infty)}(f(\mathbf{x}_i))$. This constrains the latent function values of the instances to be consistent with their preference relations. To allow some tolerance to noise in the inputs or the preference relations, one can assume the latent functions are contaminated with Gaussian noise (Chu & Ghahramani, 2005b):

$$\begin{aligned} p(\mathbf{x}_i > \mathbf{x}_j|f(\mathbf{x}_i), f(\mathbf{x}_j)) &= \int I_{[f(\mathbf{x}_i) + v_1, \infty)}(f(\mathbf{x}_i) + v_2) \phi\left(\frac{v_1}{\sigma_v}\right) \phi\left(\frac{v_2}{\sigma_v}\right) dv_1 dv_2 \\ &= \Phi\left(\frac{f(\mathbf{x}_i) - f(\mathbf{x}_j)}{\sqrt{2}\sigma_v}\right). \end{aligned} \quad (4)$$

More generally, the preferences of each input can be presented in the form of a preference graph (Chu & Ghahramani, 2005b), where the labels are the graph vertices. Some examples are shown in Chu and Ghahramani (2005b, Fig. 1). Therefore, in this more general case, $y_i = \{(c_i^+, c_i^-)\}_{j=1}^{g_i}$, where c_i^+ is the initial label vertex of the j -th edge and c_i^- is the terminal label, and g_i is the number of edges. This setting can be modelled by introducing a

¹ In Machine Learning literature, the probit function is known as sigmoid function, please see Bishop (2006, Sec. 4.3.5).

Table 1 SkewGPs are conjugate to a larger class of likelihoods

		regression	preference classification	mixed
Prior	Likelihood	Normal	Affine Probit	Product
	Posterior	Exact	SkewGP GP	SkewGP
	Approx.		GP	GP

latent function f_l for each predefined label (Chu & Ghahramani, 2005b). The observed edge (c_i^{j+}, c_i^{j-}) is modelled as the following constraint $f_{c_i^{j+}}(\mathbf{x}_i) \geq f_{c_i^{j-}}(\mathbf{x}_i)$ and the likelihood is

$$p(\{(c_i^{j+}, c_i^{j-})\}_{j=1}^{g_i} | \mathbf{f}(\mathbf{x}_i)) = \prod_{j=1}^{g_i} \Phi\left(\frac{f_{c_i^{j+}}(\mathbf{x}_i) - f_{c_i^{j-}}(\mathbf{x}_i)}{\sqrt{2}\sigma_v}\right), \quad (5)$$

where \mathbf{f} denotes the vector of latent function (one for each label). For instance, note that standard Multiclass Classification (Williams & Barber, 1998), Ordinal Regression, Hierarchical Multiclass classification, can be formulated in this way (Chu & Ghahramani, 2005b).

Mixed: In some applications, we may have scalar, binary and preference observations at the same time. Assuming independence, the likelihood model of n mixed observations can in general be written as the product of normal PDFs and normal CDFs. An example of mixed type data is shown in Fig. 1. The dotted line represents the function we used to generate the observations. The left (blue) points are numeric (non-noisy) observations and the right points represent preferences. We used the colored points (red and gold) to visualise the 30 preferential observations. The meaning of these points is as follows: (i) the value of the functions computed at the x 's corresponding to the bottom gold points is less than the value of the function computed at the x corresponding to the red point; (ii) the value of the function computed at the x 's corresponding to the top gold points is greater than the value of the function computed at the x corresponding to the red point. These 30 qualitative judgments is the only information we have on the function for $x \in [2.5, 5]$.

State-of-the-art: The state-of-the-art (nonparametric) Bayesian approach to deal with the above problems is to impose a Gaussian Process (GP) prior on the (latent) function(s) f . For scalar observations, due to the conjugacy between normal likelihood and normal prior, the posterior model is still a GP and its mean and covariance functions can be computed analytically (O'Hagan, 1978; Rasmussen & Williams, 2006, Ch.2).

For binary data (classification), the posterior process is not a GP. Several algorithms for approximate inference have been proposed, which are based on approximating the non-Gaussian posterior with a tractable Gaussian distribution. There are three main types of approximation: (i) Laplace Approximation (LP) (MacKay, 1996; Williams & Barber, 1998; ii) Expectation Propagation (EP) (Minka, 2001; iii) Kullback–Leibler divergence (KL) minimization (Opper & Archambeau, 2009), comprising Variational Bounding (VB) (Gibbs & MacKay, 2000) as a particular case. An exhaustive theoretical and empirical analysis of the above approaches was performed by Nickisch and Rasmussen (2008). They conclude that EP approximation is, in terms of accuracy, always the method of choice, except when you cannot afford the slightly longer running time compared to the fastest LP approximation.

Ordinal regression with GPs was proposed by Chu and Ghahramani (2005a) using the likelihood (3). The posterior is not a GP and so two approximations of the posterior were derived: LP and EP. The authors show that both LP and EP outperform the support vector approach (SVM), and that the EP approach is generally better than LP.

Preference learning based on GPs was proposed in Chu and Ghahramani (2005b) using the likelihoods (4) and (5). Again, the posterior is not a GP and the LP approximation is used to approximate the posterior with a GP; the approach outperforms SVMs.

In a recent paper (Benavoli et al., 2020), by extending a result derived by Durante (2019) for the parametric case, we showed that: (i) although the probit likelihood (2) and the GP are not conjugate, the posterior process can still be computed in closed form and is a Skew Gaussian Process (SkewGP); (ii) SkewGP prior and probit likelihood are conjugate. Such a novel result allowed us to compute the exact posterior for binary classification and for preference learning (Benavoli et al., 2021).

In a parallel line of research, Alodat and Al-Momani (2014) and Alodat and Shakhatreh (2020) considered a GP regression problem (with either a GP or a SkewGP prior) but with measurement errors following a skew normal distribution (resulting into a special case of “mixed likelihood” model discussed in this paper). The authors proved that the resulting posterior is SkewGP.

In the next sections, we unify these results by showing that SkewGP is conjugate with both the normal and affine probit likelihood and, more in general, with their product. This shows that SkewGP encompasses GP for both regression and classification.

The rest of the paper is organised as follows. Section 2 reviews the properties of the Skew Normal distribution and defines Skew Gaussian Processes. Section 3, which includes the main results of the paper, shows that SkewGPs provide closed-form solution to nonparametric regression, classification, preference and mixed problems. Section 4 provides algorithms to efficiently compute predictions for SkewGPs and to compute a fast approximation of the marginal likelihood. Section 5 discusses the application of SkewGPs to active learning and Bayesian optimisation. We show that SkewGPs outperform the Laplace and Expectation Propagation approximation. Finally, Sect. 6 concludes the paper.

2 Background on the Skew-Normal distribution and Skew Gaussian Processes

Skew-normal distributions have long been seen (O’Hagan & Leonard, 1976) as generalizations of the normal distribution allowing for non-zero skewness. Here we follow O’Hagan and Leonard (1976) and we say that a real-valued continuous random variable has skew-normal distribution if it has the following probability density function (PDF)

$$p(z) = \frac{2}{\sigma} \phi\left(\frac{z - \xi}{\sigma}\right) \Phi\left(\alpha\left(\frac{z - \xi}{\sigma}\right)\right), \quad z \in \mathbb{R},$$

where ϕ and Φ are the PDF and Cumulative Distribution Function (CDF), respectively, of the standard univariate Normal distribution. The numbers $\xi \in \mathbb{R}$, $\sigma > 0$, $\alpha \in \mathbb{R}$ are the location, scale and skewness parameters respectively.

The generalization of a univariate skew-normal to the multivariate case is not completely straightforward and over the years many generalisations of this distribution were proposed. Arellano and Azzalini (2006) provided a unification of those generalizations in a single and tractable multivariate *Unified Skew-Normal* distribution. This distribution

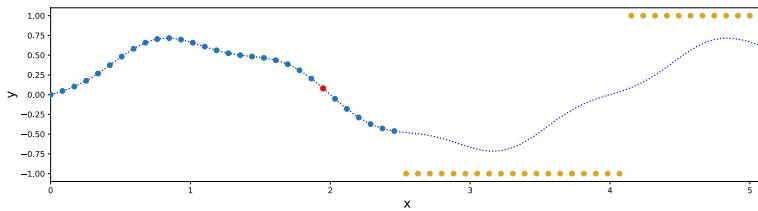


Fig. 1 Mixed numeric and preference observations

satisfies closure properties for marginals and conditionals and allows more flexibility due the introduction of additional parameters.

2.1 Unified Skew-Normal distribution

The Unified Skew-Normal is a very general family of multivariate distributions that allows for skewness on different directions through latent variables. We say that a p -dimensional vector $\mathbf{z} \in \mathbb{R}^p$ is distributed as a Unified Skew-Normal distribution with latent skewness dimension s , $\mathbf{z} \sim \text{SUN}_{p,s}(\xi, \Omega, \Delta, \gamma, \Gamma)$, if its probability density function (Azzalini, 2013, Ch.7):

$$p(\mathbf{z}) = \phi_p(\mathbf{z} - \xi; \Omega) \frac{\Phi_s(\gamma + \Delta^T \bar{\Omega}^{-1} D_{\Omega}^{-1} (\mathbf{z} - \xi); \Gamma - \Delta^T \bar{\Omega}^{-1} \Delta)}{\Phi_s(\gamma; \Gamma)}, \quad (6)$$

where $\phi_p(\mathbf{z} - \xi; \Omega)$ is the PDF of a multivariate Normal distribution with mean $\xi \in \mathbb{R}^p$ and covariance $\Omega = D_{\Omega} \bar{\Omega} D_{\Omega} \in \mathbb{R}^{p \times p}$, $\bar{\Omega}$ is a correlation matrix and D_{Ω} a diagonal matrix containing the square root of the diagonal elements in Ω . The notation $\Phi_s(\mathbf{a}; A)$ denotes the CDF of a s -dimensional multivariate Normal distribution with zero mean and covariance matrix A , $N_s(0, A)$, evaluated at $\mathbf{a} \in \mathbb{R}^s$. The distribution is parametrized by a location vector ξ , a covariance matrix Ω and the latent variable parameters $\gamma \in \mathbb{R}^s$, $\Gamma \in \mathbb{R}^{s \times s}$, $\Delta \in \mathbb{R}^{p \times s}$. In particular Δ is the skewness matrix.

The PDF (6) is well-defined provided that the matrix M below is positive definite, i.e.²

$$M := \begin{bmatrix} \Gamma & \Delta^T \\ \Delta & \bar{\Omega} \end{bmatrix} \in \mathbb{R}^{(s+p) \times (s+p)} > 0. \quad (7)$$

Note that when $\Delta = 0$, (6) reduces to $\phi_p(\mathbf{z} - \xi; \Omega)$, i.e. a skew-normal with zero skewness matrix is a normal distribution. Moreover we assume that $\Phi_0(\cdot) = 1$, so that, for $s = 0$, (6) becomes a multivariate Normal distribution.

Figure 2 shows the density of a univariate SUN distribution with latent dimensions $s = 1$ (a1) and $s = 2$ (a2). The effect of a higher latent dimension can be better observed in bivariate SUN densities as shown in Fig. 3. The contours of the corresponding bivariate

² In Azzalini (2013), it is assumed that M is a correlation matrix. This means that Γ must be a correlation matrix. However, from (6) note that, $\frac{\Phi_s(\gamma + \Delta^T \bar{\Omega}^{-1} D_{\Omega}^{-1} (\mathbf{z} - \xi); \Gamma - \Delta^T \bar{\Omega}^{-1} \Delta)}{\Phi_s(\gamma; \Gamma)}$ is equal to $\frac{\Phi_s(\mathbf{v}^{-1} \gamma + \mathbf{v}^{-1} \Delta^T \bar{\Omega}^{-1} D_{\Omega}^{-1} (\mathbf{z} - \xi); \mathbf{v}^{-1} (\Gamma - \Delta^T \bar{\Omega}^{-1} \Delta) \mathbf{v}^{-1})}{\Phi_s(\mathbf{v}^{-1} \gamma; \mathbf{v}^{-1} \Gamma \mathbf{v}^{-1})}$ where \mathbf{v} is the diagonal matrix that makes $\mathbf{v}^{-1} \Gamma \mathbf{v}^{-1}$ a correlation matrix. Therefore, in this paper, for simplicity we do not restrict Γ to be a correlation matrix. In any case, this does not have any impact on the properties of the SUN distribution we discuss in this section.

normal are dashed. We also plot the skewness directions given by $\bar{\Omega}^{-1}\Delta$. Note that a SUN with two latent dimensions (Fig. 3(a3), (a4)) has two direction of skewness.

Derivations of the moment generating function and of the first two multivariate central moments for the SUN distribution can be found in Gupta et al. (2013).

2.2 Additive representations

The role of the latent dimension s can be briefly explained as follows. Consider a random vector $\begin{bmatrix} \mathbf{x}_0 \\ \mathbf{x}_1 \end{bmatrix} \sim N_{s+p}(0, M)$ with M as in (7) and define \mathbf{y} as the vector with distribution $(\mathbf{x}_1 \mid \mathbf{x}_0 + \gamma > 0)$. The density of \mathbf{y} can be written as

$$\begin{aligned} f(\mathbf{y}) &= \frac{\int_{\mathbf{x}_0 + \gamma > 0} \phi_{s+p}((\mathbf{t}_0, \mathbf{y}); M) d\mathbf{t}_0}{\int_{\mathbf{x}_0 + \gamma > 0} \phi_s(\mathbf{t}; \Gamma) d\mathbf{t}} = \phi_p(\mathbf{y}; \bar{\Omega}) \frac{P(\mathbf{x}_0 + \gamma > 0 \mid \mathbf{x}_1 = \mathbf{y})}{\Phi_s(\gamma; \Gamma)} \\ &= \phi_p(\mathbf{y}; \bar{\Omega}) \frac{\Phi_s(\gamma + \Delta^T \bar{\Omega}^{-1} \mathbf{y}; \Gamma - \Delta^T \bar{\Omega}^{-1} \Delta)}{\Phi_s(\gamma; \Gamma)}, \end{aligned}$$

where the first equality comes from a basic property of conditional distributions, see, e.g. Azzalini (2013), Ch. 1.3, and the second equality is a consequence of the multivariate normal conditioning properties. Then we have that

$$\mathbf{z} = \xi + D_\Omega \mathbf{y} \sim \text{SUN}_{p,s}(\xi, \Omega, \Delta, \gamma, \Gamma).$$

Note that the skewness of \mathbf{z} is determined by the correlation Δ of \mathbf{x}_1 with the latent s -dimensional vector \mathbf{x}_0 .

The previous representation is useful for understanding the role of the latent dimension s in a skew-Gaussian vector. We report below another representation which is more practical for sampling. Consider the independent random vectors $\mathbf{r}_0 \sim N_p(0, \bar{\Omega} - \Delta \Gamma^{-1} \Delta^T)$ and $\mathbf{r}_{1,-\gamma}$, the truncation below $-\gamma$ of $\mathbf{r}_1 \sim N_s(0, \Gamma)$. Then the random variable

$$\mathbf{z}_u = \xi + D_\Omega(\mathbf{r}_0 + \Delta \Gamma^{-1} \mathbf{r}_{1,-\gamma}), \quad (8)$$

is distributed as (6), Azzalini (2013, Ch.7) and Benavoli et al. (2021, Sec.A.1). The additive representation introduced above can be used to draw samples from the distribution as it will be discussed in Sect. 4.

