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

Alessio Benavoli
School of Computer Science and Statistics, Trinity College, Dublin, Ireland
BENAVOLI@GMAIL.COM

Dario Azzimonti
Dalle Molle Institute for Artificial Intelligence (IDSIA)
Manno, Switzerland
DARIO.AZZIMONTI@IDSIA.CH

Dario Piga
Dalle Molle Institute for Artificial Intelligence (IDSIA)
Manno, Switzerland
DARIO.PIGA@SUPSI.CH

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.

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 exact inference is known to be intractable.

In this paper, we show that is actually possible to derive closed-form expression for the posterior process in all the above cases (not only for regression), and that the posterior process is a Skew Gaussian Process (SkewGP). SkewGPs are more general and more flexible nonparametric distributions than GPs, as 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 show that we can compute inferences for regression, classification, pref-
ference and mixed problems with computation complexity of $O(n^3)$ and storage demands of $O(n^2)$ (same as for 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.

| Prior | Likelihood | regression | preference|classification | mixed |
|-------|------------|------------|------------|---------------|-------|
| Skew Gaussian Process | Normal | SkewGP | GP | GP |
| Gaussian Process | Affine Probit | SkewGP | GP |
| Gaussian Process | Product | SkewGP |

Table 1: SkewGPs are conjugate to a larger class of likelihoods.

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 $x_i \in \mathbb{R}^d$ and output $y_i$. Depending on the type of the output variable, supervised learning problems can be divided into different categories.

**Numeric:** In the regression setting the outputs are real values $y \in \mathbb{R}$ and the input-output mapping is usually modelled as $y_i = f(x_i) + v_i$, where $v_i$ is an additive independent, identically distributed Gaussian noise with zero mean and variance $\sigma_v^2$. The likelihood model is given by:

$$p(y_i|f(x_i)) = \phi \left( \frac{y_i - f(x_i)}{\sigma_v} \right),$$

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:

$$p(y_i|f(x_i)) = \Phi(f(x_i)),$$

that is a Bernoulli distribution with probability $\Phi(f(x_i))$, where $\Phi(\cdot)$ is the CDF of the standard Normal distribution.

**Ordinal:** In ordinal regression, $y_i \in \{1, 2, \ldots, r\}$, where the integer $1, 2, \ldots, r$ are used to denote order categories (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 $\mathbb{R} = (-\infty, b_1] \cup (b_1, b_2] \cup \cdots \cup (b_{r-1}, \infty)$. The likelihood can be modelled as an indicator function $p(y_i|f(x_i)) = I_{[(b_{j-1}, b_j]}(f(x_i))$. In case this observation model is contaminated by noise, we assume a Gaussian noise with zero mean and
unknown variance $\sigma_y^2$, the likelihood model becomes (Chu and Ghahramani, 2005b):

$$p(y_i|f(x_i)) = \int I_{[b_{y_i-1},b_{y_i})}(f(x_i) + v) \phi \left( \frac{v}{\sigma_y} \right) dv = \Phi \left( \frac{b_{y_i} - f(x_i)}{\sigma_y} \right) - \Phi \left( \frac{b_{y_{i-1}} - f(x_i)}{\sigma_y} \right). \quad (3)$$

The $b_i$ defining the partition $(-\infty, b_1], (b_1, b_2], \ldots, (b_{r-1}, \infty)$ 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 $x_i$ with the label $i$, then we can define a preference relation over the inputs. The label preference $i \succ j$ means that the input $x_i$ is preferred to $x_j$. This can be modelled by assuming that there is an unobservable latent function $f_i$ associated with each input $x_i$, and that the function values $\{f(x_i)\}_k$ preserve the preference relations observed in the dataset. Then the likelihood can be modelled as an indicator function $p(x_i \succ x_j|f(x_i), f(x_j)) = I_{f(x_i) \succ f(x_j)}(f(x_i))$. 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 and Ghahramani, 2005a):

$$p(x_i \succ x_j|f(x_i), f(x_j)) = \int I_{f(x_i) + v_1 \succ f(x_j) + v_2}(f(x_i) + v_1) \phi \left( \frac{v_1}{\sigma_y} \right) \phi \left( \frac{v_2}{\sigma_y} \right) dv_1 dv_2 = \Phi \left( \frac{f(x_i) - f(x_j)}{\sqrt{2} \sigma_y} \right). \quad (4)$$

More generally, the preferences of each input can be presented in the form of a preference graph (Chu and Ghahramani, 2005a), where the labels are the graph vertices. Some examples are shown in (Chu and Ghahramani, 2005a, Fig.1). Therefore, in this more general case, $y_i = \{(c_i^+, c_i^-)\}^r_{j=1}$, 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 latent function $f_i$ for each predefined label (Chu and Ghahramani, 2005a). The observed edge $(c_i^+, c_i^-)$ is modelled as the following constraint $f_{c_i^+}(x_i) \geq f_{c_i^-}(x_i)$ and the likelihood is

$$p(\{(c_i^+, c_i^-)\}^r_{j=1}|f(x_i)) = \prod_{j=1}^{g_i} \Phi \left( \frac{f_{c_i^+}(x_i) - f_{c_i^-}(x_i)}{\sqrt{2} \sigma_y} \right), \quad (5)$$

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

**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 Figure 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 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 \(x \in [2.5, 5]\).

![Figure 1: Mixed numeric and preference observations](image)

**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 and 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 and Barber, 1998); (ii) Expectation Propagation (EP) (Minka, 2001); (iii) Kullback-Leibler divergence (KL) minimization (Opper and Archambeau, 2009), comprising Variational Bounding (VB) (Gibbs and 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 (2005b) 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, 2005a) 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., 2020a), we that, although the probit likelihood (2) and the GP are not conjugate, the posterior process can still be computed in closed form and it is a Skew Gaussian Process (SkewGP). Moreover, by extending a result derived by Durante (2019) for the parametric case, we proved that SkewGP 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., 2020b).
In the next sections, we extend this result 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, Section 6 concludes the paper.