2.3 Closure properties

Among the many interesting properties of the Skew-Normal family (see Azzalini, 2013, Ch.7 for details), here we are particularly interested in its closure under marginalization and affine transformations. Consider $\mathbf{z} \sim \text{SUN}_{p,s}(\xi, \Omega, \Delta, \gamma, \Gamma)$ and partition $\mathbf{z} = [\mathbf{z}_1, \mathbf{z}_2]^T$, where $\mathbf{z}_1 \in \mathbb{R}^{p_1}$ and $\mathbf{z}_2 \in \mathbb{R}^{p_2}$ with $p_1 + p_2 = p$, then

$$\begin{aligned} \mathbf{z}_1 &\sim \text{SUN}_{p_1,s}(\xi_1, \Omega_{11}, \Delta_1, \gamma, \Gamma), \\ \text{with } \xi &= \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix}, \quad \Delta = \begin{bmatrix} \Delta_1 \\ \Delta_2 \end{bmatrix}, \quad \Omega = \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix}. \end{aligned} \quad (9)$$

Moreover, Azzalini (2013, Ch.7) the conditional distribution is a unified skew-Normal, i.e., $(\mathbf{Z}_2 \mid \mathbf{Z}_1 = \mathbf{z}_1) \sim \text{SUN}_{p_2,s}(\xi_{2|1}, \Omega_{2|1}, \Delta_{2|1}, \gamma_{2|1}, \Gamma_{2|1})$, where

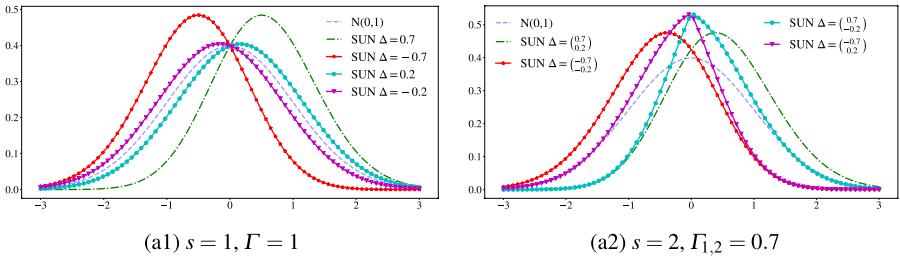


Fig. 2 Density plots for $\text{SUN}_{1,s}(0, 1, \Delta, \gamma, \Gamma)$. For all plots Γ is a correlation matrix, $\gamma = 0$, dashed lines are the contour plots of $y \sim N_1(0, 1)$

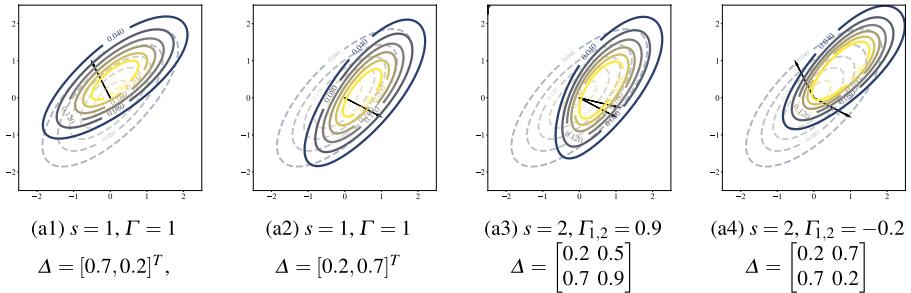


Fig. 3 Contour density plots for four unified skew-normal. For all plots $p = 2$, $\xi = [0, 0]^T$, Ω and Γ are correlation matrices with $\Omega_{1,2} = 0.7$, $\gamma = 0$, dashed lines are the contour plots of $y \sim N_2(\xi, \Omega)$

$$\begin{aligned}\xi_{2|1} &:= \xi_2 + \Omega_{21}\Omega_{11}^{-1}(\mathbf{z}_1 - \xi_1), \quad \Omega_{2|1} := \Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12}, \\ \Delta_{2|1} &:= \Delta_2 - \bar{\Omega}_{21}\bar{\Omega}_{11}^{-1}\Delta_1, \\ \gamma_{2|1} &:= \gamma + \Delta_1^T\Omega_{11}^{-1}(z_1 - \xi_1), \quad \Gamma_{2|1} := \Gamma - \Delta_1^T\bar{\Omega}_{11}^{-1}\Delta_1\end{aligned}$$

and $\bar{\Omega}_{11}^{-1} := (\bar{\Omega}_{11})^{-1}$.

In Sect. 3, we exploit this property to obtain samples from the predictive posterior distribution at a new input \mathbf{x}^* given samples of the posterior at the training inputs.

2.4 SkewGP

The unified skew-normal distribution can be generalized (Benavoli et al., 2020) to a stochastic process. We briefly recall here its construction.

Consider a location function $\xi : \mathbb{R}^d \rightarrow \mathbb{R}$, a scale (kernel) function³ $\Omega : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$, a skewness vector function $\Delta : \mathbb{R}^d \rightarrow \mathbb{R}^s$ and the parameters $\gamma \in \mathbb{R}^s, \Gamma \in \mathbb{R}^{s \times s}$. We say $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is a SkewGP with latent dimension s , if for any sequence of n points $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$, the vector $[f(\mathbf{x}_1), \dots, f(\mathbf{x}_n)] \in \mathbb{R}^n$ is skew-normal distributed with parameters γ, Γ and location, scale and skewness matrices, respectively, given by

³ With an abuse of notation, we denote the mean, covariance and skewness vector functions by ξ, Ω, Δ as for the parametric case in the previous sections.

$$\xi(X) := \begin{bmatrix} \xi(\mathbf{x}_1) \\ \xi(\mathbf{x}_2) \\ \vdots \\ \xi(\mathbf{x}_n) \end{bmatrix}, \quad \Omega(X, X) := \begin{bmatrix} \Omega(\mathbf{x}_1, \mathbf{x}_1) & \Omega(\mathbf{x}_1, \mathbf{x}_2) & \dots & \Omega(\mathbf{x}_1, \mathbf{x}_n) \\ \Omega(\mathbf{x}_2, \mathbf{x}_1) & \Omega(\mathbf{x}_2, \mathbf{x}_2) & \dots & \Omega(\mathbf{x}_2, \mathbf{x}_n) \\ \vdots & \vdots & \dots & \vdots \\ \Omega(\mathbf{x}_n, \mathbf{x}_1) & \Omega(\mathbf{x}_n, \mathbf{x}_2) & \dots & \Omega(\mathbf{x}_n, \mathbf{x}_n) \end{bmatrix}, \quad \Delta(X) := [\Delta(\mathbf{x}_1) \quad \Delta(\mathbf{x}_2) \quad \dots \quad \Delta(\mathbf{x}_n)]. \quad (10)$$

The skew-normal distribution is well defined if the matrix $M = \begin{bmatrix} \Gamma & \Delta(X) \\ \Delta(X)^T & \bar{\Omega}(X, X) \end{bmatrix}$ is positive definite for all $X = \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subset \mathbb{R}^d$ and for any n . Benavoli et al. (2020) shows that SkewGp is a well defined stochastic process. In that case we write $f \sim \text{SkewGP}_s(\xi, \Omega, \Delta, \gamma, \Gamma)$.

We briefly review here a possible choice for the functions Ω, Δ and the matrix Γ that guarantees that M is always positive definite. We follow Benavoli et al. (2020) and we choose a positive definite kernel stationary function⁴ $K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ which we will use to generate both Ω and Δ . Given n points and $X = \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subset \mathbb{R}^d$, s pseudo-points $U = \{\mathbf{u}_1, \dots, \mathbf{u}_s\} \subset \mathbb{R}^d$ and a phase diagonal matrix $L \in \mathbb{R}^{s \times s}$ with elements $L_{ii} \in \{-1, 1\}$, we build the matrix M as

$$M = \begin{bmatrix} \Gamma & \Delta(X, U) \\ \Delta(U, X) & \Omega(X, X) \end{bmatrix} = \begin{bmatrix} L\bar{K}(U, U)L & L\bar{K}(U, X) \\ \bar{K}(X, U)L & \bar{K}(X, X) \end{bmatrix}, \quad (11)$$

where $\bar{K}(\mathbf{x}, \mathbf{x}') = \frac{1}{\sigma^2} K(\mathbf{x}, \mathbf{x}')$. This structure guarantees that M is a positive definite matrix for any X . Pseudo-points allow for a flexible handling of the skewness determined by Γ and Δ . See Benavoli et al. (2020) for several examples of the effect of the pseudo-points positions.

SkewGPs can then be used as prior for f in a Bayesian model. Note that, with $s = 0$, this construction recovers a GP with covariance kernel $\Omega = K$. Below we show that this larger family of prior distributions for f present remarkable conjugacy properties with many common likelihoods.

3 Conjugacy of SkewGP

This section includes the main results of the paper: we will prove that SkewGP is conjugate with both the normal and probit affine likelihood and, more in general, with their product.

3.1 Normal likelihood

Consider n input points $X = \{\mathbf{x}_i : i = 1, \dots, n\}$, with $\mathbf{x}_i \in \mathbb{R}^d$, and m_r output points $Y = \{y_i : i = 1, \dots, m_r\}$, with $y_i \in \mathbb{R}$ and the likelihood

$$p(Y | f(X)) = \prod_{i=1}^{m_r} \phi\left(\frac{y_i - \mathbf{c}_i^T f(X)}{\sigma_v}\right) = \phi_{m_r}(Y - Cf(X); R), \quad (12)$$

⁴ This construction can easily be generalised to non-stationary kernels.

with $\phi_{m_r}(Y - Cf(X); R)$ denotes a multivariate normal PDF with zero mean and covariance R computed at $Y - Cf(X)$, where $\mathbb{R}^{m_r \times n} \ni C = [\mathbf{c}_i^T]_{i=1}^{m_r}$ and $\mathbf{c}_i \in \mathbb{R}^n$ is a data dependent vector and $R = \sigma_v^2 I_{m_r}$ is a covariance matrix, with I_{m_r} being the identity matrix of dimension m_r .⁵

Lemma 1 *Let us assume that $f(\mathbf{x}) \sim \text{SkewGP}_s(\xi(\mathbf{x}), \Omega(\mathbf{x}, \mathbf{x}'), \Delta(\mathbf{x}), \gamma, \Gamma)$ and consider the Normal likelihood $p(Y | f(X)) = \phi_{m_r}(Y - Cf(X); R)$ where $C \in \mathbb{R}^{m_r \times n}$ and $R = \mathbb{R}^{m_r \times m_r}$. The posterior distribution of $f(X)$ is a SUN:*

$$\begin{aligned} p(f(X) | Y) &= \text{SUN}_{m_r, s}(\xi_p, \Omega_p, \Delta_p, \gamma_p, \Gamma_p) \quad \text{with} \\ \xi_p &= \xi + \Omega C^T (C \Omega C^T + R)^{-1} (Y - C \xi), \end{aligned} \quad (13)$$

$$\Omega_p = \Omega - \Omega C^T (C \Omega C^T + R)^{-1} C \Omega, \quad (14)$$

$$\Delta_p = \bar{\Omega}_p D_{\Omega_p} D_{\Omega}^{-1} \bar{\Omega}^{-1} \Delta, \quad (15)$$

$$\gamma_p = \gamma + \Delta^T \bar{\Omega}^{-1} D_{\Omega}^{-1} (\xi_p - \xi), \quad (16)$$

$$\Gamma_p = \Gamma - \Delta^T \bar{\Omega}^{-1} \Delta + \Delta_p^T \bar{\Omega}_p^{-1} \Delta_p, \quad (17)$$

where, for simplicity of notation, we denoted $\xi(X), \Omega(X, X), \Delta(X)$ as ξ, Ω, Δ , and $\Omega = D_{\Omega} \bar{\Omega} D_{\Omega}$ and $\Omega_p = D_{\Omega_p} \bar{\Omega}_p D_{\Omega_p}$.

All the proofs are in the Appendix. For either $s = 0$ or $\Delta = 0$, the SkewGP prior becomes a GP and it can be noted that, in this case, the posterior is Gaussian with posterior mean ξ_p and posterior covariance Ω_p (that is the terms in (14)–(16) disappear).

In practical applications of SkewGP, the hyperparameters of the scale function $\Omega(\mathbf{x}, \mathbf{x}')$, of the skewness vector function $\Delta(\mathbf{x}) \in \mathbb{R}^s$ and the hyperparameters $\gamma \in \mathbb{R}^s, \Gamma \in \mathbb{R}^{s \times s}$ must be selected. As for GPs, we use Bayesian model selection to set such hyperparameters and this requires the maximization of the marginal likelihood with respect to these hyperparameters.

Corollary 1 *Consider the probabilistic model in Lemma 1, the marginal likelihood of the observations Y is*

$$p(Y) = \phi_{m_r}(Y - C\xi(X); C\Omega(X, X)C^T + R) \frac{\Phi_s(\gamma_p; \Gamma_p)}{\Phi_s(\gamma; \Gamma)}. \quad (18)$$

Observe again that for either $s = 0$ or $\Delta = 0$, the marginal likelihood coincides with that of a GP because the ratio $\Phi_s(\gamma_p; \Gamma_p)/\Phi_s(\gamma; \Gamma)$ disappears. The computation of the marginal

⁵ Note that, in regression, one usually has that $m_r = n$. However, in this paper, we are interested in mixed-type data problems and so we will consider cases where one has numeric observations for a subset of dimension m_r (the subscript r stands for regression) of the x 's (the subset is selected via the matrix C), and binary/preference observations for another subset of dimension m_a (the subscript a stands for affine) of the x 's and so on. This will become clear in Sect. 3.3 where we consider mixed-type data. In this section, we will consider the case $m_r \neq n$ to keep this generality.

likelihood (18) requires the calculation of two multivariate CDFs. We will address this point in Sect. 4.

We now prove that, a-posteriori, for a new test point \mathbf{x} , the function $f(\mathbf{x})$ is SkewGP distributed under the Normal likelihood in (12).

Theorem 1 *Let us assume a SkewGP prior $f(\mathbf{x}) \sim \text{SkewGP}_s(\xi(\mathbf{x}), \Omega(\mathbf{x}, \mathbf{x}'), \Delta(\mathbf{x}), \gamma, \Gamma)$, the likelihood $p(Y | f(X)) = \phi_{m_r}(Y - Cf(X); R)$, then a-posteriori f is SkewGP with mean, scale, and skew functions:*

$$\tilde{\xi}(\mathbf{x}) = \xi(\mathbf{x}) + \Omega(\mathbf{x}, X)C^T(C\Omega(X, X)C^T + R)^{-1}(Y - C\xi(X)), \quad (19)$$

$$\tilde{\Omega}(\mathbf{x}, \mathbf{x}) = \Omega(\mathbf{x}, \mathbf{x}) - \Omega(\mathbf{x}, X)C^T(C\Omega(X, X)C^T + R)^{-1}C\Omega(X, \mathbf{x}), \quad (20)$$

$$\tilde{\Delta}(\mathbf{x}) = D_{\tilde{\Omega}(\mathbf{x}, \mathbf{x})}^{-1}D_{\Omega(\mathbf{x}, \mathbf{x})}\Delta(\mathbf{x}) - D_{\tilde{\Omega}(\mathbf{x}, \mathbf{x})}^{-1}\Omega(\mathbf{x}, X)C^T(C\Omega(X, X)C^T + R)^{-1}CD_{\Omega(X, X)}\Delta(X), \quad (21)$$

and parameters γ_p, Γ_p as in Lemma 1.

The posterior mean (19) and posterior covariance (20) coincide with those of the posterior GP. The term (21) is the posterior skewness. This results proves that SkewGP process and the Normal likelihood are conjugate.

3.2 Probit affine likelihood

Consider n input points $X = \{\mathbf{x}_i : i = 1, \dots, n\}$, with $\mathbf{x}_i \in \mathbb{R}^d$, and a data-dependent matrix $W \in \mathbb{R}^{m_a \times n}$ and vector $Z \in \mathbb{R}^{m_a}$. We define an affine probit likelihood as

$$p(W, Z | f(X)) = \Phi_{m_a}(Z + Wf(X); \Sigma), \quad (22)$$

where $\Phi_{m_a}(\mathbf{x}; \Sigma)$ is the m_a -variate Gaussian CDF evaluated at $\mathbf{x} \in \mathbb{R}^m$ with covariance $\Sigma \in \mathbb{R}^{m_a \times m_a}$. Note that this likelihood model includes the classic GP probit classification model (Rasmussen & Williams, 2006) with binary observations $y_1, \dots, y_n \in \{0, 1\}$ encoded in the matrix $W = \text{diag}(2y_1 - 1, \dots, 2y_n - 1)$, where $m = n$, $Z = 0$ and $\Sigma = I_{m_a}$ (the identity matrix of dimension m_a), that is

$$\Phi_{m_a} \left(\begin{bmatrix} 2y_1 - 1 \\ \vdots \\ 2y_n - 1 \end{bmatrix} | f(X); I_{m_a} \right). \quad (23)$$

Moreover, the likelihood in (22) is equal to the preference likelihood (4) for $Z = 0$ and a particular choice of W . In fact, consider the likelihood

$$\Phi_{m_a} \left(\begin{bmatrix} \frac{f(\mathbf{v}_1) - f(\mathbf{u}_1)}{\sigma_v} \\ \vdots \\ \frac{f(\mathbf{v}_m) - f(\mathbf{u}_m)}{\sigma_v} \end{bmatrix} | I_{m_a} \right), \quad (24)$$

where $\mathbf{v}_i, \mathbf{u}_j \in X$. If we denote by $W \in \mathbb{R}^{m_a \times n}$ the matrix defined as $W_{i,j} = V_{i,j} - U_{i,j}$ where $V_{i,j} = \frac{1}{\sigma_v}$ if $\mathbf{v}_i = \mathbf{x}_j$ and 0 otherwise and $U_{i,j} = \frac{1}{\sigma_v}$ if $\mathbf{u}_i = \mathbf{x}_j$ and 0 otherwise. Then we can

write the likelihood (24) as in (22).⁶ The conjugacy between SkewGP and the likelihood (23) was proved in Benavoli et al. (2020) extending to the nonparametric setting a result proved in Durante (2019, Th.1 and Co.4) for the parametric setting. The conjugacy between SkewGP and the likelihood (24) was derived in Benavoli et al. (2021) extending the results for classification. In this paper, we further expand those results by including the shifting term Z (which, for instance, allows us to deal with the likelihood (3)) and a general covariance matrix Σ , as in (22).

Lemma 2 *Let us assume that $f(\mathbf{x}) \sim \text{SkewGP}_s(\xi(\mathbf{x}), \Omega(\mathbf{x}, \mathbf{x}'), \Delta(\mathbf{x}), \gamma, \Gamma)$ and consider the likelihood $\Phi_{m_a}(Z + Wf(X); \Sigma)$ where $\Sigma \in \mathbb{R}^{m_a \times m_a}$ is symmetric positive definite, $W \in \mathbb{R}^{m_a \times n}$ and $Z \in \mathbb{R}^{m_a}$ are an arbitrary data dependent matrix and, respectively, vector. The posterior of $f(X)$ is a SUN:*

$$p(f(X)|W, Z) = \text{SUN}_{n, s+m_a}(\xi_p, \Omega_p, \Delta_p, \gamma_p, \Gamma_p) \quad (25)$$

$$\xi_p = \xi, \quad \Omega_p = \Omega, \quad (26)$$

$$\Delta_p = [\Delta, \bar{\Omega}D_\Omega W^T], \quad (27)$$

$$\gamma_p = [\gamma, Z + W\xi]^T, \quad (28)$$

$$\Gamma_p = \begin{bmatrix} \Gamma & \Delta^T D_\Omega W^T \\ WD_\Omega \Delta & (W\Omega W^T + \Sigma) \end{bmatrix}, \quad (29)$$

where, for simplicity of notation, we denoted $\xi(X), \Omega(X, X), \Delta(X)$ as ξ, Ω, Δ respectively and $\Omega = D_\Omega \bar{\Omega} D_\Omega$.

In this case, the posterior is a skew normal distribution even for either $s = 0$ or $\Delta(\mathbf{x}) = 0$. This shows that, for GP classification and GP preference learning, the true posterior is a skew normal distribution.⁷

Corollary 2 *Consider the probabilistic model in Lemma 2, the marginal likelihood of the observations Z, W is*

$$p(Z, W) = \frac{\Phi_{s+m_a}(\gamma_p; \Gamma_p)}{\Phi_s(\gamma; \Gamma)}. \quad (30)$$

Note that, both the posterior in (13) and the marginal likelihood in (30) require the accurate computation of Φ_{s+m_a} – a multi-dimensional integral whose dimension increases with the number of observations m_a . We will discuss how to tackle this issue in Sect. 4.

⁶ We omitted the constant $\sqrt{2}$ for simplicity.

⁷ Since here we are basically considering a parametric setting (the posterior of f computed at the training points), this lemma extends the results derived in Durante (2019, Th.1 and Co.4), for the parametric case to general probit affine likelihoods $\Phi_{m_a}(Z + Wf; \Sigma)$. In the proofs of Theorem 1 and Corollary 4 in Durante (2019), Γ_p (posterior Γ) is rescaled to make it a correlation matrix. As explained in Sect. 2.1, we do not do it for simplicity of notation.

We now prove that, a-posteriori, for a new test point \mathbf{x} , the function $f(\mathbf{x})$ is SkewGP distributed under the affine probit likelihood in (22).

Theorem 2 *Let us assume that $f(\mathbf{x}) \sim \text{SkewGP}_s(\xi(\mathbf{x}), \Omega(\mathbf{x}, \mathbf{x}'), \Delta(\mathbf{x}), \gamma, \Gamma)$ and consider the likelihood $\Phi_{m_a}(Z + Wf(X); \Sigma)$ where $\Sigma \in \mathbb{R}^{m_a \times m_a}$ is symmetric positive definite, $W \in \mathbb{R}^{m_a \times n}$ and $Z \in \mathbb{R}^{m_a}$, then a-posteriori f is SkewGP with mean, covariance and skewness functions:*

$$\begin{aligned}\tilde{\xi}(\mathbf{x}) &= \xi(\mathbf{x}) \\ \tilde{\Omega}(\mathbf{x}, \mathbf{x}) &= \Omega(\mathbf{x}, \mathbf{x}), \\ \tilde{\Delta}(\mathbf{x}) &= [\Delta(\mathbf{x}) \quad D_{\Omega(\mathbf{x}, \mathbf{x})}^{-1} \Omega(\mathbf{x}, X) W^T],\end{aligned}\tag{31}$$

and parameters γ_p, Γ_p as in Lemma 2.

This results proves that SkewGP process and the probit affine likelihood are conjugate, which provides closed-form expression for the posterior for both classification and preference learning.

Example 1D preference learning. In this paragraph, we provide a one-dimensional example of how we can use the above derivations to compute the exact posterior for preference learning. Consider the non-linear function $g(x) = \cos(5x) + e^{-\frac{x^2}{2}}$ which is in Fig. 4.

We use this function to generate 45 random pairwise preferences between 25 random points $x_i \in [-2.6, 2.6]$. Our aim is to infer $f(x)$, that is the latent function that models the preference relations observed in the dataset. We will then use the learned model to compute $f(x) - f(0.05)$, which can tell us the points x which are preferred to $x_r = 0.05$ (0.05 is the point corresponding to the maximum value of g in the dataset).

In all cases we consider a GP prior with zero mean and radial basis function (RBF) kernel over the unknown function f . We will compare the exact posterior computed via SkewGP with two approximations: Laplace (LP) and Expectation Propagation (EP). We discuss how to compute the predictions for SkewGP in Sect. 4.

Figure 5(top) shows the predicted posterior distribution $f(x) - f(0.05)$ (and relative 95% credible region) computed according to LP, EP and SkewGP. All the methods use the same prior: a GP with zero mean and RBF covariance function (the hyperparameters are the same: lengthscale 0.33 and variance $\sigma^2 = 50$). Therefore, the only difference between the exact posterior (SkewGP) and the posteriors of EP and LP is due to the different approximations. The true posterior (SkewGP) of the preference function is skewed as it can be seen in Fig. 5(bottom-left), which reports the *skewness statistics* for $h(x) = f(x) - f(x_r)$ as a function of x , defined as:

$$SS(h(x)) := \frac{\mathbb{E}[(h(x) - \mu)^3]}{(\mathbb{E}[(h(x) - \mu)^2])^{3/2}},$$

with $\mu := \mathbb{E}[h(x)]$, and \mathbb{E} is computed via Monte Carlo sampling from the posterior.

The strong skewness of the posterior is the reason why both the LP and EP approximations are not able to correctly approximate the posterior as shown in Fig. 5(top). Figure 5(bottom-right) shows the predictive posterior distribution for $f(0.77)$ for the three models. It can be noticed that the true posterior (SkewGP) is skewed, which explains why the LP and EP approximations are not accurate.

3.3 Mixed likelihood

Consider n input points $X = \{\mathbf{x}_i : i = 1, \dots, n\}$, with $\mathbf{x}_i \in \mathbb{R}^d$, and the product likelihood:

$$p(Y, Z, W | f(X)) = \Phi_{m_a}(Z + Wf(X); \Sigma) \phi_{m_r}(Y - Cf(X); R). \quad (32)$$

where C, R, Y, W, Z, Σ have the same dimensions as before. To obtain the posterior we apply first Theorem 2 and then Theorem 1.

Theorem 3 *Let us assume a SkewGP prior $f(\mathbf{x}) \sim \text{SkewGP}_s(\xi(\mathbf{x}), \Omega(\mathbf{x}, \mathbf{x}'), \Delta(\mathbf{x}), \gamma, \Gamma)$, the likelihood (32), then a-posteriori f is SkewGP with mean, covariance and skewness functions:*

$$\tilde{\xi}(\mathbf{x}) = \xi(\mathbf{x}) + \Omega(\mathbf{x}, X)C^T(C\Omega(X, X)C^T + R)^{-1}(Y - C\xi(X)), \quad (33)$$

$$\tilde{\Omega}(\mathbf{x}, \mathbf{x}) = \Omega(\mathbf{x}, \mathbf{x}) - \Omega(\mathbf{x}, X)C^T(C\Omega(X, X)C^T + R)^{-1}C\Omega(X, \mathbf{x}), \quad (34)$$

$$\begin{aligned} \tilde{\Delta}(\mathbf{x}) &= D_{\tilde{\Omega}(\mathbf{x}, \mathbf{x})}^{-1} \left[\Delta(\mathbf{x}) - D_{\Omega(\mathbf{x}, \mathbf{x})}^{-1} \Omega(\mathbf{x}, X)W^T \right] \\ &\quad - D_{\tilde{\Omega}(\mathbf{x}, \mathbf{x})}^{-1} \Omega(\mathbf{x}, X)C^T(C\Omega(X, X)C^T + R)^{-1} \\ &\quad CD_{\Omega(X, X)} \left[\Delta(X) - D_{\Omega(X, X)}^{-1} \Omega(X, X)W^T \right], \end{aligned} \quad (35)$$

$$\tilde{\gamma} = \gamma_p + \left[\Delta(X) - D_{\Omega(X, X)}^{-1} \Omega(X, X)W^T \right]^T \tilde{\Omega}(X, X)^{-1} D_{\Omega(X, X)}^{-1} (\tilde{\xi}(X) - \xi(X)) \quad (36)$$

$$\begin{aligned} \tilde{\Gamma} &= \Gamma_p - \left[\Delta(X) - D_{\Omega(X, X)}^{-1} \Omega(X, X)W^T \right]^T \tilde{\Omega}^{-1}(X, X) \left[\Delta(X) - D_{\Omega(X, X)}^{-1} \Omega(X, X)W^T \right] \\ &\quad + \Delta_p^T \tilde{\Omega}(X, X)^{-1} \Delta_p, \end{aligned} \quad (37)$$

with

$$\Delta_p = \tilde{\Omega}(X, X)D_{\tilde{\Omega}(X, X)}D_{\Omega(X, X)}^{-1}\tilde{\Omega}^{-1}(X, X) \left[\Delta(X) - D_{\Omega(X, X)}^{-1} \Omega(X, X)W^T \right], \quad (38)$$

and γ_p, Γ_p as in Lemma 2.

Corollary 3 *The marginal likelihood of the observations Y is*

$$p(Y) = \phi_{m_r}(Y - C\xi(X); C\Omega(X, X)C^T + R) \frac{\Phi_{s+m_a}(\tilde{\gamma}; \tilde{\Gamma})}{\Phi_s(\gamma; \Gamma)}, \quad (39)$$

with $\tilde{\gamma}, \tilde{\Gamma}$ in Theorem 3.

This general setting shows that SkewGPs are conjugate with a larger class of likelihoods and, therefore, encompass GPs. Note that, also in this case, the posterior and the marginal likelihood in (39) require the accurate computation of Φ_{s+m_a} – a multi-dimensional integral whose dimension increases with the number of observations m_a . We will discuss how to tackle this issue in Sect. 4.

In the next paragraphs, we provide two one-dimensional examples of (i) mixed numeric and preference regression; (ii) mixed numeric and binary output regression. Another

Fig. 4 1D function $g(x) = \cos(5x) + e^{-\frac{x^2}{2}}$

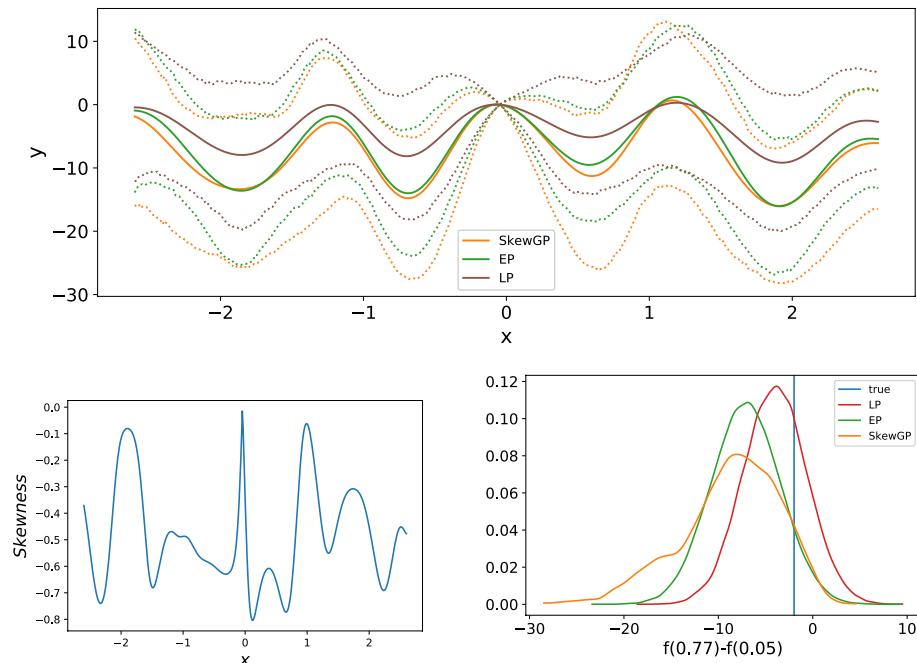
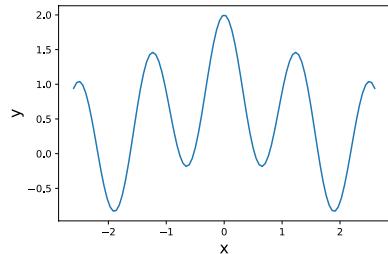


Fig. 5 Preference learning. The top plot shows the posterior means and credible intervals for LP, EP and SkewGP. The bottom-left plot shows the skewness statistics for the SkewGP posterior as a function of x . The right plot the predictive distribution of $f(0.77) - f(0.05)$ for the three models

application where the “mixed likelihood” model arises is GP regression with skew-normal measurement errors, see Alodat and Shakhatreh (2020).

Example 1D mixed numeric and preference regression Consider again the regression problem discussed in Sect. 1 and the dataset shown in Fig. 1.

Figure 6(top) shows the posterior mean and relative credible region of the regression function f computed according to SkewGP, LP and EP. All the methods use the same prior: a GP with zero mean and RBF covariance function (the hyperparameters are the same: lengthscale 0.19, variance $\sigma^2 = 1$ and noise variance $\sigma^2 = 0.001$).

Figure 6(bottom-left) reports the skewness statistics for SkewGP and Fig. 6(bottom-right) the predictive posterior distribution for $f(4.7)$ for the three models. It can be noticed that the true posterior (SkewGP) is very skewed for $x \geq 2.5$, which explains why the LP and EP approximations are not accurate. The LP approximation heavily underestimates the

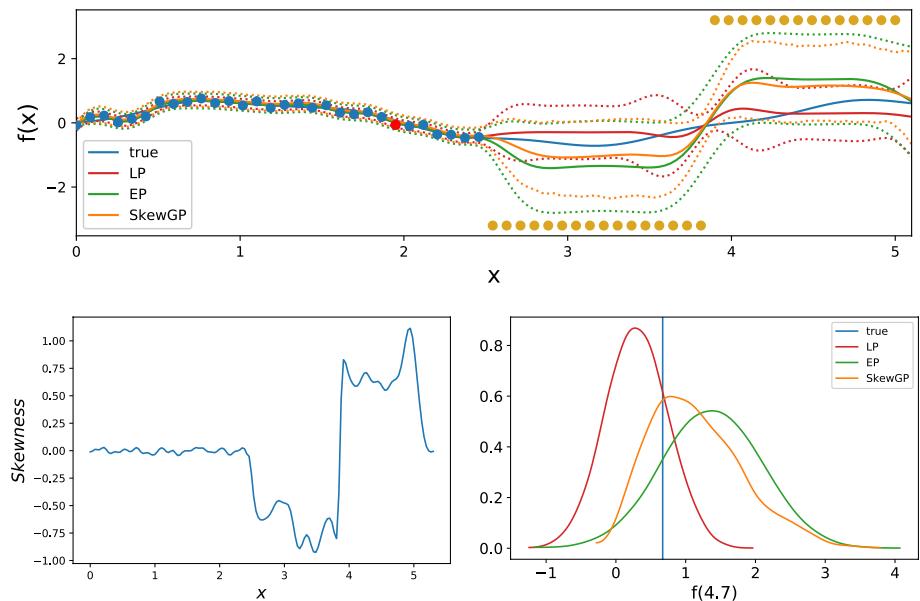


Fig. 6 Mixed numeric and preference learning. Top: dataset, true function (blue), mean and credible intervals for LP, EP, SkewGP. Bottom-left: skewness statistics for the SkewGP posterior as a function of x . Bottom-right: predictive distribution of $f(4.7)$ for the three models (color figure online)

mean and the “support” of the true posterior (SkewGP), as also evident from Fig. 6(bottom-right). The EP approximation estimates a large variance to “fit” the skewed posterior. This confirms, also for the mixed case, what was observed by Kuss and Rasmussen (2005) and Nickisch and Rasmussen (2008) for classification.⁸ The symmetry assumption for the posterior for LP and EP affects the coverage of their credible intervals (regions).

It can be noticed that the three models coincide for $x < 2.5$, that is in the region including the numeric observations. The SkewGP posterior is indeed symmetric in this region, as it can be seen from Fig. 6(bottom-left).

Example 1D mixed numeric and binary regression This mixed problem arises when binary judgments (valid or non-valid) together with scalar observations $f(x)$ are available. Such situation often comes up in industrial applications. For example imagine a process f that produces a certain noisy output $f(x) + v$ which depends on input parameters x . Given a certain input x_k , assume now that when $f(x_k) + v \leq h$ no output is produced. In this case, observations are pairs and the space of possibility is $\{(valid, y), (non\text{-}valid, None)\}$, with $y = f(x) + v$. The above setting can be modelled by the likelihood:

⁸ In Kuss and Rasmussen (2005) and Nickisch and Rasmussen (2008), it was observed that, in classification, the LP mean is always closer to zero than the true mean. This is what we mean here with “underestimating” the mean. Moreover, it was observed that the credible intervals of the LP approximation are usually too small, while EP tends to overestimate their width.

$$\begin{cases} \phi\left(\frac{y-f(x)}{\sigma_v}\right) \Phi\left(\frac{f(x)-h}{\sigma_v}\right), & (\text{valid}, y) \\ \Phi\left(\frac{h-f(x)}{\sigma_v}\right), & (\text{non-valid}, \text{None}), \end{cases} \quad (40)$$

and, therefore, formulated as a mixed regression problem.⁹ Figure 7(top) shows the true function we used to generate the data (in blue), the numeric data (blue points) and binary data (gray points where $y_1 = 1$ means valid and $y_i = -1$ means non-valid). The threshold h has been set to $h = 0$ and so the non-valid zone is the region in gray. We report the posterior means and credible regions of LP, EP and SkewGP. All the methods use the same prior: a GP with zero mean and Radial Basis Function (RBF) covariance function (the hyperparameters were estimated using EP: lengthscale 0.497, variance $\sigma^2 = 0.147$ and noise variance $\sigma^2 = 0.0021$). Figure 7(bottom-left) reports the skewness statistics for SkewGP and Fig. 7(bottom-right) the predictive posterior distribution for $f(3.3)$ for the three models. It can be noticed that the true posterior (SkewGP) is very skewed in the non-valid zone ($2 \leq x \leq 4$), which explains why the LP and EP approximations are not accurate. Again the LP approximation heavily underestimates the mean and the support of the true posterior (SkewGP) and the EP approximation estimates a large variance to “fit” the skewed posterior.