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

Skew-normal distributions have long been seen (O’Hagan and 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 \( \phi \) and \( \Phi \) 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 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 \( z \in \mathbb{R}^p \) is distributed as a Unified Skew-Normal distribution with latent skewness dimension \( s \), \( z \sim \text{SUN}_{p,s}(\xi, \Omega, \Delta, \gamma, \Gamma) \), if its probability density function (Azzalini, 2013, Ch.7) is:

\[ p(z) = \phi_p(z - \xi; \Omega) \Phi_s \left( \gamma + \Delta^T \bar{\Omega}^{-1} D_{\Omega}^{-1} (z - \xi); \Gamma - \Delta^T \bar{\Omega}^{-1} \Delta \right) \Phi_s (\gamma; \Gamma), \quad (6) \]

where \( \phi_p(z - \xi; \Omega) \) 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 \( D_{\Omega} \) a diagonal matrix containing the square root of the diagonal elements in \( \Omega \). The notation \( \Phi_s (\mathbf{a}; M) \) denotes the CDF of \( \phi_s (0; M) \) evaluated at \( \mathbf{a} \in \mathbb{R}^s \). The distribution is parametrized by a location vector \( \xi \), a covariance matrix \( \Omega \) and the latent variable parameters \( \gamma \in \mathbb{R}^s, \Gamma \in \mathbb{R}^{s \times s}, \Delta^p \times \Delta^{s \times s} \). In particular \( \Delta \) is the skewness matrix.
Figure 2: Density plots for $SUN_{1,s}(0, 1, \Delta, \gamma, \Gamma)$. For all plots $\Gamma$ is a correlation matrix, $\gamma = 0$, dashed lines are the contour plots of $y \sim N_1(0, 1)$.

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

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

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

Note that when $\Delta = 0$, (6) reduces to $\phi_p(z - \xi; \Omega)$, 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 Figure 3. The contours of the corresponding bivariate 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.

### 2.2 Additive representations

The role of the latent dimension $s$ can be briefly explained as follows. Consider a random vector $[x_0 \ x_1] \sim \phi_{s+p}(0; M)$ with $M$ as in (7) and define $y$ as the vector with distribution $(x_1 | x_0 + \gamma > 0)$. The density of $y$ can be written as

$$f(y) = \frac{\int_{x_0+\gamma>0} \phi_{s+p}((t_0, y); M) dt_0}{\int_{x_0+\gamma>0} \phi_s(t; \Gamma) dt} = \frac{\phi_p(y; \hat{\Omega}) \Phi_{s}(\gamma > 0 | x_0 = y)}{\Phi_{s}(\gamma; \Gamma)}$$

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 $z = \xi + D_{\Omega}y \sim \text{SUN}_{p,s}(\xi, \Omega, \Delta, \gamma, \Gamma)$. Note that the skewness of $z$ is determined by the correlation $\Delta$ of $x_1$ with the latent $s$-dimensional vector $x_0$.

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 $r_0 \sim \phi_p(0; \hat{\Omega} - \Delta^T \Delta^{-1} \Delta)$ and $r_{1,-\gamma}$, the truncation below $\gamma$ of $r_1 \sim \phi_s(0, \Gamma)$. Then the random variable

$$z_u = \xi + D_{\Omega}(r_0 + \Delta \Omega^{-1} r_{1,-\gamma}), \quad (8)$$

is distributed as (6), (Azzalini, 2013, Ch.7) and (Benavoli et al., 2020b, Sec.A.1). The additive representation introduced above can be used to draw samples from the distribution as it will be discussed in Section 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 $z \sim \text{SUN}_{p,s}(\xi, \Omega, \Delta, \gamma, \Gamma)$ and partition $z = [z_1, z_2]^T$, where $z_1 \in \mathbb{R}^{p_1}$ and $z_2 \in \mathbb{R}^{p_2}$.
with \( p_1 + p_2 = p \), then
\[
z_1 \sim SUN_{p_1,s}(\xi_1, \Omega_{11}, \Delta_1, \gamma, \Gamma),
\]
with \( \xi = \begin{bmatrix} \xi_1 \\ \xi_2 \end{bmatrix} \), \( \Delta = \begin{bmatrix} \Delta_1 \\ \Delta_2 \end{bmatrix} \), \( \Omega = \begin{bmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{bmatrix} \).

Moreover, (Azzalini, 2013, Ch.7) the conditional distribution is a unified skew-Normal, i.e., \((Z_2 | Z_1 = z_1) \sim SUN_{p_2,s}(\xi_{2|1}, \Omega_{2|1}, \Delta_{2|1}, \gamma_{2|1}, \Gamma_{2|1})\), where
\[
\xi_{2|1} := \xi_2 + \Omega_{21} \Omega_{11}^{-1} (z_1 - \xi_1), \quad \Omega_{2|1} := \Omega_{22} - \Omega_{21} \Omega_{11}^{-1} \Omega_{12}, \\
\Delta_{2|1} := \Delta_2 - \Omega_{21} \Omega_{11}^{-1} \Delta_1, \\
\gamma_{2|1} := \gamma + \Delta_1^T \Omega_{11}^{-1} (z_1 - \xi_1), \quad \Gamma_{2|1} := \Gamma - \Delta_1^T \Omega_{11}^{-1} \Delta_1,
\]
and \( \Omega_{11}^{-1} := (\Omega_{11})^{-1} \).