4 Sampling from the posterior and hyperparameters selection

The computation of predictive inference (mean, credible intervals etc.) is achieved by sampling the posterior SkewGP. This can be obtained using the additive representation discussed in Sect. 2.2. Recall that $\mathbf{z} \sim \text{SUM}_{p,s}(\xi, \Omega, \Delta, \gamma, \Gamma)$ can be written as $\mathbf{z} = \xi + D_\Omega(\mathbf{r}_0 + \Delta\Gamma^{-1}\mathbf{r}_{1,-\gamma})$ with $\mathbf{r}_0 \sim \phi_p(0; \bar{\Omega} - \Delta\Gamma^{-1}\Delta^T)$ and $\mathbf{r}_{1,-\gamma}$ is the truncation below γ of $\mathbf{r}_1 \sim \phi_s(0; \Gamma)$. Note that sampling \mathbf{r}_0 can be achieved efficiently with standard methods, however using standard rejection sampling for the variable $\mathbf{r}_{1,-\gamma}$ would incur in exponentially growing sampling time as the dimension m_a increases. In Benavoli et al. (2020, 2021) we used the recently introduced sampling technique *linear elliptical slice sampling (lin-ess, Gessner et al., 2020)* which improves Elliptical Slice Sampling (*ess, Murray et al., 2010*) for multivariate Gaussian distributions truncated on a region defined by linear constraints. In particular this approach derives analytically the acceptable regions on the elliptical slices used in *ess* and guarantees rejection-free sampling. We report the pseudo-code for *lin-ess* in Algorithm 1.¹⁰ Since *lin-ess* is rejection-free,¹¹ we can compute exactly (deterministically) the computation complexity of (8): $O(n^3)$ with storage demands of $O(n^2)$. SkewGPs have similar bottleneck computational complexity of full GPs. Finally, it is **important** to observe that $\mathbf{r}_{1,-\gamma}$ does not depend on test point \mathbf{x} and, therefore, we do not need to re-sample $\mathbf{r}_{1,-\gamma}$ to sample f at another test point

⁹ In Benavoli et al. (2021), we considered mixed preferential-binary observations, where binary judgments (valid or non-valid) together with preference judgments are available.

¹⁰ The pseudo-code reports the main steps of the algorithm. We omitted some implementation details for simplicity, for instance how to select \mathbf{x}_0 or how to deal with empty intersections. The full code of the algorithm is available at <https://github.com/benavoli/SkewGP>.

¹¹ Its computational bottleneck is the Cholesky factorization of the covariance matrix $\tilde{\Gamma}$, same as for sampling from a multivariate Gaussian.

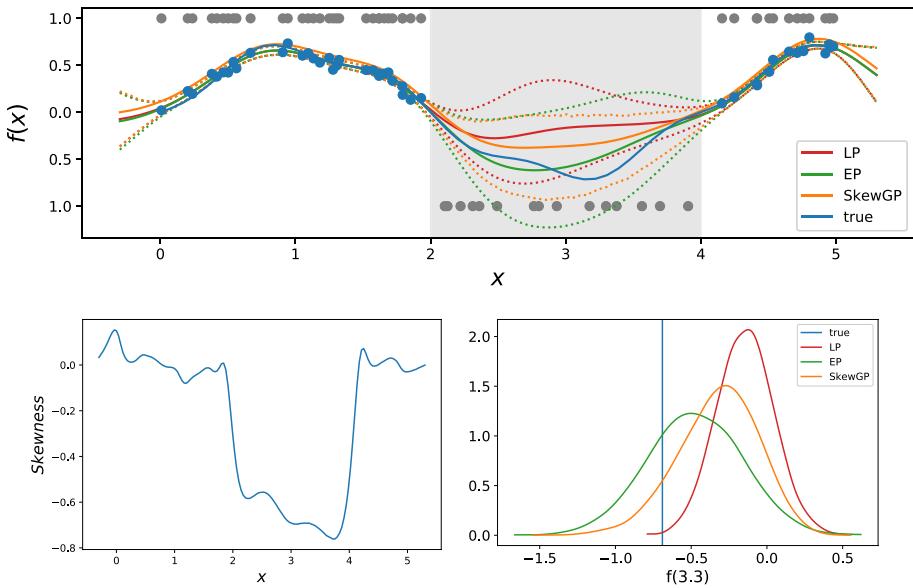


Fig. 7 Mixed numeric and binary regression. Top: dataset, true function (blue), mean and credible intervals for LP, EP, SkewGP. Bottom-left: skewness statistics for the SkewGP posterior as a function of x . Bottom-right: predictive distribution of $f(3.3)$ for the three models (color figure online)

\mathbf{x}' . This is fundamental in active learning and Bayesian optimisation because acquisition functions are functions of \mathbf{x} and we need to optimize them with respect to \mathbf{x} .

Algorithm 1: Lin-Ess to sample from $N_s(\mu, \Gamma)$ truncated below γ

```

input :  $\Gamma \in \mathbb{R}^{s \times s}$  symmetric p.d.,  $\mu, \gamma \in \mathbb{R}^s, m > 0$ 
output:  $\mathbf{x}_{sample} \in \mathbb{R}^{m \times s}, m$  samples from the  $N_s(0, \Gamma)$  truncated below  $\gamma$ .
 $L = \text{Cholesky}(\Gamma);$ 
select  $\mathbf{x}_0$  s.t.  $L\mathbf{x}_0 - \gamma > 0;$ 
while  $i < m$  do
    sample  $\boldsymbol{\nu} \sim N_s(0, \Gamma);$ 
    compute  $\mathbf{r}^2 = \mathbf{x}_0^2 + \boldsymbol{\nu}^2$  (componentwise);
    compute the angles  $\theta_1 = 2 \arctan\left(\frac{\nu - \sqrt{\mathbf{r}^2 - \gamma^2}}{\mathbf{x}_0 + \gamma}\right)$  and
         $\theta_2 = 2 \arctan\left(\frac{\nu + \sqrt{\mathbf{r}^2 - \gamma^2}}{\mathbf{x}_0 + \gamma}\right)$  (componentwise);
    find the largest intesection angles in the domain boundary  $[\theta_{min}, \theta_{max}]$ ;
    sample  $\theta \sim U(\theta_{min}, \theta_{max});$ 
    compute  $\mathbf{x}_{i,sample} = \mathbf{x}_0 \cos(\theta) + \boldsymbol{\nu} \sin(\theta);$ 
    set  $\mathbf{x}_0 = \mathbf{x}_{i,sample};$ 
     $i = i + 1;$ 

```

We demonstrate this computational savings in a 1D classification task – probit likelihood and GP prior on the latent function. We generated 8 datasets of size

$$n \in \{1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500\}$$

as follows $\mathbf{x}_i \sim N(0, I_3)$ for $i = 1, \dots, n$, $\sigma^2 \in \mathcal{U}(1, 10)$, $\ell_j \in \mathcal{U}(0.1, 1.1)$ for $j = 1, \dots, 3$, the latent function f is sampled from a GP with zero mean and Radial Basis Function (RBF) kernel:

$$\Omega(\mathbf{x}, \mathbf{x}') := \sigma^2 \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{x}')^T \begin{bmatrix} \frac{1}{\ell_1^2} & 0 & 0 \\ 0 & \frac{1}{\ell_2^2} & 0 \\ 0 & 0 & \frac{1}{\ell_3^2} \end{bmatrix} (\mathbf{x} - \mathbf{x}')\right), \quad (41)$$

and $y_i \sim \text{Bernoulli}(\Phi(f(x_i)))$ for $i = 1, \dots, n$.

We evaluated the computational load to approximate the posterior using Laplace's method (GP-LP), Expectation Propagation (GP-EP), Hamiltonian Monte Carlo (GP-HMC) and Elliptical Slice Sampling (GP-ess).¹² We compared these four approaches against our SkewGP formulation of the posterior (denoted as SkewGP-LinEss), which allows us exploit the decomposition discussed above to speed-up sampling time. For each value of n , we generated 3 datasets and averaged the running time to: (i) compute 3100 samples (for SkewGP-liness, GP-HMC and GP-ess) using the first 100 samples for tuning; (ii) analytically approximate the posterior for GP-LP and GP-EP.¹³

The computational time in seconds (on a standard laptop) for the 5 methods is shown in Fig. 8(left). It can be noticed that GP-HMC is the slowest method and GP-LP is the fastest; however it is well-known that GP-LP provides a bad approximation of the posterior. Among the remaining three methods, SkewGP-LinEss is the fastest for $n > 2200$. The computational cost of drawing one sample using LinEss is $O(n^2)$. This is comparable with that of ess. However, for large n , the constant can be much smaller for Lin-Ess because, contrarily to ess, Lin-Ess does not need to compute the likelihood at each iteration. This is also the main reason why both GP-EP and GP-ess become much slower at the increase of n . It is **important** to notice that this computational advantage derives by our analytical formulation of the posterior process as a SkewGP.

Our goal was only to compare the running time to obtain 3000 samples but we also checked the convergence of the posterior. Figure 8(right) reports the Gelman–Rubin (GR) (Gelman & Rubin, 1992) statistics for GP-ess and SkewGP-LinEss.¹⁴ (Gelman & Rubin, 1992) suggest that GR values greater than 1.2 should indicate nonconvergence. The GR statistics is below 1.2 for SkewGP-LinEss, while for GP-ess is greater than 1.2 for $n > 2500$. This means that GP-ess needs additional tuning for $n > 2500$ (so it is even slower).

It is also worth to notice that, for SkewGP, two alternative approaches to approximate the posterior mean have been proposed in Cao et al. (2020): (i) using a tile-low-rank Monte

¹² For GP-LP and GP-EP, we use the implementations in GPy (GPy, 2012), while for GP-HMC and GP-Ess we implemented the probabilistic model in PyMC3 (Salvatier et al., 2016) using the default values for acceptance rate. Note that, for GP-HMC and GP-Ess, we apply the Monte Carlo methods directly to the probabilistic model comprising the probit likelihood and the GP prior on the latent function f . In other words we do not exploit the fact that the posterior process is SkewGP.

¹³ Note that, all methods use the true kernel hyperparameters which are kept fixed.

¹⁴ To compute the Gelman–Rubin statistics, we have sampled an extra chain (3100 samples) in each experiment.

Carlo methods for computing multivariate Gaussian probabilities; (ii) approximating functionals of multivariate truncated normals which via a mean-field variational approach. The cost of the tile-low-rank version of is expensive in high dimensional setting (Cao et al., 2020), while the variational approximation scales up to large n . The drawback of the variational approximation is that it is query dependent (it approximates an expectation). Conversely, the *lin-ess* based approach is both efficient and agnostic of the query.

Hyperparameters selection In the GP literature (Rasmussen & Williams, 2006), a GP is usually parametrised with a zero mean function $\xi(\mathbf{x}) = 0$ and a covariance kernel $\Omega(\mathbf{x}, \mathbf{x}')$ which depends on hyperparameters $\theta \in \Theta$. Typically θ contains lengthscale parameters and a variance parameter. For instance, the RBF kernel in (41) has $\theta = [\ell_1, \ell_2, \ell_3, \sigma, \sigma_v]$, where σ_v is the variance of the measurement noise. To fully define a SkewGP, we must also select the latent dimensions s , the additional parameters γ, Γ and the skewness function $\Delta(\mathbf{x})$. The function $\Delta(\mathbf{x})$ may also have hyperparameters, which we denote by \mathbf{u} , so that $\theta = [\ell, \sigma, \sigma_v, \gamma, \Gamma, \mathbf{u}]$. A choice of $\gamma, \Gamma, \mathbf{u}$ is discussed in Benavoli et al. (2020) for classification.

The parameters θ are chosen by maximizing the marginal likelihood, that for (30) and (39), involves computing the CDF Φ_{s+m_a} of a multivariate normal distribution, whose dimensions grows with m_a . Quasi-randomized Monte Carlo methods (Genz, 1992; Genz & Bretz, 2009; Botev, 2017) have been proposed to calculate Φ_{s+m_a} for small m_a (few hundreds observations). These approaches are not in general suitable for medium and large m_a (apart from special cases Phinikettos & Gandy, 2011; Genton et al., 2018; Azzimonti & Ginsbourger, 2018). Another approach is to use EP approximations as in Cunningham et al. (2013), which has a similar computational load of GP-EP for learning the hyperparameters.

We overcome this issue by using the approximation introduced in Benavoli et al. (2020, [Prop.2]):

Proposition 1 *Lower bound of the CDF:*

$$\Phi_m(\tilde{\gamma}; \tilde{\Gamma}) \geq \sum_{i=1}^b \Phi_{|B_i|}(\tilde{\gamma}_{B_i}; \tilde{\Gamma}_{B_i}) - (b-1), \quad (42)$$

where B_1, \dots, B_b are a partition of the training dataset into b random disjoint subsets, $|B_i|$ denotes the number of observations in the i -th element of the partition, $\tilde{\gamma}_{B_i}, \tilde{\Gamma}_{B_i}$ are the parameters of the posterior computed using only the subset B_i of the data (in the experiments $|B_i| = 70$).

To compute $\Phi_{|B_i|}(\cdot)$ in (42), we use the routine proposed in Trinh and Genz (2015), that computes multivariate normal probabilities using bivariate conditioning. This is very fast (fractions of a second for $|B_i| = 70$) and returns a deterministic quantity. We therefore optimise the hyperparameters of the kernel by maximising the lower bound in (42) using the *Dual Annealing* optimization routine.¹⁵ Another possible way to learn the hyperparameters is to exploit the following result.

Proposition 2 *Assume that $\tilde{\gamma}$ does not depend on θ_i ,¹⁶ the derivative of $\log \Phi_m(\tilde{\gamma}; \tilde{\Gamma})$ with respect to θ_i is*

¹⁵ For local search, we use the L-BFGS algorithm with numerical computation of the gradient.

¹⁶ The result can easily be extended to the case where $\tilde{\gamma}$ depends on θ_i by using a change of variables.

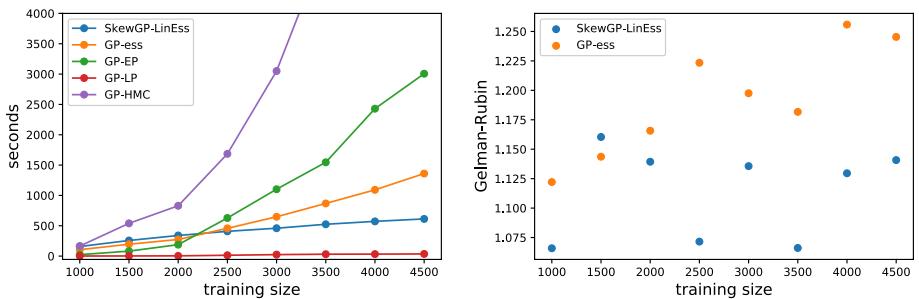


Fig. 8 Left: running time for sampling from the posterior latent function at the training point as a function of the number of training points for the five methods. Right: Gelman–Rubin statistics for GP-ess and SkewGP-ess

$$\frac{\partial}{\partial \theta_i} \log \Phi_m(\tilde{y}; \tilde{\Gamma}) = -\frac{1}{2} \text{Tr}\left(\tilde{\Gamma}^{-1} \tilde{\Gamma}'_{\theta_i}\right) + \frac{1}{2} \text{Tr}\left(\tilde{\Gamma}^{-1} \tilde{\Gamma}'_{\theta_i} \tilde{\Gamma}^{-1} N\right),$$

where $\tilde{\Gamma}'_{\theta_i} = \frac{\partial}{\partial \theta_i} \tilde{\Gamma}$ and $N = \frac{1}{\Phi_m(\tilde{y}; \tilde{\Gamma})} \int \int_{-\infty}^{\tilde{y}} \mathbf{z} \mathbf{z}^T \frac{1}{|2\pi\Gamma|} e^{-\frac{1}{2} \mathbf{z}^T \tilde{\Gamma}^{-1} \mathbf{z}} d\mathbf{z}$.

Note that, M is the second order moment matrix of a multivariate truncated normal. We plan to investigate the use of (batch) stochastic gradient descent methods based on Proposition 2 to learn the hyperparameters on future work. In the next sections, we will use the approximation in (42) to optimise the hyperparameters.

5 Application to active learning and optimisation

The ability of providing a calibrated measure of uncertainty is fundamental for a Bayesian model. In the previous one-dimensional examples, we showed that the true posterior (SkewGP) can be highly skewed and so very different from LP and EP's posteriors. LP tends underestimating the uncertainty and EP tends overestimating it. Both methods are not able to capture the skewness (asymmetry) of the posterior which, in turn, can lead to significant differences in the computed posterior means. In this section we will compare LP, EP and SkewGP in two tasks, *Bayesian Active Learning* and *Bayesian Optimization*, where a wrong representation of uncertainty can lead to a significant performance degradation. In all experiments, we employ a RBF kernel and we estimate its parameters by maximising the respective marginal likelihood for LP, EP and SkewGP. The Python implementation of SkewGPs for regression, classification and mixed problems is available at <https://github.com/benavoli/SkewGP>. Although the derivations in the previous sections were carried out for a generic SkewGP, in the numerical experiments below we will consider as a prior a SkewGP with latent dimension $s = 0$, that is a GP prior. In this way the only difference between LP, EP and SkewGP is the computation of the posterior. For all the models (LP, EP and SkewGP) we compute the acquisition functions via Monte Carlo sampling from the posterior (using 3000 samples).¹⁷

¹⁷ Although for LP and EP some analytical formulas are available (for instance to compute the Upper Credible Bound used in the Bayesian Optimisation section), by using random sampling for all models we remove any advantage of SkewGP over LP and EP due to the additional exploration effect of Monte Carlo sampling.

5.1 Bayesian active learning

A key problem in machine learning and statistics is data efficiency. Active learning is a powerful technique for achieving data efficiency. Instead of collecting and labelling a large dataset, which often can be very costly, in active learning one sequentially acquires labels from an expert only for the most informative data points from a pool of available unlabeled data. After each acquisition step, the newly labelled point is added to the training set, and the model is retrained. This process is repeated until a suitable level of accuracy is achieved. As for optimal experimental design, the goal of active learning is to produce the best model with the least possible data. In a nutshell, Bayesian Active Learning consists of four steps.

1. Train a Bayesian model on the labeled training dataset.
2. Use the trained model to select the next input from the unlabeled data pool.
3. Send the selected input to be labeled (by human experts).
4. Add the labeled samples to the training dataset and repeat the steps.

In step 2, the next input point is usually selected by maximising an information criterion. In this paper, we focus on active learning for binary classification and, consider the *Bayesian Active Learning by Disagreement* (BALD) information criterion introduced in Houlsby et al. (2011).¹⁸

$$\text{Bald}(\mathbf{x}) = h\left(E_{f \sim p(f|W)}(\Phi(f(\mathbf{x}))) - E_{f \sim p(f|W)}(h(\Phi(f(\mathbf{x}))))\right) \quad (43)$$

with $h(p) = -p \log(p) - (1-p) \log(1-p)$ being the binary entropy function of the probability p . The next point \mathbf{x} is therefore selected by maximising $\text{Bald}(\mathbf{x})$.

As Bayesian model we consider a GP classifier (GP prior and probit likelihood) and we compute the expectations in (43) via Monte Carlo sampling from the posterior distribution of the latent function f .

Our aim is to compare the two approximations for the posterior LP, EP versus the full model SkewGP in the above active learning task. We consider eight UCI classification datasets.¹⁹ Table 2 displays the characteristics of the considered datasets.

They are multiclass dataset which we transformed into binary classification problems considering the first class (class 0) versus rest (class 1).²⁰ We start with a randomly selected initial labelled pool of 10 data points and we run active learning sequentially for 90 steps. We perform 10 trials by starting from a different randomly selected initial pool. Figure 9 shows, for each dataset, the accuracy (averaged over the 10 trials) as a function of the number of iterations for LP, EP and SkewGP.

LP performs poorly, in many cases its accuracy does not increase with the number of iterations. SkewGP is always the best algorithm and clearly outperforms EP in 6 datasets. The reason of this significant difference in performance can be understood from Fig. 10

¹⁸ It corresponds to the information gain computed in y space.

¹⁹ These datasets have recently been used for GP classification in Villacampa-Calvo et al. (2020).

²⁰ Our goal in this section is to assess the effect of the approximation of the posterior process in an active learning task by comparing the full posterior (SkewGP) versus LP and EP approximations. For this reason, we have preferred to focus on simple binary classification problems where such effect is easier to understand. However, note that, SkewGP can be applied to multiclass classification problems (without binarization) using the likelihood in (5).

which reports, for each dataset, the percentage (averaged over 10 trials) of unique input instances in the final labelled pool (that is at the end of the 90 iterations). We call it *pool diversity*. A low pool diversity means that the same instance has been included in the pool more than once, which is the result of a poor representation of uncertainty. Note in fact that LP has always the lowest pool diversity. EP, which provides a better approximation of the posterior, performs better, but SkewGP has always the highest pool diversity. SkewGP performs a better exploration of the input space.

5.2 Bayesian optimisation

We consider the problem of finding the global maximum of an unknown function which is expensive to evaluate. For instance, evaluating the function requires conducting an experiment. Mathematically, for a function g on a domain X , the goal is to find a global maximiser \mathbf{x}^o :

$$\mathbf{x}^o = \arg \max_{\mathbf{x} \in X} g(\mathbf{x}).$$

Bayesian optimisation (BO) poses this as a sequential decision problem – a trade-off between learning about the underlying function g (exploration) and capitalizing on this information in order to find the optimum \mathbf{x}^o (exploitation).

In BO, it is usually assumed that $g(\mathbf{x})$ can be evaluated directly (numeric (noisy) observations). However, in many applications, measuring $g(\mathbf{x})$ can be either costly or not always possible. In these cases, the objective function g may only be accessed via preference judgments, such as “this is better than that” between two candidate solutions $\mathbf{x}_i, \mathbf{x}_j$ (like in A/B tests or recommender systems). In such situations, Preferential Bayesian optimization (PBO) (Shahriari et al., 2015; González et al., 2017) or more general algorithms for active preference learning should be adopted (Brochu et al., 2008; Zoghi et al., 2014; Sadigh et al., 2017; Bemporad & Piga, 2021). These approaches require the users to simply compare the final outcomes of two different candidate solutions and indicate which they prefer.

There are also applications where either

1. numeric (noisy) measurements and preference, or
2. numeric (noisy) measurements and binary observations,

may be available together;

In the next sections, we show how we can use SkewGP as surrogated model for BO in these situations and how it outperforms BO based on LP and EP.

5.2.1 Preferential optimisation

The state-of-the-art approach for PBO (Shahriari et al., 2015; González et al., 2017) uses a GP as a prior distribution of the latent preference function and a probit likelihood to model the observed pairwise comparisons. The posterior distribution of the preference function is approximated via the LP approach. In a recent paper (Benavoli et al., 2021), we showed that, by computing the exact posterior (SkewGP), we can outperform LP in PBO. In this section, we further compare SkewGP with EP (but we also report LP for completeness). For EP, we use the formulation for preference learning discussed by Houlsby et al. (2011), which shows that GP preference learning is equivalent to GP classification with a particular

Table 2 Characteristics of the datasets

Dataset	#Instances	#Attributes	#Classes
Glass	214	9	6
New-thyroid	215	5	3
Satellite	6435	36	6
Svmguide2	391	20	3
Vehicle	846	18	4
Vowel	540	10	6
Waveform	5000	21	3
Wine	178	13	3

transformed kernel function. Therefore, we compare three different implementation of Bayesian preferential optimisation based on LP, EP, SkewGP.²¹

In PBO, since g can only be queried via preferences, the next candidate solution \mathbf{x} is selected by optimizing (w.r.t. \mathbf{x}) a *dueling acquisition function* $\alpha(\mathbf{x}, \mathbf{x}_r)$, where \mathbf{x}_r (reference point) is the best point found so far.²² By optimizing $\alpha(\mathbf{x}, \mathbf{x}_r)$, one aims to find a point that is better than \mathbf{x}_r (but also considering the trade-off between exploration and exploitation). After computing the optimum of the the acquisition function, denoted with \mathbf{x}_n , we query the black-box function for $\mathbf{x}_n > \mathbf{x}_r$. If $\mathbf{x}_n > \mathbf{x}_r$, then \mathbf{x}_n becomes the new reference point (\mathbf{x}_r) for the next iteration.

We consider two dueling acquisition functions: (i) Upper Credible Bound (UCB); (ii) Expected Improvement Info Gain (EIIG).

UCB: The dueling UCB acquisition function is defined as the upper bound of the minimum width $\gamma\%$ (in the experiments we use $\gamma = 95$) credible interval of $f(\mathbf{x}) - f(\mathbf{x}_r)$.

EIIG: The dueling EIIG was proposed in Benavoli et al. (2021) by combining the expected probability of improvement (in log-scale) and the dueling BALD (Houlsby et al., 2011) (information gain):

$$k \log \left(E_{f \sim p(f|W)} \left(\Phi \left(\frac{f(\mathbf{x}) - f(\mathbf{x}_r)}{\sqrt{2}\sigma} \right) \right) \right) - IG(\mathbf{x}, \mathbf{x}_r),$$

where $IG(\mathbf{x}, \mathbf{x}_r) = h \left(E_{f \sim p(f|W)} \left(\Phi \left(\frac{f(\mathbf{x}) - f(\mathbf{x}_r)}{\sqrt{2}\sigma} \right) \right) \right) - E_{f \sim p(f|W)} \left(h \left(\Phi \left(\frac{f(\mathbf{x}) - f(\mathbf{x}_r)}{\sqrt{2}\sigma} \right) \right) \right)$. This last acquisition function balances exploration-exploitation by means of the nonnegative scalar k (in the experiments we use $k = 0.1$).

We have considered for $g(\mathbf{x})$ six benchmark functions: $g(x) = \cos(5x) + e^{-\frac{x^2}{2}}$ denoted as ‘1D’ (1D), ‘Six-Hump Camel’ (2D), ‘Langer’ (2D), ‘Colville’ (4D), ‘Rosenbrock5’ (5D) and ‘Hartman6’ (6D). These are minimization problems.²³ Each experiment begins with 5 initial (randomly selected) duels and a total budget of 100 duels are run. Further, each experiment is repeated 20 times with different initialization (the same for all methods) as in González et al. (2017) and Benavoli et al. (2021).

²¹ For LP and EP, we use GPy (2012).

²² More precisely, we assume that preferential observation compares the current input with the best input found so far \mathbf{x}_r .

²³ We converted them into maximizations so that the acquisition functions are well-defined.

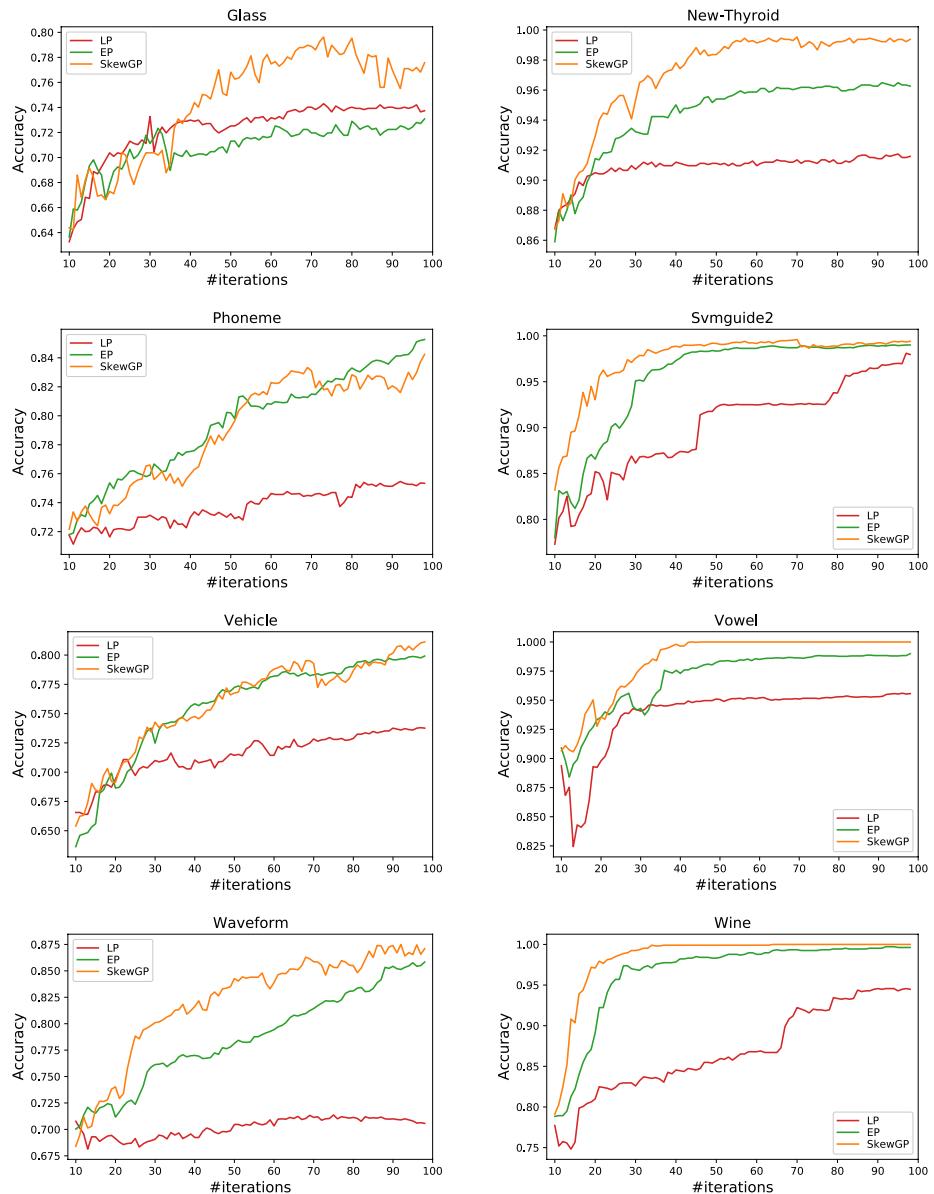


Fig. 9 Averaged results over 10 trials for LP, EP and SkewGP on the 8 UCI datasets. The x-axis represents the number of iterations and the y-axis represents the classification accuracy

Figure 11 reports the performance of the different methods. Consistently across all benchmarks PBO-SkewGP outperforms both PBO-LP and PBO-EP (no matter the acquisition function.) This confirms for PBO what previously noticed for active learning: a wrong uncertainty representation leads to a non optimal exploration of the input space and, therefore, to a slower convergence in active learning tasks.

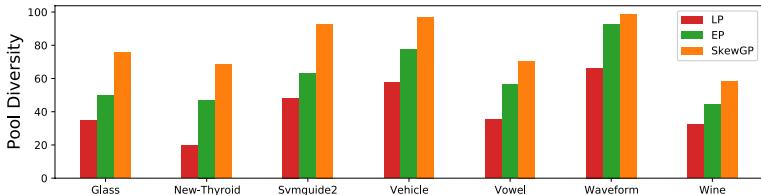


Fig. 10 Averaged pool diversity (percentage) in the active learning task

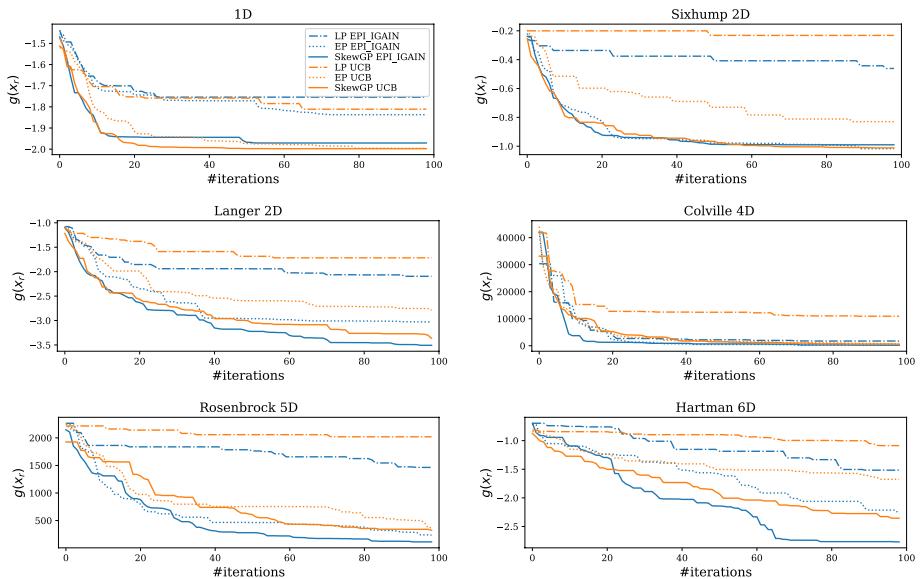


Fig. 11 Averaged results over 20 trials for PBO for LP, EP and SkewGP on the 6 benchmark functions considering 2 different acquisition functions. The x-axis represents the number of iterations and the y-axis represents the value of the true objective function at the current optimum x_r

5.2.2 Mixed numerical and preferential BO

We repeat the previous experiments considering mixed type observations, that is g can be queried directly (numeric data) and via preferences. Each experiment begins with 5 initial (randomly selected) duels (preferences) and 5 numeric (scalar) observations of g . A total budget of 100 iterations is considered and we assume that at the iteration 4, 8, 12, 16, ... g is queried directly and in all the other iterations is queried via preferences (with respect to x_r). Each experiment is repeated 20 times with different initialization (the same for all methods). In this case, we only consider the UCB acquisition function, which is valid for both numeric and preference data.²⁴

²⁴ Observe in fact that the Bald criterion assumes binary observations.