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

### 2.4 SkewGP

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

Consider a location function \( \bar{\xi} : \mathbb{R}^d \rightarrow \mathbb{R} \), a scale (kernel) function \( \Omega : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R} \), 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 \( x_1, \ldots, x_n \in \mathbb{R}^d \), the vector \( [f(x_1), \ldots, f(x_n)] \in \mathbb{R}^n \) is skew-normal distributed with parameters \( \gamma, \Gamma \) and location, scale and skewness matrices, respectively, given by
\[
\bar{\xi}(X) := \begin{bmatrix} \bar{\xi}(x_1) \\ \bar{\xi}(x_2) \\ \vdots \\ \bar{\xi}(x_n) \end{bmatrix}, \quad \Omega(X,X) := \begin{bmatrix} \Omega(x_1,x_1) & \Omega(x_1,x_2) & \cdots & \Omega(x_1,x_n) \\ \Omega(x_2,x_1) & \Omega(x_2,x_2) & \cdots & \Omega(x_2,x_n) \\ \vdots & \vdots & \cdots & \vdots \\ \Omega(x_n,x_1) & \Omega(x_n,x_2) & \cdots & \Omega(x_n,x_n) \end{bmatrix}, \quad \Delta(X) := \begin{bmatrix} \Delta(x_1) \\ \Delta(x_2) \\ \vdots \\ \Delta(x_n) \end{bmatrix}.
\]

The skew-normal distribution is well defined if the matrix \( M = \begin{bmatrix} \Gamma & \Delta(X) \\ \Delta(X)^T & \Omega(X,X) \end{bmatrix} \) is positive definite for all \( X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d \) and for any \( n \). Benavoli et al. (2020a) shows that SkewGP is a well defined stochastic process. In that case we write \( f \sim \text{SkewGP}(\bar{\xi}, \Omega, \Delta, \gamma, \Gamma) \).

We briefly review here a possible choice for the functions \( \bar{\xi}, \Delta \) and the matrix \( \Gamma \) that guarantees that \( M \) is always positive definite. We follow Benavoli et al. (2020a) and we choose a positive definite kernel stationary function \( K : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R} \) which we will use to generate both \( \Omega \) and \( \Delta \). Given \( n \) points and \( X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d \), \( s \) pseudo-points \( U = \{u_1, \ldots, u_s\} \subset \mathbb{R}^d \) 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
\[
M = \begin{bmatrix} \Gamma & \Delta(U,U) \\ \Delta(U,X) & \Omega(X,X) \end{bmatrix} = \begin{bmatrix} LK(U,U)L & LK(U,X) \\ K(U,X)L & K(X,X) \end{bmatrix},
\]

1. This construction can easily be generalised to non-stationary kernels.
where $\tilde{K}(\mathbf{x}, \mathbf{x}') = \frac{1}{\sigma^2}K(\mathbf{x}, \mathbf{x}')$. 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 $\Gamma$ and $\Delta$. See Benavoli et al. (2020a) 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 $X = \{\mathbf{x}_i : i = 1, \ldots, n\}$, with $\mathbf{x}_i \in \mathbb{R}^d$, and $m_r$ output points $Y = \{y_i : i = 1, \ldots, m_r\}$, with $y_i \in \mathbb{R}$ and the likelihood

$$p(Y \mid f(X)) = \prod_{i=1}^{m_r} \phi \left( \frac{y_i - \mathbf{c}_i^T f(\mathbf{x}_i)}{\sigma_v} \right) = \phi_m(\mathbf{y} - C f(X); R),$$

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

**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 $p(Y \mid f(X)) = \phi_m(\mathbf{y} - C f(X); R)$ 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:

$$p(f(X) \mid Y) = \text{SUN}_{m_r,s}(\xi_p, \Omega_p, \Delta_p, \gamma_p, \Gamma_p) \quad \text{with}$$

$$\xi_p = \xi + \Omega C^T(C\Omega C^T + R)^{-1}(Y - C\xi),$$
$$\Omega_p = \Omega - \Omega C^T(C\Omega C^T + R)^{-1}C\Omega,$$
$$\Delta_p = \tilde{\Omega}_p D_\Delta \tilde{\Omega}^{-1} \Delta,$$
$$\gamma_p = \gamma + \Delta^T \tilde{\Omega}^{-1} D_\Omega^{-1}(\xi_p - \xi),$$
$$\Gamma_p = \Gamma - \Delta^T \tilde{\Omega}^{-1} \Delta + \Delta^T \tilde{\Omega}_p^{-1} \Delta_p,$$

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

All the proofs are in the Appendix. For either $s = 0$ or $\Delta = 0$, the SkewGP prior becomes a GP and it can noted that, in this case, the posterior is Gaussian with posterior mean $\xi_p$ and posterior covariance $\Omega_p$ (that is the terms in (15)–(17) 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 2 Consider the probabilistic model in Lemma 1, the marginal likelihood of the observations \( Y \) is

\[
p(Y) = \phi_{m_x}(Y - C\xi(X); C\Omega(X,X)C^T + R) \frac{\Phi_s(\gamma_p; \Gamma_p)}{\Phi_s(\gamma; \Gamma)}.
\]  

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

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

Theorem 3 Let us assume a SkewGP prior \( f(x) \sim \text{SkewGP}_s(\xi(x), \Omega(x,x'), \Delta(x), \gamma, \Gamma) \), the likelihood \( p(Y|f(X)) = \phi_{m_y}(Y - Cf(X); R) \), then a-posteriori \( f \) is SkewGP with mean, scale, and skew functions:

\[
\tilde{\xi}(x) = \xi(x) + \Omega(x,x)C^T(C\Omega(X,X)C^T + R)^{-1}(Y - C\xi(X)),
\]

(19)

\[
\tilde{\Omega}(x,x) = \Omega(x,x) - \Omega(x,x)C^T(C\Omega(X,X)C^T + R)^{-1}C\Omega(X,x),
\]

(20)

\[
\tilde{\Delta}(x) = D^{-1}_{\tilde{\Omega}(x,x)}D_{\tilde{\Omega}(x,x)}\Delta(x)
\]

\[
- D^{-1}_{\tilde{\Omega}(x,x)}\Omega(x,x)C^T(C\Omega(X,X)C^T + R)^{-1}CD_{\tilde{\Omega}(x,x)}\Delta(X),
\]

(21)

and parameters \( \gamma_p, \Gamma_p \) 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 \( X = \{x_i : i = 1, \ldots, n\} \), with \( x_i \in \mathbb{R}^d \), and a data-dependent matrix \( W \in \mathbb{R}^{m_x \times n} \) and vector \( Z \in \mathbb{R}^{m_x} \). We define an affine probit likelihood as

\[
p(W,Z|f(X)) = \Phi_{m_x}(Z + Wf(X); \Sigma),
\]

(22)

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

\[
\Phi_{m_x} \left( \begin{bmatrix} 2y_1 - 1 \\ \vdots \\ 2y_n - 1 \end{bmatrix} ; I_{m_x} \right).
\]

(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

\[
\Phi_{m_x} \left( \begin{bmatrix} \frac{f(y_1) - f(u_1)}{\sigma_y} \\ \vdots \\ \frac{f(y_n) - f(u_m)}{\sigma_y} \end{bmatrix} ; I_{m_x} \right),
\]

(24)
where \( v_i, u_j \in X \). If we denote by \( W \in \mathbb{R}^{m_u \times n} \) the matrix defined as \( W_{ij} = V_{i,j} - U_{i,j} \) where \( V_{i,j} = \frac{1}{\sigma_i} \) if \( v_i = x_j \) and 0 otherwise and \( U_{i,j} = \frac{1}{\sigma_i} \) if \( u_i = x_j \) and 0 otherwise. Then we can write the likelihood (24) as in (22). The conjugacy between: (i) SkewGP and the likelihood (23), (ii) SkewGP and the likelihood (24), were derived in (Benavoli et al., 2020a) and, respectively, (Benavoli et al., 2020b). Here we extend those results by including the shifting term \( Z \) (which, for instance, allows us to deal with the likelihood (3)) and a general covariance matrix \( \Sigma \), as in (22).

**Lemma 4** Let us assume that \( f(x) \sim \text{SkewGP}_s(\xi(x), \Omega(x,x'), \Delta(x), \gamma, \Gamma) \) and consider the likelihood \( \Phi_{m_u}(Z + W f(x); \Sigma) \) where \( \Sigma \in \mathbb{R}^{m_u \times m_u} \) is symmetric positive definite, \( W \in \mathbb{R}^{m_u \times n} \) and \( Z \in \mathbb{R}^{m_u} \) are an arbitrary data dependent matrix and, respectively, vector. The posterior of \( f(x) \) is a SUN:

\[
p(f(X)|W, Z) = \text{SUN}_{n,s+m_u}(\xi_p, \Omega_p, \Delta_p, \gamma_p, \Gamma_p) \tag{25}
\]

\[
\xi_p = \xi, \quad \Omega_p = \Omega, \tag{26}
\]

\[
\Delta_p = [\Delta, \bar{\Omega}D\Omega W^T], \tag{27}
\]

\[
\gamma_p = [\gamma, Z + W\bar{\xi}]^T, \tag{28}
\]

\[
\Gamma_p = \begin{bmatrix}
\Gamma & D\bar{\Omega}W^T \\
W\bar{\Omega}\Delta & (W\Omega W^T + \Sigma)
\end{bmatrix}, \tag{29}
\]

where, for simplicity of notation, we denoted \( \xi(X), \Omega(X,X), \Delta(X) \) as \( \xi, \Omega, \Delta \) respectively and \( \Omega = D\bar{\Omega}\bar{\Omega}D\bar{\Omega} \).

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

**Corollary 5** Consider the probabilistic model in Lemma 4, the marginal likelihood of the observations \( Z, W \) is

\[
p(Z, W) = \frac{\Phi_{s+m_u}(\gamma_p; \Gamma_p)}{\Phi_s(\gamma; \Gamma)}. \tag{30}
\]

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

**Theorem 6** Let us assume a SkewGP prior \( f(x) \sim \text{SkewGP}_s(\xi(x), \Omega(x,x'), \Delta(x), \gamma, \Gamma) \), the likelihood \( \Phi_{m_u}(Z + W f(X)) \) with \( W \in \mathbb{R}^{m_u \times n} \) and \( Z \in \mathbb{R}^{m_u} \), then a-posteriori \( f \) is SkewGP with mean, covariance and skewness functions:

\[
\bar{\xi}(x) = \xi(x) \tag{31}
\]

\[
\bar{\Omega}(x,x) = \Omega(x,x),
\]

\[
\bar{\Delta}(x) = \begin{bmatrix} \Delta(x) & D^{-1}_{\Omega(x,x)}\Omega(x,X)W^T \end{bmatrix},
\]

and parameters \( \gamma_p, \Gamma_p \) as in Lemma 4.

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.

---

2. We omitted the constant \( \sqrt{2} \) for simplicity.
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 \( g(x) = \cos(5x) + e^{-x^2} \) which is in Figure 4.

![Figure 4: 1D function \( g(x) = \cos(5x) + e^{-x^2} \)](image)

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 will consider a GP prior (with zero mean and RBF kernel) over the unknown function \( f \) and we will compare the exact posterior computed via SkewGP with two approximations: Laplace (LP) and Expectation Propagation (EP). We will discuss how to compute the predictions for SkewGP in Section 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 Figure 5(bottom-left), which reports the skewness statistics for \( h(x) = f(x) - f(x_r) \) as a function of \( x \), defined as:

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

with \( \mu := E[h(x)] \), and \( E \) 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 Figure 5(bottom-left). 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 \( X = \{x_i : i = 1, \ldots, n\} \), with \( x_i \in \mathbb{R}^d \), and the product likelihood:

\[
p(Y, Z, W \mid f(X)) = \Phi_{n_y}(Z + W f(X); \Sigma) \phi_{n_y}(Y - C f(X); R).
\]
Let us assume a SkewGP prior $f(x) \sim \text{SkewGP}_s(\tilde{\xi}(x), \Omega(x,x'), \Delta(x), \gamma, \Gamma)$, the likelihood (32), then a-posteriori $f$ is SkewGP with mean, covariance and skewness functions:

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

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

$$\tilde{\Delta}(x) = D_{\Omega(x,x)}^{-1}D_{\Omega(x,x)}\begin{bmatrix} \Delta(x) & D_{\Omega(x,x)}^{-1}\Omega(x,x)W^T \\
-D_{\Omega(x,x)}^{-1}C\Omega(x,x)C^T(C\Omega(x,x)C^T + R)^{-1}CD_{\Omega(x,x)} & \tilde{D}_{\Omega(x,x)}^{-1}\Omega(x,x)W^T \end{bmatrix},$$

$$\gamma = \gamma_p + \begin{bmatrix} \Delta(X) & D_{\Omega(X,X)}^{-1}\Omega(X,X)W^T \end{bmatrix}^T\tilde{\Omega}(X,X)^{-1}D_{\Omega(X,X)}^{-1}\tilde{\xi}(X) - \xi(X),$$

$$\Gamma = \Gamma_p - \begin{bmatrix} \Delta(X) & D_{\Omega(X,X)}^{-1}\Omega(X,X)W^T \end{bmatrix}^T\tilde{\Omega}(X,X)^{-1}(X,X)\begin{bmatrix} \Delta(X) & D_{\Omega(X,X)}^{-1}\Omega(X,X)W^T \end{bmatrix} + \Delta_p^T\tilde{\Omega}(X,X)^{-1}\Delta_p,$$  

with

$$\Delta_p = \tilde{\Omega}(X,X)D_{\tilde{\Omega}(X,X)}D_{\Omega(X,X)}^{-1}(X,X)\begin{bmatrix} \Delta(X) & D_{\Omega(X,X)}^{-1}\Omega(X,X)W^T \end{bmatrix}.$$  

and $\gamma_p, \Gamma_p$ as in Lemma 4.
Corollary 8 The marginal likelihood of the observations $Y$ is

$$p(Y) = \phi_m(Y - C\xi(X);C\Omega(X,X)C^T + R) \frac{\Phi_{x+m}(\tilde{\gamma}; \tilde{\Gamma})}{\Phi_{\gamma}(\gamma; \Gamma)}.$$ \hspace{1cm} (39)

with $\tilde{\gamma}, \tilde{\Gamma}$ in Theorem 7.

This general setting shows that SkewGPs are conjugate with a larger class of likelihoods and, therefore, encompass GPs.

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

Example 1D mixed numeric and preference regression Consider again the regression problem discussed in Section 1 and the dataset shown in Figure 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 $\sigma^2_v = 0.001$).

Figure 6(bottom-left) reports the skewness statistics for SkewGP and Figure 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 mean and the support of the true posterior (SkewGP), as also evident from Figure 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); Nickisch and Rasmussen (2008) for classification. 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 Figure 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 $\{(\text{valid}, y), (\text{non-valid}, \text{None})\}$, with $y = f(x) + v$. The above setting can be modelled by the likelihood:

$$
\begin{cases}
\phi \left( \frac{y-f(x)}{\sigma_v} \right) \Phi \left( \frac{f(x)-h}{\sigma_v} \right), & (\text{valid}, y) \\
\Phi \left( \frac{h-f(x)}{\sigma_v} \right), & (\text{non-valid}, \text{None}),
\end{cases}
$$ \hspace{1cm} (40)

and, therefore, formulated as a mixed regression problem.\footnote{In (Benavoli et al., 2020b), we considered mixed preferential-binary observations, where binary judgments (valid or non-valid) together with preference judgments are available.}

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 RBF covariance function (the hyperparameters were estimated using EP: lengthscale 0.497, variance $\sigma^2 = 0.147$ and noise variance $\sigma_v^2 = 0.0021$). Figure 7(bottom-left) reports the skewness statistics for SkewGP and Figure 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. Computing the marginal likelihood and sampling from the posterior

In the GP literature (Rasmussen and Williams, 2006), a GP is usually parametrised with a zero mean function $\xi(x) = 0$ and a covariance kernel $\Omega(x, x')$ which depends on hyperparameters $\theta \in \Theta$. Typically $\theta$ contains lengthscale parameters and a variance parameter. For instance, the Radial Basis Function (RBF) kernel

$$\Omega(x, x') := \sigma^2 \exp\left(-\frac{\|x - x'\|^2}{2\ell^2}\right),$$

has $\theta = [\ell, \sigma, \sigma_v]$, where $\sigma_v^2$ is the variance of the measurement noise. To fully define a SkewGP, we must also select the latent dimensions $s$, the additional parameters $\gamma, \Gamma$ and the skewness function $\Delta(x)$. The function $\Delta(x)$ may also have hyperparameters, which we denote by $u$, so that $\theta = [\ell, \sigma, \sigma_v, \gamma, \Gamma, u]$. A choice of $\gamma, \Gamma, u$ is discussed in (Benavoli et al., 2020a) for classification.

4. In the experiments, we use a RBF kernel with ARD, then there is a lengthscale parameter for each component of $x$ and $\ell$ is a vector.
Figure 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.

The parameters $\theta$ are chosen by maximizing the marginal likelihood, that for the general mixed problem (39), involves computing the CDF $\Phi_{s+m_a}$ of a multivariate normal distribution, whose dimensions grows with $m_a$. Quasi-randomized Monte Carlo methods (Genz, 1992; Genz and Bretz, 2009; Botev, 2017) have been proposed to calculate $\Phi_{s+m_a}$ 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 and Gandy (2011); Genton et al. (2018); Azzimonti and Ginsbourger (2018)). We overcome this issue by using the approximation introduced in (Benavoli et al., 2020a, Prop.2):

**Proposition 9** Lower bound of the CDF:

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

(41)

where $B_1, \ldots, B_b$ 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 (41), we use the routine proposed in Trinh and Genz (2015), that computes multivariate normal probabilities using bivariate conditioning. This is very fast and returns a deterministic quantity. We therefore optimise the hyperparameters of the kernel by maximising the lower bound in (41) using the Dual Annealing optimization routine.\(^5\)

\(^5\) For local search, we use the L-BFGS algorithm with numerical computation of the gradient.
4.1 Predictive inference

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 Section 2.2. Recall that \( z \sim \text{SUN}_p(\xi, \Omega, d, \gamma, \Gamma) \) can be written as \( z = \xi + D_\Omega(r_0 + \Delta \Gamma^{-1} r_{1,-\gamma}) \) with \( r_0 \sim \phi_p(0; \Omega - \Delta \Gamma^{-1} \Gamma^T) \) and \( r_{1,-\gamma} \) is the truncation below \( \gamma \) of \( r_1 \sim \phi_r(0; \Gamma) \). Note that sampling \( r_0 \) can be achieved efficiently with standard methods, however using standard rejection sampling for the variable \( r_{1,-\gamma} \) would incur in exponentially growing sampling time as the dimension \( m_o \) increases. In (Benavoli et al., 2020a,b) 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

\[
\text{Algorithm 1: Lin-Ess to sample from } \phi_r(\mu; \Gamma) \text{ truncated below } \gamma
\]

input: \( \Gamma \in \mathbb{R}^{d \times d} \) symmetric p.d., \( \mu, \gamma \in \mathbb{R}^d, m > 0 \)

output: \( x_{\text{sample}} \in \mathbb{R}^{m \times d} \), \( m \) samples from the \( \phi_r(0; \Gamma) \) truncated below \( \gamma \).