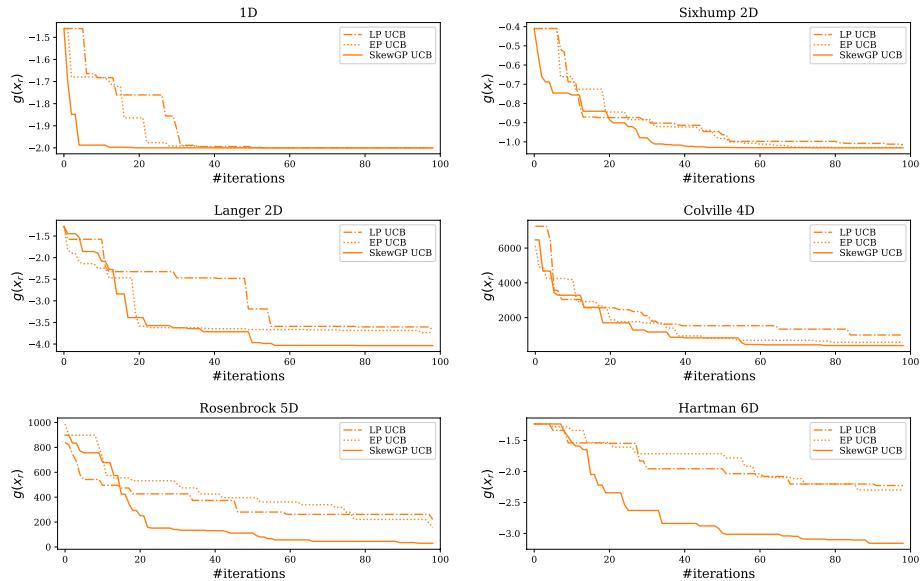


Fig. 12 Averaged results over 20 trials for mixed-PBO for LP, EP and SkewGP on the 6 benchmark functions considering the UCB acquisition function. The x-axis represents the number of iterations and the y-axis represents the value of the true objective function at the current optimum \mathbf{x}_*

Figure 12 shows the performance of the different methods. As expected,²⁵ the differences between the three approaches tend to be smaller than preference-only BO, especially in the low dimensional problems. However, also in this case, BO-SkewGP outperforms both BO-LP and BO-EP.

5.2.3 Safe Bayesian optimisation

Safe BO (Sui et al., 2015) is an extension of BO, which aims to solve the constrained optimization problem:

$$\mathbf{x}^o = \arg \max_{\mathbf{x} \in X} g(\mathbf{x}) \quad \text{s.t.} \quad g(\mathbf{x}) \geq h,$$

where g is an unknown function and $h \in \mathbb{R}$ is a safety threshold. The set $\mathcal{S} = \{\mathbf{x} \in X \mid g(\mathbf{x}) \geq h\}$ is called the *safe* set. This safe set is not known initially (because g is unknown), but is estimated after each function evaluation. Note that, in safe optimisation, the algorithm should avoid (with high probability) unsafe points. Therefore, the challenge is to find an appropriate strategy, which at each iteration not only aims to find the global maximum within the currently known safe set $\hat{\mathcal{S}}$ (exploitation), but also aims to increase the size of $\hat{\mathcal{S}}$ including inputs \mathbf{x} that are known to be safe (exploration). Different strategies and approaches are discussed in Sui et al. (2015).

²⁵ LP, EP and SkewGP coincides in case the observations are all numeric.

In Safe BO, it is usually assumed that g can be queried directly. Given a certain input \mathbf{x}_k , we instead assume that when $g(\mathbf{x}_k) < h$ no output is produced. As discussed for (40), in this case the space of possibility is {(valid, $g(\mathbf{x}_k)$), (non-valid, *None*)} – an input \mathbf{x} is *valid* whether $\mathbf{x} \in \mathcal{S}$.

We can address this BO problem using the general framework (Berkenkamp et al., 2016a, b), that is we employ a GP regression model to learn g using only *valid* data and a GP classifier (with two classes *valid* and *non-valid*) to learn the sign of the constraint.²⁶

We compare the above approach with a BO strategy that uses SkewGP as conjugate prior for the likelihood (40) and, therefore, addresses the two types of observations as a mixed numeric and binary regression problem. As acquisition function we use UCB plus a term that penalises violations of the constraint with high probability $1000(P[f < 0] < 0.7)$, where P is computed by sampling the function f from SkewGP.

In the experiments, we sample (using rejection sampling) a set of 100 random functions g from a zero-mean GP with RBF kernel (with variance 2 and lengthscale uniformly sampled in the interval [1, 2]), which satisfies $g(0) \geq 0$. This ensures that $x = 0$ is an initial safe input. This point serves as a starting point for both the algorithms.

The dependence between the kernel hyperparameters and the acquired data can lead to poor results in BO (especially in the first iterations). In fact, hyperparameters estimated by maximising the marginal likelihood can lead to a GP estimate which does not have a calibrated uncertainty. In Safe BO (Berkenkamp et al., 2016a, b), one critically relies on the uncertainty to guarantee safety (avoiding constraint violation with high probability). As a consequence, the hyperparameters are kept fixed and we treat the kernel as a prior over functions in the true Bayesian sense – the kernel hyperparameters encode our prior knowledge about the functions. Therefore, in the experiment we set the variance of the RBF kernel to 2 and the lengthscale to 1.5.²⁷

Figures 13 and 14 report 7 iterations of the two approaches, which we call as SafeOpt and SkewGP-BO, for one of the 100 trials. Figure 13(Iter. 0, left) shows the true function g in gray; the estimated GP regression model (mean and 99.7 credible interval) after observing the safe point $x = 0$ (top plot); the estimated GP classifier (mean and 99.7 credible interval of the latent function) after observing the valid point $x = 0$ (bottom plot). The x in the plots corresponds the value of $g(0)$ in the top-plot and to the binary observation 1 (which means valid) in the bottom-plot. Figure 13(Iter. 0, right) shows the posterior mean and 99.7 credible interval for SkewGP, which solves the mixed numeric and binary regression problem. We use xs to mark numeric observations and circle to mark binary observations (the ordinate +4 corresponds to valid and -4 to non-valid).

At iteration 1, SkewGP-BO selects a *non-valid* point Fig. 13(Iter. 1, right), while SafeOpt selects a *valid* one, finding a better maximum Fig. 13(Iter. 1, left).

At iteration 2, SkewGP-BO selects a *valid* point and finds a better maximum Fig. 13(Iter. 2, right), while GP-SafeOpt violates the constraint Fig. 13(Iter. 2, left).

At iteration 3, SkewGP-BO selects a *non-valid* point Fig. 13(Iter. 3, right), while SafeOpt selects a *valid* point Fig. 13(Iter. 3, left).

At iteration 4 and 5, SkewGP-BO finds the maximum Fig. 14(Iter. 4 and 5, right), while SafeOpt explores the *valid* region Fig. 14(Iter. 4 and 5, left).

At iteration 6, SafeOpt violates the constraint Fig. 14(Iter. 6, left).

²⁶ We have used the SafeOpt library (Berkenkamp et al., 2016a, b) with $\beta = 0.85$ and used the EP approximation for GP classification.

²⁷ The lengthscale can be different from the true one which is uniformly sampled in [1, 2].

At iteration 7, SafeOpt finally finds the maximum Fig. 14(Iter. 7, left).

Figure 15 reports the averaged results over 100 trials for SafeOpt versus SkewGP-BO. The x-axis represents the number of iterations and the y-axis represents the value of the true objective function at the current optimum \mathbf{x}_r . SkewGP-BO achieves the best performance. Both SafeOpt and SkewGP-BO are very unlikely to explore the unsafe region (51 violations for SafeOpt and 53 for SkewGP-BO in the 3000 acquisitions). This shows that, when Safe BO can be modelled as a mixed problem,²⁸ SkewGP provides faster convergence than using two separated surrogated models (GPs) for objective and constraint.

6 Conclusions

We have shown that Skew Gaussian process (SkewGP) are conjugate to the normal likelihood, the probit affine likelihood and with their product, and provided marginals and closed form conditionals. SkewGPs allow us to compute the exact posterior in all these cases and, therefore, provide an accurate representation of the uncertainty. From a theoretical point of view, this shows that SkewGPs encompass GPs in nonparametric regression, classification, preference learning and mixed problems. From a practical point of view, we compared SkewGPs with GPs (Laplace and Expectation Propagation approximation) in two tasks, Bayesian Active Learning and Bayesian Optimization, where a wrong representation of uncertainty can lead to a significant performance degradation. SkewGP achieved an improved performance over GPs (Laplace’s method and Expectation Propagation approximations).

As future work, we plan to investigate the possibility of using inducing points, as for sparse GPs (Quiñonero-Candela & Rasmussen, 2005; Snelson & Ghahramani, 2006; Titsias, 2009; Hensman et al., 2013; Hernández-Lobato & Hernández-Lobato, 2016; Bauer et al., 2016; Schuerch et al., 2020), to reduce the computational load for matrix operations (complexity $O(n^3)$ with storage demands of $O(n^2)$). We also plan to derive tighter approximations of the marginal likelihood. Moreover, we plan to study different ways to parametrize the skewness matrix Δ and select the prior parameters γ, Γ , which allows one to fully exploit the flexibility of SkewGPs. We also plan to investigate the possibility of extending these results to a larger class of processes, *Extended SkewGPs*, studied in Alodat and Al-Rawwash (2014).

Appendix: Proofs of the results in the paper

Proof Lemma 1 and Corollary 1 The likelihood is $\phi_{m_r}(Y - Cf; R)$, where $Y \in \mathbb{R}^{m_a}$ is the vector of observations; $f \in \mathbb{R}^n$ is the vector of function values at the input points \mathbf{x}_i for $i = 1, \dots, n$ and the prior is:

$$p(f) = \phi_n(f - \xi; \Omega) \frac{\Phi_s(\gamma + \Delta^T \bar{\Omega}^{-1} D_\Omega^{-1} (f - \xi); \Gamma - \Delta^T \bar{\Omega}^{-1} \Delta)}{\Phi_s(\gamma; \Gamma)}. \quad (44)$$

²⁸ The framework for safe BO in Berkenkamp et al. (2016a, b) is more general and cannot always be expressed as a mixed problem.

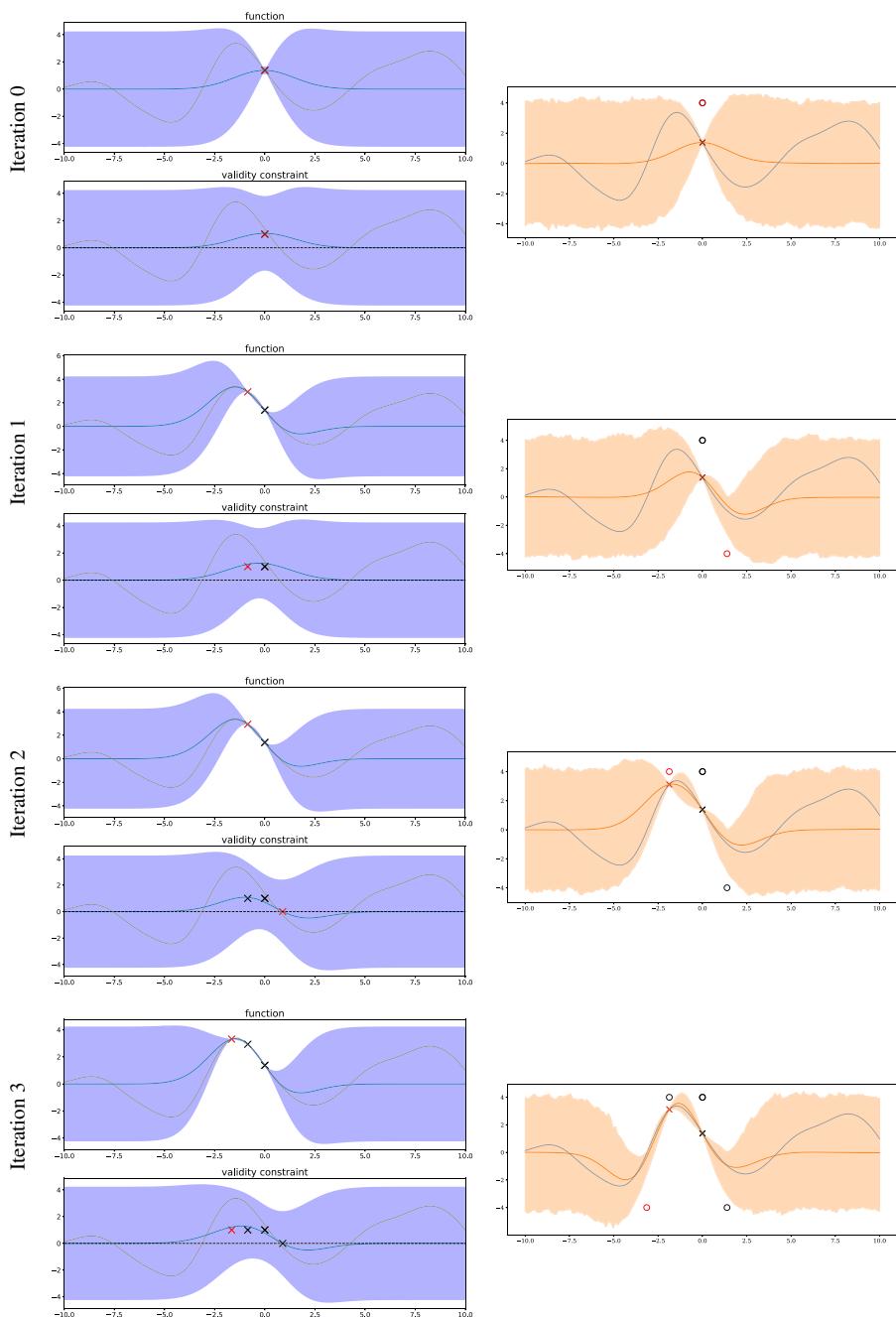


Fig. 13 Iterations 0–3, GP-SafeOpt (left) and SkewGP-SafeOpt (right)

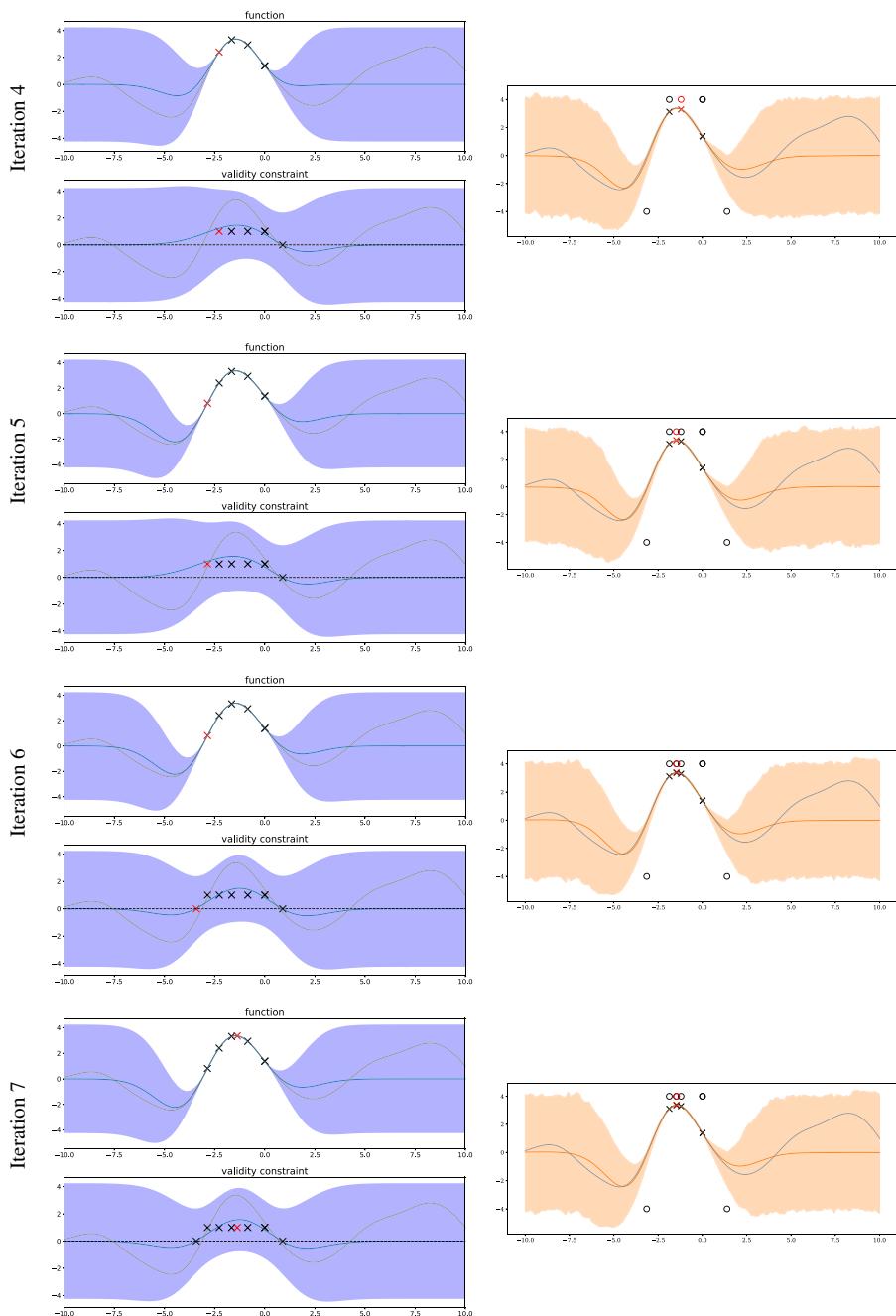
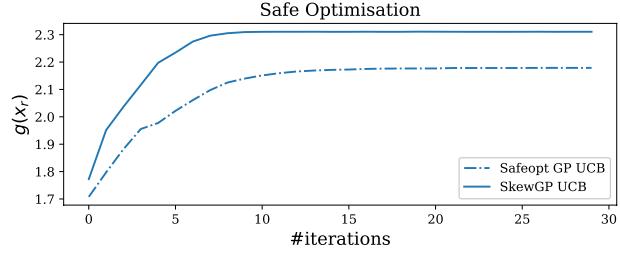


Fig. 14 Iterations 4–7, GP-SafeOpt (left) and SkewGP-SafeOpt (right)

Fig. 15 Averaged results over 100 trials for GP-SafeOpt versus SkewGP. The x-axis represents the number of iterations and the y-axis represents the value of the true objective function at the current optimum \mathbf{x}_r