\( L = \text{Cholesky}(\Gamma) \);

select \( x_0 \) s.t. \( Lx_0 - \gamma > 0 \);

while \( i < m \) do

sample \( v \sim \phi_r(0; \Gamma) \);

compute \( r^2 = x_0^2 + v^2 \);

compute the angles \( \theta_1 = 2 \arctan(\frac{v - \sqrt{r^2 - \gamma}}{x_0 + \gamma}) \) and \( \theta_2 = 2 \arctan(\frac{v + \sqrt{r^2 - \gamma}}{x_0 + \gamma}) \);

find the largest intersection angles in the domain boundary \([\theta_{\text{min}}, \theta_{\text{max}}]\);

sample \( \theta \sim U(\theta_{\text{min}}, \theta_{\text{max}}) \);

compute \( x_{i,\text{sample}} = x_0 \cos(\theta) + v \sin(\theta) \);

set \( x_0 = x_{i,\text{sample}} \);

\( i = i + 1 \);

Two alternative approaches to approximate the posterior mean have been proposed in (Cao et al., 2020): (i) using a tile-low-rank Monte 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

6. The pseudo-code reports the main steps of the algorithm. We omitted some implementation details for simplicity, for instance how to select \( x_0 \) or how to deal with empty intersections. The full code of the algorithm is available at https://github.com/benavoli/SkewGP.

7. Its computational bottleneck is the Cholesky factorization of the covariance matrix \( \Gamma \), same as for sampling from a multivariate Gaussian.
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.

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 in the computation of the posterior.

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 unlabelled 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):\(^8\)

$$\text{Bald}(x) = h(E_{f \sim p(f|W)}(\Phi(f(x)))) - E_{f \sim p(f|W)}(h(\Phi(f(x))))$$  \hspace{1cm} (42)

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

\(^8\) It corresponds to the information gain computed in $y$ space.
As Bayesian model we consider a GP classifier (GP prior and probit likelihood) and we compute the expectations in (42) 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. Table 2 displays the characteristics of the considered 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        |

Table 2: Characteristics of the datasets.

They are multiclass dataset which we transformed into binary classification problems considering the first class (class 0) versus rest (class 1). 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 8 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 Figure 9 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 $x^*$:

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

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 $x^*$ (exploitation).

9. These datasets have recently been used for GP classification in (Villacampa-Calvo et al., 2020).
Figure 8: 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.
In BO, it is usually assumed that $g(x)$ can be evaluated directly (numeric (noisy) observations). However, in many applications, measuring $g(x)$ 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 $x_i, x_j$ (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 and Piga, 2019). 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., 2020b), 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 transformed kernel function. Therefore, we compare three different implementation of Bayesian preferential optimisation based on LP, EP, SkewGP.\(^\text{10}\)

In PBO, since $g$ can only be queried via preferences, the next candidate solution $x$ is selected by optimizing (w.r.t. $x$) a dueling acquisition function $\alpha(x, x_r)$, where $x_r$ (reference point) is the best point found so far.\(^\text{11}\). By optimizing $\alpha(x, x_r)$, one aims to find a point that is better than $x_r$ (but

\(^{10}\) For LP and EP, we use (GPy, since 2012).

\(^{11}\) More precisely, we assume that preferential observation compares the current input with the best input found so far $x_r$. 

---

Figure 9: Averaged pool diversity (percentage) in the active learning task.
also considering the trade-off between exploration and exploitation). After computing the optimum of the acquisition function, denoted with \( \mathbf{x}_n \), we query the black-box function for \( \mathbf{x}_n \sim \mathbf{x}_r \). If \( \mathbf{x}_n \succ \mathbf{x}_r \), 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., 2020b) by combining the expected improvement 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 \( k = 0.1 \)).

Figure 10: 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 \( \mathbf{x}_r \).

We have considered for \( g(\mathbf{x}) \) six benchmark functions: \( g(\mathbf{x}) = \cos(5\mathbf{x}) + e^{-\mathbf{x}^2} \) denoted as ‘1D’ (1D), ‘Six-Hump Camel’ (2D), ‘Langer’ (2D), ‘Colville’ (4D), ‘Rosenbrock5’ (5D) and ‘Hartman6’ (6D). These are minimization problems.\(^{12} \) 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.

\(^{12} \) We converted them into maximizations so that the acquisition functions are well-defined.
with different initialization (the same for all methods) as in (González et al., 2017; Benavoli et al., 2020b).

Figure 10 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.

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 $4, 8, 12, 16, \ldots$ $g$ 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.

Figure 11 shows the performance of the different methods. As expected, 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:

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

where $g$ is an unknown function and $h \in \mathbb{R}$ is a safety threshold. The set $\mathcal{S} = \{x \in X \mid g(x) \geq h\}$ 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 $\mathcal{S}$ (exploitation), but also aims to increase the size of $\mathcal{S}$ including inputs $x$ that are known to be safe (exploration). Different strategies and approaches are discussed in (Sui et al., 2015).

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

We can address this BO problem using the general framework (Berkenkamp et al., 2016b,a), 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.

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 a term that penalises

13. Observe in fact that the Bald criterion assumes binary observations.
14. LP, EP and SkewGP coincides in case the observations are all numeric.
15. We have used the SafeOpt library (Berkenkamp et al., 2016b,a) with $\beta = 0.85$ and used the EP approximation for GP classification.
Figure 11: 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 $x_r$. 

24
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., 2016b,a), 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.\textsuperscript{16}

Figure 12 and 13 report 7 iterations of the two approaches, which we call as SafeOpt and SkewGP-BO, for one of the 100 trials. Figure 12(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 12(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 $x$s 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 Figure 12(Iter. 1, right), while SafeOpt selects a valid one, finding a better maximum Figure 12(Iter. 1, left).

At iteration 2, SkewGP-BO selects a valid point and finds a better maximum Figure 12(Iter. 2, right), while GP-SafeOpt violates the constraint Figure 12(Iter. 2, left).

At iteration 3, SkewGP-BO selects a non-valid point Figure 12(Iter. 3, right), while SafeOpt selects a valid point 12(Iter. 3, left).

At iteration 4 and 5, SkewGP-BO finds the maximum Figure 13(Iter. 4 and 5, right), while SafeOpt explores the valid region Figure 13(Iter. 4 and 5, left).

At iteration 6, SafeOpt violates the constraint Figure 13(Iter. 6, left).

At iteration 7, SafeOpt finally finds the maximum Figure 13(Iter. 7, left).

Figure 14 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 $x_r$. 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,\textsuperscript{17} SkewGP provides faster convergence than using two separated surrogated models (GPs) for objective and constraint.

\textsuperscript{16} The lengthscale can be different from the true one which is uniformly sampled in $[1, 2]$.

\textsuperscript{17} The framework for safe BO in (Berkenkamp et al., 2016b,a) is more general and cannot always be expressed as a mixed problem.
Figure 12: Iterations 0-3, GP-SafeOpt (left) and SkewGP-SafeOpt (right).
Figure 13: Iterations 4-7, GP-SafeOpt (left) and SkewGP-SafeOpt (right).
Figure 14: 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 $x_r$.

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, 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 $\Delta$ and select the prior parameters $\gamma, \Gamma$, which allows one to fully exploit the flexibility of SkewGPs.
Appendix A. Proofs of the results in the paper

Proof Lemma 1 and Corollary 2 The likelihood is $\phi_{m_n}(Y - Cf; R)$, where $Y \in \mathbb{R}^{n_y}$ is the vector of observations; $f \in \mathbb{R}^n$ is the vector of function values at the input points $x_i$ for $i = 1, \ldots, n$ and the prior is:

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

(43)

First, note that

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

with

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

(44)

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

(45)

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

$$\gamma + \Delta^T \tilde{\Omega}^{-1}D_\Omega^{-1}(f - \xi) = \gamma + \Delta^T \tilde{\Omega}^{-1}D_\Omega^{-1}(f - \xi_p + \xi_p - \xi)$$

$$= \gamma + \Delta^T \tilde{\Omega}^{-1}D_\Omega^{-1}(\xi_p - \xi) + \Delta^T \tilde{\Omega}^{-1}D_\Omega^{-1}(f - \xi_p)$$

$$= \gamma + \Delta^T \tilde{\Omega}^{-1}D_\Omega^{-1}(\xi_p - \xi) + \Delta^T \tilde{\Omega}^{-1}D_\Omega^{-1}D_{\Omega_p}^{-1}D_\Omega^{-1}D_{\Omega_p}^{-1}(f - \xi_p)$$

$$= \gamma_p + \Delta^T \tilde{\Omega}_p^{-1}D_{\Omega_p}^{-1}(f - \xi_p)$$

with

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

(46)

$$\Delta^T_p = \Delta^T \tilde{\Omega}_p^{-1}D_{\Omega_p}^{-1}D_{\Omega_p}^{-1}D_{\Omega_p}^{-1}(f - \xi_p)$$

(47)

Finally observe that

$$\Gamma - \Delta^T \tilde{\Omega}^{-1}\Delta = \Gamma - \Delta^T \tilde{\Omega}^{-1}\Delta + \Delta^T_p \tilde{\Omega}_p^{-1}D_{\Omega_p}^{-1} - \Delta^T_p \tilde{\Omega}_p^{-1}D_{\Omega_p}^{-1} = \Gamma_p - \Delta^T_p \tilde{\Omega}_p^{-1}D_{\Omega_p}^{-1}$$

with $\Gamma_p = \Gamma - \Delta^T \tilde{\Omega}_p^{-1}D_{\Omega_p}^{-1}D_{\Omega_p}^{-1}D_{\Omega_p}^{-1}$. Putting everything together

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

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 3** Consider the test point \( \mathbf{x} \in \mathbb{R}^d \) and the vector \( \hat{f} = \left[ f(X) \right] := [f \ f_s] \) we have

\[
p(\mathbf{f}, f_s) = \text{SkewGP} \left( \left[ \frac{\xi(X)}{\xi(\mathbf{x})} \right], \left[ \begin{array}{c} \Omega(X, X) \\ \Omega(\mathbf{x}, X) \end{array} \right], \left[ \begin{array}{c} \Delta(X) \\ \Delta(\mathbf{x}) \end{array} \right], \gamma, \Gamma \right)
\]

and the predictive distribution is by definition

\[
p(f_s \mid Y) = \int p(f_s \mid \mathbf{f})p(\mathbf{f} \mid Y)d\mathbf{f} = \int p(f_s \mid \mathbf{f})p(\mathbf{f} \mid Y)p(\mathbf{f})d\mathbf{f} = \int p(f_s, \mathbf{f})p(\mathbf{f} \mid Y)d\mathbf{f}
\]

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

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

with

\[
\hat{\xi} = \left[ \frac{\xi(X)}{\xi(\mathbf{x})} \right] + \left[ \begin{array}{c} \Omega(X, X) \\ \Omega(\mathbf{x}, X) \end{array} \right] C^T (C\Omega C^T + R)^{-1} (Y - C\xi(X)) \\
\hat{\Omega} = \left[ \begin{array}{c} \Omega(X, X) \\ \Omega(\mathbf{x}, X) \end{array} \right] - \left[ \begin{array}{c} \Omega(X, X) \\ \Omega(\mathbf{x}, X) \end{array} \right] C^T (C\Omega C^T + R)^{-1} C \left[ \begin{array}{c} \Omega(X, X) \\ \Omega(\mathbf{x}, X) \end{array} \right]
\]