First, note that

$$\phi_{m_r}(Y - Cf; R)\phi_n(f; \xi, \Omega) \propto \phi_n(f - \xi_p; \Omega_p),$$

with

$$\xi_p = \xi + \Omega C^T (C \Omega C^T + R)^{-1} (Y - C \xi), \quad (45)$$

$$\Omega_p = \Omega - \Omega C^T (C \Omega C^T + R)^{-1} C \Omega. \quad (46)$$

Now consider $\Phi_s(\gamma + \Delta^T \bar{\Omega}^{-1} D_{\Omega}^{-1} (f - \xi); \Gamma - \Delta^T \bar{\Omega}^{-1} \Delta)$ and observe that

$$\begin{aligned} \gamma + \Delta^T \bar{\Omega}^{-1} D_{\Omega}^{-1} (f - \xi) &= \gamma + \Delta^T \bar{\Omega}^{-1} D_{\Omega}^{-1} (f - \xi_p + \xi_p - \xi) \\ &= \gamma + \Delta^T \bar{\Omega}^{-1} D_{\Omega}^{-1} (\xi_p - \xi) + \Delta^T \bar{\Omega}^{-1} D_{\Omega}^{-1} (f - \xi_p) \\ &= \gamma + \Delta^T \bar{\Omega}^{-1} D_{\Omega}^{-1} (\xi_p - \xi) + \Delta^T \bar{\Omega}^{-1} D_{\Omega}^{-1} D_{\Omega p} \bar{\Omega}_p \bar{\Omega}_p^{-1} D_{\Omega p}^{-1} (f - \xi_p) \\ &= \gamma_p + \Delta_p^T \bar{\Omega}_p^{-1} D_{\Omega p}^{-1} (f - \xi_p) \end{aligned}$$

with

$$\gamma_p = \gamma + \Delta^T \bar{\Omega}^{-1} D_{\Omega}^{-1} (\xi_p - \xi) \quad (47)$$

$$\Delta_p^T = \Delta^T \bar{\Omega}^{-1} D_{\Omega}^{-1} D_{\Omega p} \bar{\Omega}_p \quad (48)$$

Finally observe that

$$\Gamma - \Delta^T \bar{\Omega}^{-1} \Delta = \Gamma - \Delta^T \bar{\Omega}^{-1} \Delta + \Delta_p^T \bar{\Omega}_p^{-1} \Delta_p - \Delta_p^T \bar{\Omega}_p^{-1} \Delta_p = \Gamma_p - \Delta_p^T \bar{\Omega}_p^{-1} \Delta_p$$

with $\Gamma_p = \Gamma - \Delta^T \bar{\Omega}^{-1} \Delta + \Delta_p^T \bar{\Omega}_p^{-1} \Delta_p$. Putting everything together

$$p(f|Y) = \phi_n(f - \xi_p; \Omega_p) \frac{\Phi_s(\gamma_p + \Delta_p^T \bar{\Omega}_p^{-1} D_{\Omega p}^{-1} (f - \xi_p); \Gamma_p - \Delta_p^T \bar{\Omega}_p^{-1} \Delta_p)}{\Phi_s(\gamma_p; \Gamma_p)}$$

By considering the derivations of the above expression, we can derive the marginal likelihood

$$\phi_n(Y - C\xi; C\Omega C^T + R) \frac{\Phi_s(\gamma_p, \Gamma_p)}{\Phi_s(\gamma, \Gamma)}.$$

□

Proof Theorem 1 Consider the test point $\mathbf{x} \in \mathbb{R}^d$ and the vector $\hat{f} = \begin{bmatrix} f(X) \\ f(\mathbf{x}) \end{bmatrix} := [\mathbf{f}^\top \ f_*]^\top$ we have

$$p(\mathbf{f}, f_*) = \text{SkewGP}\left(\begin{bmatrix} \xi(X) \\ \xi(\mathbf{x}) \end{bmatrix}, \begin{bmatrix} \Omega(X, X) & \Omega(X, \mathbf{x}) \\ \Omega(\mathbf{x}, X) & \Omega(\mathbf{x}, \mathbf{x}) \end{bmatrix}, \begin{bmatrix} \Delta(X) \\ \Delta(\mathbf{x}) \end{bmatrix}, \gamma, \Gamma\right)$$

and the predictive distribution is by definition

$$\begin{aligned} p(f_* \mid Y) &= \int p(f_* \mid \mathbf{f}) p(\mathbf{f} \mid Y) d\mathbf{f} \\ &= \int p(f_* \mid \mathbf{f}) \frac{p(Y \mid \mathbf{f}) p(\mathbf{f})}{p(Y)} d\mathbf{f} \\ &\propto \int p(f_*, \mathbf{f}) p(Y \mid \mathbf{f}) d\mathbf{f} \end{aligned}$$

We can then apply Lemma 1 with \hat{f} and the likelihood $p([C \mid \mathbf{0}] \mid \hat{f}) = \phi_n\left(Y - [C \mid \mathbf{0}], \begin{bmatrix} f(X) \\ f(\mathbf{x}) \end{bmatrix}; R\right)$ which results in a posterior distribution

$$p\left(\begin{bmatrix} f(X) \\ f(\mathbf{x}) \end{bmatrix} \mid [Y \mid \mathbf{0}]\right) = \text{SUN}_{n+1,s}(\hat{\xi}, \hat{\Omega}, \hat{\Delta}, \hat{\gamma}, \hat{\Gamma})$$

with

$$\hat{\xi} = \begin{bmatrix} \xi(X) \\ \xi(\mathbf{x}) \end{bmatrix} + \begin{bmatrix} \Omega(X, X) \\ \Omega(\mathbf{x}, X) \end{bmatrix} C^T (C\Omega C^T + R)^{-1} (Y - C\xi(X)) \quad (49)$$

$$\hat{\Omega} = \begin{bmatrix} \Omega(X, X) & \Omega(X, \mathbf{x}) \\ \Omega(\mathbf{x}, X) & \Omega(\mathbf{x}, \mathbf{x}) \end{bmatrix} - \begin{bmatrix} \Omega(X, X) \\ \Omega(\mathbf{x}, X) \end{bmatrix} C^T (C\Omega C^T + R)^{-1} C \begin{bmatrix} \Omega(X, X) & \Omega(X, \mathbf{x}) \end{bmatrix} \quad (50)$$

For $\hat{\Delta}$, from Lemma 1, one has

$$\begin{aligned} \hat{\Delta} &= D_{\hat{\Omega}}^{-1} \hat{\Omega} \begin{bmatrix} \Omega(X, X) & \Omega(X, \mathbf{x}) \\ \Omega(\mathbf{x}, X) & \Omega(\mathbf{x}, \mathbf{x}) \end{bmatrix}^{-1} D_\Omega \begin{bmatrix} \Delta(X) \\ \Delta(\mathbf{x}) \end{bmatrix} = D_{\hat{\Omega}}^{-1} \begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix} D_\Omega \begin{bmatrix} \Delta(X) \\ \Delta(\mathbf{x}) \end{bmatrix} \\ &= \begin{bmatrix} D_{\hat{\Omega}(X,X)}^{-1} A_1 D_{\Omega(X,X)} \Delta(X) + D_{\hat{\Omega}(X,X)}^{-1} A_2 D_{\Omega(\mathbf{x},X)} \Delta(\mathbf{x}) \\ D_{\hat{\Omega}(\mathbf{x},X)}^{-1} A_3 D_{\Omega(X,X)} \Delta(X) + D_{\hat{\Omega}(\mathbf{x},X)}^{-1} A_4 D_{\Omega(\mathbf{x},X)} \Delta(\mathbf{x}) \end{bmatrix} \end{aligned}$$

where

$$\begin{aligned} \begin{bmatrix} A_1 & A_2 \\ A_3 & A_4 \end{bmatrix} &= \hat{\Omega} \begin{bmatrix} \Omega(X, X) & \Omega(X, \mathbf{x}) \\ \Omega(\mathbf{x}, X) & \Omega(\mathbf{x}, \mathbf{x}) \end{bmatrix}^{-1} \\ &= \hat{\Omega} \begin{bmatrix} \Omega(X, X)^{-1} + \Omega(X, X)^{-1} \Omega(X, \mathbf{x}) (\Omega(\mathbf{x}, \mathbf{x}) - \Omega(\mathbf{x}, X) \Omega(X, X)^{-1} \Omega(X, \mathbf{x}))^{-1} \Omega(\mathbf{x}, X) \Omega(X, X)^{-1} \\ -\Omega(X, X)^{-1} \Omega(X, \mathbf{x}) (\Omega(\mathbf{x}, \mathbf{x}) - \Omega(\mathbf{x}, X) \Omega(X, X)^{-1} \Omega(X, \mathbf{x}))^{-1} \\ -(\Omega(\mathbf{x}, \mathbf{x}) - \Omega(\mathbf{x}, X) \Omega(X, X)^{-1} \Omega(X, \mathbf{x}))^{-1} \Omega(\mathbf{x}, X) \Omega(X, X)^{-1} (\Omega(\mathbf{x}, \mathbf{x}) - \Omega(\mathbf{x}, X) \Omega(X, X)^{-1} \Omega(X, \mathbf{x}))^{-1} \end{bmatrix} \end{aligned}$$

Since we are interested in the marginal, we only need to compute A_3 and A_4 :

$$\begin{aligned} A_3 &= \left(\Omega(\mathbf{x}, X) - \Omega(\mathbf{x}, X) C^T (C \Omega(X, X) C^T + R)^{-1} C \Omega(X, X) \right) \\ &\quad \left(\Omega(X, X)^{-1} + \Omega(X, X)^{-1} \Omega(X, \mathbf{x}) (\Omega(\mathbf{x}, \mathbf{x}) - \Omega(\mathbf{x}, X) \Omega(X, X)^{-1} \Omega(X, \mathbf{x}))^{-1} \right. \\ &\quad \left. \Omega(\mathbf{x}, X) \Omega(X, X)^{-1} \right) \\ &\quad - \left(\Omega(\mathbf{x}, \mathbf{x}) - \Omega(\mathbf{x}, X) C^T (C \Omega(X, X) C^T + R)^{-1} C \Omega(X, \mathbf{x}) \right) \\ &\quad \left(\Omega(\mathbf{x}, \mathbf{x}) - \Omega(\mathbf{x}, X) \Omega(X, X)^{-1} \Omega(X, \mathbf{x}) \right)^{-1} \Omega(\mathbf{x}, X) \Omega(X, X)^{-1} \\ &= -\Omega(\mathbf{x}, X) C^T (C \Omega(X, X) C^T + R)^{-1} C \end{aligned}$$

and

$$\begin{aligned} A_4 &= -\left(\Omega(\mathbf{x}, X) - \Omega(\mathbf{x}, X) C^T (C \Omega(X, X) C^T + R)^{-1} C \Omega(X, X) \right) \\ &\quad \Omega(X, X)^{-1} \Omega(X, \mathbf{x}) (\Omega(\mathbf{x}, \mathbf{x}) - \Omega(\mathbf{x}, X) \Omega(X, X)^{-1} \Omega(X, \mathbf{x}))^{-1} \\ &\quad + \left(\Omega(\mathbf{x}, \mathbf{x}) - \Omega(\mathbf{x}, X) C^T (C \Omega(X, X) C^T + R)^{-1} C \Omega(X, \mathbf{x}) \right) \\ &\quad \left(\Omega(\mathbf{x}, \mathbf{x}) - \Omega(\mathbf{x}, X) \Omega(X, X)^{-1} \Omega(X, \mathbf{x}) \right)^{-1} = I_{\mathbf{x}} \end{aligned}$$

Finally, we will show that:

$$\begin{aligned} \hat{\gamma} &= \gamma_p, \\ \hat{\Gamma} &= \Gamma_p. \end{aligned}$$

By Lemma 1

$$\begin{aligned} \hat{\gamma} &= \gamma + [\Delta(X), \Delta(\mathbf{x})] D_{\Omega} \Omega^{-1} (\xi_p - \xi) \\ &= \gamma + [\Delta(X), \Delta(\mathbf{x})] D_{\Omega} \begin{bmatrix} \Omega(X, X) & \Omega(X, \mathbf{x}) \\ \Omega(\mathbf{x}, X) & \Omega(\mathbf{x}, \mathbf{x}) \end{bmatrix}^{-1} \\ &\quad \begin{bmatrix} \Omega(X, X) \\ \Omega(\mathbf{x}, X) \end{bmatrix} C^T (C \Omega C^T + R)^{-1} (Y - C \xi(X)) \end{aligned}$$

Note that

$$\begin{bmatrix} \Omega(X, X) & \Omega(X, \mathbf{x}) \\ \Omega(\mathbf{x}, X) & \Omega(\mathbf{x}, \mathbf{x}) \end{bmatrix}^{-1} \begin{bmatrix} \Omega(X, X) \\ \Omega(\mathbf{x}, X) \end{bmatrix} = \begin{bmatrix} \Omega(X, X)^{-1} \\ 0 \end{bmatrix}$$

and so $\hat{\gamma} = \gamma_p$ as in Lemma 1. Similarly, one can show that $\hat{\Gamma} = \Gamma_p$.

By exploiting the marginalization properties of the SUN distribution, see Sect. 2.3 and in particular (9), we can derive that

$$p(f(\mathbf{x}) \mid C, Y, f(X)) = \text{SUN}_{1,s}(\tilde{\xi}(\mathbf{x}), \tilde{\Omega}(\mathbf{x}, \mathbf{x}), \tilde{\Delta}(\mathbf{x}), \gamma_p, \Gamma_p). \quad (51)$$

with mean, scale, and skewness functions:

$$\tilde{\xi}(\mathbf{x}) = \xi(\mathbf{x}) + \Omega(\mathbf{x}, X)C^T(C\Omega(X, X)C^T + R)^{-1}(Y - C\xi(X)), \quad (52)$$

$$\tilde{\Omega}(\mathbf{x}, \mathbf{x}) = \Omega(\mathbf{x}, \mathbf{x}) - \Omega(\mathbf{x}, X)C^T(C\Omega(X, X)C^T + R)^{-1}C\Omega(X, \mathbf{x}), \quad (53)$$

$$\tilde{\Delta}(\mathbf{x}) = D_{\hat{\Omega}(\mathbf{x}, \mathbf{x})}^{-1}D_{\Omega(\mathbf{x}, \mathbf{x})}\Delta(\mathbf{x}) - D_{\hat{\Omega}(\mathbf{x}, \mathbf{x})}^{-1}\Omega(\mathbf{x}, X)C^T(C\Omega(X, X)C^T + R)^{-1}CD_{\Omega(X, X)}\Delta(X). \quad (54)$$

□

Lemma 2 This proof is based on the proofs in Durante (2019, Th.1 and Co.4), and consider the more general class of affine probit likelihoods.

The joint distribution of $f(X), W, Z$ is

$$\begin{aligned} & p(W, Z[f(X)])p(f(X)) \\ &= \Phi_{m_a}(Z + W\mathbf{f}; \Sigma) \text{SUN}_{n,s}(\xi, \Omega, \Delta, \gamma, \Gamma) \\ &= \Phi_{m_a}(Z + W\mathbf{f}; \Sigma) \phi_n(\mathbf{f} - \xi; \Omega) \frac{\Phi_s(\gamma + \Delta^T \bar{\Omega}^{-1} D_{\bar{\Omega}}^{-1}(\mathbf{f} - \xi); \Gamma - \Delta^T \bar{\Omega}^{-1} \Delta)}{\Phi_s(\gamma; \Gamma)} \\ &\propto \phi_n(\mathbf{f} - \xi; \Omega) \Phi_{m_a}(Z + W\mathbf{f}; \Sigma) \Phi_s(\gamma + \Delta^T \bar{\Omega}^{-1} D_{\bar{\Omega}}^{-1}(\mathbf{f} - \xi); \Gamma - \Delta^T \bar{\Omega}^{-1} \Delta), \end{aligned} \quad (55)$$

where we denoted $\mathbf{f} = f(X) \in \mathbb{R}^n$ and omitted the dependence on X . First, note that

$$\begin{aligned} \Phi_{m_a}(Z + W\mathbf{f}; \Sigma) &= \Phi_{m_a}(Z + W\xi + (\bar{\Omega}D_{\bar{\Omega}}W^T)^T \bar{\Omega}^{-1} D_{\bar{\Omega}}^{-1}(\mathbf{f} - \xi); (W\Omega W^T + \Sigma) \\ &\quad - WD_{\bar{\Omega}}\bar{\Omega}D_{\bar{\Omega}}W^T). \end{aligned}$$

Therefore, we can write

$$\begin{aligned} & \Phi_{m_a}(Z + W\mathbf{f}; \Sigma) \Phi_s(\gamma + \Delta^T \bar{\Omega}^{-1} D_{\bar{\Omega}}^{-1}(\mathbf{f} - \xi); \Gamma - \Delta^T \bar{\Omega}^{-1} \Delta) \\ &= \Phi_{m_a}(Z + W\xi + (\bar{\Omega}D_{\bar{\Omega}}W^T)^T \bar{\Omega}^{-1} D_{\bar{\Omega}}^{-1}(\mathbf{f} - \xi); (W\Omega W^T + \Sigma) - WD_{\bar{\Omega}}\bar{\Omega}D_{\bar{\Omega}}W^T) \\ &\quad \cdot \Phi_s(\gamma + \Delta^T \bar{\Omega}^{-1} D_{\bar{\Omega}}^{-1}(\mathbf{f} - \xi); \Gamma - \Delta^T \bar{\Omega}^{-1} \Delta) \\ &= \Phi_{s+m_a}(m; M) \end{aligned} \quad (56)$$

with

$$m = \begin{bmatrix} \gamma + \Delta^T \bar{\Omega}^{-1} D_{\bar{\Omega}}^{-1}(\mathbf{f} - \xi) \\ Z + W\xi + (\bar{\Omega}D_{\bar{\Omega}}W^T)^T \bar{\Omega}^{-1} D_{\bar{\Omega}}^{-1}(\mathbf{f} - \xi) \end{bmatrix},$$

and

$$M = \begin{bmatrix} \Gamma - \Delta^T \bar{\Omega}^{-1} \Delta & 0 \\ 0 & (W\Omega W^T + \Sigma) - WD_{\bar{\Omega}}\bar{\Omega}D_{\bar{\Omega}}W^T \end{bmatrix}.$$

From (55)–(56) and the definition of the PDF of the SUN distribution (6), we can easily show that we can rewrite (55) as a SUN distribution with updated parameters:

$$\begin{aligned}\tilde{\xi} &= \xi, \\ \tilde{\Omega} &= \Omega, \\ \tilde{\Delta} &= [\Delta, \tilde{\Omega} D_\Omega W^T], \\ \tilde{\gamma} &= [\gamma, Z + W\xi]^T, \\ \tilde{\Gamma} &= \begin{bmatrix} \Gamma & \Delta^T D_\Omega W^T \\ WD_\Omega \Delta & (W\Omega W^T + \Sigma) \end{bmatrix}.\end{aligned}$$

Theorem 2 Consider the test point $\mathbf{x} \in \mathbb{R}^d$ and the vector $\hat{f} = \begin{bmatrix} f(X) \\ f(\mathbf{x}) \end{bmatrix} := [\mathbf{f} \ f_*]$ we have

$$p(\mathbf{f}, f_*) = \text{SkewGP}\left(\begin{bmatrix} \xi(X) \\ \xi(\mathbf{x}) \end{bmatrix}, \begin{bmatrix} \Omega(X, X) & \Omega(X, \mathbf{x}) \\ \Omega(\mathbf{x}, X) & \Omega(\mathbf{x}, \mathbf{x}) \end{bmatrix}, \begin{bmatrix} \Delta(X) \\ \Delta(\mathbf{x}) \end{bmatrix}, \gamma, \Gamma\right)$$

and the predictive distribution is by definition

$$\begin{aligned}p(f_* | W, Z) &= \int p(f_* | \mathbf{f}) p(\mathbf{f} | Z, W) d\mathbf{f} \\ &= \int p(f_* | \mathbf{f}) \frac{p(Z, W | \mathbf{f}) p(\mathbf{f})}{p(Z, W)} d\mathbf{f} \\ &\propto \int p(f_*, \mathbf{f}) p(Z, W | \mathbf{f}) d\mathbf{f}\end{aligned}$$

We can then apply Lemma 2 with \hat{f} and the likelihood

$$p\left(\begin{bmatrix} Z \\ \mathbf{0} \end{bmatrix}, [W \ \mathbf{0}] | \hat{f}\right) = \Phi_{m_a+1}\left(\begin{bmatrix} Z \\ \mathbf{0} \end{bmatrix} + [W \ \mathbf{0}] \begin{bmatrix} f(X) \\ f(\mathbf{x}) \end{bmatrix}; \begin{bmatrix} \Sigma & \mathbf{0} \\ \mathbf{0} & 1 \end{bmatrix}\right)$$

which does not depend on $f(\mathbf{x})$. This results in a posterior distribution

$$p\left(\begin{bmatrix} f(X) \\ f(\mathbf{x}) \end{bmatrix} | \begin{bmatrix} Z \\ \mathbf{0} \end{bmatrix}, [W \ \mathbf{0}]\right) = \text{SUN}_{n+1, m_a}(\hat{\xi}, \hat{\Omega}, \hat{\Delta}, \hat{\gamma}, \hat{\Gamma})$$

with

$$\begin{aligned}\hat{\xi} &= [\xi(X) \ \xi(\mathbf{x})]^T \\ \hat{\Omega} &= \begin{bmatrix} \Omega(X, X) & \Omega(X, \mathbf{x}) \\ \Omega(\mathbf{x}, X) & \Omega(\mathbf{x}, \mathbf{x}) \end{bmatrix} \\ \hat{\Delta} &= \begin{bmatrix} \Delta(X) \\ \Delta(\mathbf{x}) \end{bmatrix} - D_\Omega^{-1} \begin{bmatrix} \Omega(X, X) & \Omega(X, \mathbf{x}) \\ \Omega(\mathbf{x}, X) & \Omega(\mathbf{x}, \mathbf{x}) \end{bmatrix} [W \ \mathbf{0}]^T = \begin{bmatrix} \Delta(X) & D_{\Omega(X,X)}^{-1} \Omega(X, X) W^T \\ \Delta(\mathbf{x}) & D_{\Omega(\mathbf{x},\mathbf{x})}^{-1} \Omega(\mathbf{x}, X) W^T \end{bmatrix} \\ \hat{\gamma} &= \tilde{\gamma}, \\ \hat{\Gamma} &= \tilde{\Gamma},\end{aligned}$$

with $\tilde{\gamma}, \tilde{\Gamma}$ from Lemma 2. By exploiting the marginalization properties of the SUN distribution, see Sect. 2.3, we obtain

$$p(f(\mathbf{x}) | W, f(X)) = \text{SUN}_{1, s+m_a}(\xi(\mathbf{x}), \Omega(\mathbf{x}, \mathbf{x}), [\Delta(\mathbf{x}) \ D_{\Omega(\mathbf{x},\mathbf{x})}^{-1} \Omega(\mathbf{x}, X) W^T], \tilde{\gamma}, \tilde{\Gamma}). \quad (57)$$

Proposition 2

$$\frac{\partial}{\partial \theta_i} \log \Phi_m(\tilde{\gamma}; \tilde{\Gamma}) = \frac{1}{\Phi_m(\tilde{\gamma}; \tilde{\Gamma})} \iint_{-\infty}^{\tilde{\gamma}} \frac{\partial}{\partial \theta_i} \left(\frac{1}{|2\pi\Gamma|} e^{-\frac{1}{2}\mathbf{z}^T \tilde{\Gamma}^{-1} \mathbf{z}} d\mathbf{z} \right) \quad (58)$$

$$= -\frac{1}{2} \frac{1}{\Phi_m(\tilde{\gamma}; \tilde{\Gamma})} \text{Tr} \left(\tilde{\Gamma}^{-1} \tilde{\Gamma}'_{\theta_i} \right) \iint_{-\infty}^{\tilde{\gamma}} \frac{1}{|2\pi\Gamma|} e^{-\frac{1}{2}\mathbf{z}^T \tilde{\Gamma}^{-1} \mathbf{z}} d\mathbf{z} \quad (59)$$

$$+ \frac{1}{2} \frac{1}{\Phi_m(\tilde{\gamma}; \tilde{\Gamma})} \iint_{-\infty}^{\tilde{\gamma}} \text{Tr} \left(\tilde{\Gamma}^{-1} \tilde{\Gamma}'_{\theta_i} \tilde{\Gamma}^{-1} \mathbf{z} \mathbf{z}^T \right) \frac{1}{|2\pi\Gamma|} e^{-\frac{1}{2}\mathbf{z}^T \tilde{\Gamma}^{-1} \mathbf{z}} d\mathbf{z} \quad (60)$$

$$= -\frac{1}{2} \text{Tr} \left(\tilde{\Gamma}^{-1} \tilde{\Gamma}'_{\theta_i} \right) + \frac{1}{2} \text{Tr} \left(\tilde{\Gamma}^{-1} \tilde{\Gamma}'_{\theta_i} \tilde{\Gamma}^{-1} N \right), \quad (61)$$

where $\tilde{\Gamma}'_{\theta_i} = \frac{\partial}{\partial \theta_i} \tilde{\Gamma}$ and $N = \frac{1}{\Phi_m(\tilde{\gamma}; \tilde{\Gamma})} \iint_{-\infty}^{\tilde{\gamma}} \mathbf{z} \mathbf{z}^T \frac{1}{|2\pi\Gamma|} e^{-\frac{1}{2}\mathbf{z}^T \tilde{\Gamma}^{-1} \mathbf{z}} d\mathbf{z}$. In the derivations, we have exploited that $\frac{\partial}{\partial \theta_i} |\tilde{\Gamma}| = |\tilde{\Gamma}| \text{Tr}(\tilde{\Gamma}^{-1} \frac{\partial}{\partial \theta_i} \tilde{\Gamma})$ and that $\frac{\partial}{\partial \theta_i} \tilde{\Gamma}^{-1} = -\tilde{\Gamma}^{-1} \left(\frac{\partial}{\partial \theta_i} \tilde{\Gamma} \right) \tilde{\Gamma}^{-1}$.

Funding Open Access funding provided by the IReL Consortium.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Alodat, M., & Al-Momani, E. (2014). Skew Gaussian process for nonlinear regression. *Communications in Statistics-Theory and Methods*, 43(23), 4936–4961.
- Alodat, M., & Al-Rawwash, M. (2014). The extended skew Gaussian process for regression. *Metron*, 72(3), 317–330.
- Alodat, M., & Shakhatreh, M. K. (2020). Gaussian process regression with skewed errors. *Journal of Computational and Applied Mathematics*, 370, 112665.
- Arellano, R. B., & Azzalini, A. (2006). On the unification of families of skew-normal distributions. *Scandinavian Journal of Statistics*, 33(3), 561–574.
- Azzalini, A. (2013). *The skew-normal and related families* (Vol. 3). Cambridge University Press.
- Azzimonti, D., & Ginsbourger, D. (2018). Estimating orthant probabilities of high-dimensional gaussian vectors with an application to set estimation. *Journal of Computational and Graphical Statistics*, 27(2), 255–267.
- Bauer, M., van der Wilk, M., & Rasmussen, C. E. (2016). Understanding probabilistic sparse Gaussian process approximations. *Advances in neural information processing systems* (pp. 1533–1541).
- Bemporad, A., & Piga, D. (2021). Global optimization based on active preference learning with radial basis functions. *Machine Learning*, 110(2), 417–448.
- Benavoli, A., Azzimonti, D., & Piga, D. (2020). Skew Gaussian processes for classification. *Machine Learning*, 109, 1877–1902.
- Benavoli, A., Azzimonti, D., Piga, D., & (2021). Preferential Bayesian optimisation with skew Gaussian Processes. In *Genetic and evolutionary computation conference companion (GECCO'21 companion)*, July 10–14, 2021, Lille, France, New York, NY. ACM.

- Berkenkamp, F., Krause, A., & Schoellig, A. P. (2016a). Bayesian optimization with safety constraints: Safe and automatic parameter tuning in robotics. arXiv preprint [arXiv:1602.04450](https://arxiv.org/abs/1602.04450).
- Berkenkamp, F., Schoellig, A. P., & Krause, A. (2016b). Safe controller optimization for quadrotors with gaussian processes. In *2016 IEEE international conference on robotics and automation (ICRA)* (pp. 491–496). IEEE.
- Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
- Botev, Z. I. (2017). The normal law under linear restrictions: Simulation and estimation via minimax tilting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1), 125–148.
- Brochu, E., de Freitas, N., & Ghosh, A. (2008). Active preference learning with discrete choice data. *Advances in neural information processing systems* (pp. 409–416).
- Cao, J., Durante, D., & Genton, M. G. (2020). Scalable computation of predictive probabilities in probit models with gaussian process priors. [arXiv:2009.01471](https://arxiv.org/abs/2009.01471).
- Chu, W., & Ghahramani, Z. (2005a). Gaussian processes for ordinal regression. *Journal of Machine Learning Research*, 6, 1019–1041.
- Chu, W., & Ghahramani, Z. (2005b). *Preference learning with gaussian processes*. Association for Computing Machinery.
- Cunningham, J. P., Hennig, P., & Lacoste-Julien, S. (2013). Gaussian probabilities and expectation propagation.
- Durante, D. (2019). Conjugate Bayes for probit regression via unified skew-normal distributions. *Biometrika*, 106(4), 765–779.
- Gelman, A., Rubin, D. B., et al. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4), 457–472.
- Genton, M. G., Keyes, D. E., & Turkiiyah, G. (2018). Hierarchical decompositions for the computation of high-dimensional multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 27(2), 268–277.
- Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 1(2), 141–149.
- Genz, A., & Bretz, F. (2009). *Computation of multivariate normal and t probabilities* (Vol. 195). Springer Science & Business Media.
- Gessner, A., Kanjilal, O., & Hennig, P. (2020). Integrals over Gaussians under linear domain constraints. In S. Chiappa, & R. Calandri (Eds.), *Proceedings of the twenty third international conference on artificial intelligence and statistics* (Vol. 108, pp. 2764–2774). PMLR.
- Gibbs, M. N., & MacKay, D. J. (2000). Variational gaussian process classifiers. *IEEE Transactions on Neural Networks*, 11(6), 1458–1464.
- González, J., Dai, Z., Damianou, A., & Lawrence, N. D. (2017). Preferential Bayesian optimization. In *Proceedings of the 34th international conference on machine learning* (Vol. 70, pp. 1282–1291). JMLR.org.
- GPy (2012). GPy: A gaussian process framework in python. <http://github.com/SheffieldML/GPy>.
- Gupta, A. K., Aziz, M. A., & Ning, W. (2013). On some properties of the unified skew normal distribution. *Journal of Statistical Theory and Practice*, 7(3), 480–495.
- Hensman, J., Fusi, N., & Lawrence, N. D. (2013). Gaussian processes for big data. In *Proceedings of the twenty-ninth conference on uncertainty in artificial intelligence, UAI'13, Arlington, Virginia, USA* (pp. 282–290). AUAI Press.
- Hernández-Lobato, D., & Hernández-Lobato, J. M. (2016). Scalable gaussian process classification via expectation propagation. *Artificial Intelligence and Statistics* (pp. 168–176).
- Houlsby, N., Huszár, F., Ghahramani, Z., & Lengyel, M. (2011). Bayesian active learning for classification and preference learning. arXiv preprint [arXiv:1112.5745](https://arxiv.org/abs/1112.5745).
- Kuss, M., & Rasmussen, C. E. (2005). Assessing approximate inference for binary Gaussian process classification. *Journal of Machine Learning Research*, 6, 1679–1704.
- MacKay, D. J. (1996). In Models of neural networks III. *Bayesian methods for backpropagation networks* (pp. 211–254). Springer.
- Minka, T. P. (2001). A family of algorithms for approximate bayesian inference. In UAI (pp. 362–369). Morgan Kaufmann.
- Murray, I., Adams, R., & MacKay, D. (2010). Elliptical slice sampling. In *Proceedings of the 13th international conference on artificial intelligence and statistics, Chia, Italy* (pp. 541–548). PMLR.
- Nickisch, H., & Rasmussen, C. E. (2008). Approximations for binary gaussian process classification. *Journal of Machine Learning Research*, 9, 2035–2078.
- O'Hagan, A. (1978). Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society: Series B (Methodological)*, 40(1), 1–24.

- O'Hagan, A., & Leonard, T. (1976). Bayes estimation subject to uncertainty about parameter constraints. *Biometrika*, 63(1), 201–203.
- Opper, M., & Archambeau, C. (2009). The variational gaussian approximation revisited. *Neural Computation*, 21(3), 786–792.
- Phinikettos, I., & Gandy, A. (2011). Fast computation of high-dimensional multivariate normal probabilities. *Computational Statistics & Data Analysis*, 55(4), 1521–1529.
- Quiñonero-Candela, J., & Rasmussen, C. E. (2005). A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6, 1939–1959.
- Rasmussen, C. E., & Williams, C. K. (2006). *Gaussian processes for machine learning*. MIT Press.
- Sadigh, D., Dragan, A. D., Sastry, S., & Seshia, S. A. (2017). Active preference-based learning of reward functions. In *Robotics: Science and systems*.
- Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). Probabilistic programming in python using pymc3. *PeerJ Computer Science*, 2, e55.
- Schuerch, M., Azzimonti, D., Benavoli, A., & Zaffalon, M. (2020). Recursive estimation for sparse Gaussian process regression. *Automatica*, 120, 109–127.
- Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., & De Freitas, N. (2015). Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1), 148–175.
- Snelson, E., & Ghahramani, Z. (2006). Sparse Gaussian processes using pseudo-inputs. *Advances in neural information processing systems* (pp. 1257–1264).
- Sui, Y., Gotovos, A., Burdick, J., & Krause, A. (2015). Safe exploration for optimization with gaussian processes. In *International conference on machine learning* (pp. 997–1005). PMLR.
- Titsias, M. (2009). Variational learning of inducing variables in sparse Gaussian processes. In D. van Dyk, & M. Welling (Eds.), *Proceedings of the twelfth international conference on artificial intelligence and statistics, volume 5 of proceedings of machine learning research. Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA* (pp. 567–574). PMLR.
- Trinh, G., & Genz, A. (2015). Bivariate conditioning approximations for multivariate normal probabilities. *Statistics and Computing*, 25(5), 989–996.
- Villacampa-Calvo, C., Zaldivar, B., Garrido-Merchán, E. C., & Hernández-Lobato, D. (2020). Multi-class gaussian process classification with noisy inputs. arXiv preprint [arXiv:2001.10523](https://arxiv.org/abs/2001.10523).
- Williams, C. K., & Barber, D. (1998). Bayesian classification with gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12), 1342–1351.
- Zoghi, M., Whiteson, S., Munos, R., & Rijke, M. (2014). Relative upper confidence bound for the k-armed dueling bandit problem. In *Proceedings of the 31st international conference on machine learning, Beijing, China* (pp. 10–18).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.