(48) (49)

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

\[
\hat{\Delta} = D^{-1}_\Omega \hat{\Omega} \left[ \begin{array}{c} \Omega(X, X) \\ \Omega(\mathbf{x}, X) \end{array} \right] A_1 A_2 \left[ \begin{array}{c} \Delta(X) \\ \Delta(\mathbf{x}) \end{array} \right] = D^{-1}_\Omega \left[ \begin{array}{c} A_1 \\ A_3 \\ A_4 \end{array} \right] D\Omega \left[ \begin{array}{c} \Delta(X) \\ \Delta(\mathbf{x}) \end{array} \right]
\]

\[
\left[ \begin{array}{c} A_1 \\ A_3 \\ A_4 \end{array} \right] = \hat{\Omega} \left[ \begin{array}{c} \Omega(X, X) \\ \Omega(\mathbf{x}, X) \\ \Omega(\mathbf{x}, X) \end{array} \right]^{-1} \left[ \begin{array}{c} \Omega(X, X) \\ \Omega(\mathbf{x}, X) \end{array} \right]
\]

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

\[
A_3 = \left( \Omega(X, X) - \Omega(\mathbf{x}, X) C^T (C\Omega(X, X) C^T + R)^{-1} C\Omega(X, X) \right)
\]

\[
\left( \Omega(X, X)^{-1} + \Omega(X, X)^{-1} \Omega(\mathbf{x}, X) \left( \Omega(\mathbf{x}, X) - \Omega(\mathbf{x}, X) \Omega(X, X)^{-1} \Omega(X, X) \right) \right)^{-1} \Omega(\mathbf{x}, X) \Omega(X, X)^{-1} \Omega(X, X) \Omega(X, X)^{-1} \Omega(X, X)
\]

\[
= -\Omega(X, X) C^T (C\Omega(X, X) C^T + R)^{-1} C\Omega(X, X)
\]

\[
(\Omega(X, x) - \Omega(x, X) \Omega(X, X)^{-1} \Omega(X, x))^{-1} \Omega(X, x) \Omega(X, x)^{-1}
\]

\[
= -\Omega(X, X) C^T (C\Omega(X, X) C^T + R)^{-1} C
\]

30
Finally, we will show that:

\[ A_4 = -\left( \Omega(x,X) - \Omega(x,X) C^T (C \Omega(x,X) C^T + R)^{-1} C \Omega(x,X) \right) \]
\[ \Omega(X,X)^{-1} \Omega(X,x) (\Omega(x,x) - \Omega(x,X) \Omega(X,X)^{-1} \Omega(x,X))^{-1} \]
\[ + \left( \Omega(x,x) - \Omega(x,X) C^T (C \Omega(x,X) C^T + R)^{-1} C \Omega(x,X) \right) \]
\[ (\Omega(x,x) - \Omega(x,X) \Omega(X,X)^{-1} \Omega(x,X))^{-1} = I_x \]

and

By Lemma 1

\[ \hat{\gamma} = \gamma + [\Delta(x), \Delta(x)] D_\Omega \Omega^{-1} (\xi_p - \bar{\xi}) \]
\[ = \gamma + [\Delta(x), \Delta(x)] D_\Omega \left[ \begin{array}{c}
\Omega(x,X) \\
\Omega(x,x)
\end{array} \right]^{-1} \left[ \begin{array}{c}
\Omega(x,X) \\
\Omega(x,x)
\end{array} \right] C^T (C \Omega C^T + R)^{-1} (Y - C \xi(X)) \]

Note that

\[ \left[ \begin{array}{cc}
\Omega(x,X) & \Omega(x,x) \\
\Omega(x,x) & \Omega(x,x)
\end{array} \right]^{-1} \left[ \begin{array}{c}
\Omega(x,X) \\
\Omega(x,x)
\end{array} \right] = \left[ \begin{array}{c}
\Omega(X,X)^{-1} \\
0
\end{array} \right] \]

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 Section 2.3 and in particular (9), we can derive that

\[ p(f(x) \mid C, Y, f(X)) = \text{SUN}_{1,s} \left( \bar{\xi}(x), \bar{\Omega}(x,x), \bar{\Delta}(x), \gamma_p, \Gamma_p \right). \]  

with mean, scale, and skewness functions:

\[ \bar{\xi}(x) = \xi(x) + \Omega(x,X) C^T (C \Omega(x,X) C^T + R)^{-1} (Y - C \xi(X)), \]  

\[ \bar{\Omega}(x,x) = \Omega(x,x) - \Omega(x,X) C^T (C \Omega(x,X) C^T + R)^{-1} C \Omega(x,x), \]  

\[ \bar{\Delta}(x) = D_{\bar{\Omega}(x,x)}^{-1} D_{\bar{\Omega}(x,x)} \Delta(x) - D_{\bar{\Omega}(x,x)}^{-1} \Omega(x,X) C^T (C \Omega(x,X) C^T + R)^{-1} C \Omega(x,X) \Delta(x). \]

**Lemma 4** We aim to derive the posterior of \( f(X) \). The joint distribution of \( f(X), W, Z \) is

\[ p(W,Z \mid f(X))p(f(X)) \]
\[ = \Phi_{m_u}(Z + Wf; \Sigma) \text{SUN}_{m,s}(\xi, \Omega, \Delta, \gamma, \Gamma) \]
\[ = \Phi_{m_u}(Z + Wf; \Sigma) \phi_f(f - \xi; \Omega) \frac{\Phi_s(\gamma + \Delta^T \bar{\Omega}^{-1} D_{\bar{\Omega}}^{-1} (f - \xi); \Gamma - \Delta^T \bar{\Omega}^{-1} \Delta)}{\Phi_s(\gamma; \Gamma)} \]
\[ \propto \phi_f(f - \xi; \Omega) \Phi_{m_u}(Z + Wf; \Sigma) \Phi_s(\gamma + \Delta^T \bar{\Omega}^{-1} D_{\bar{\Omega}}^{-1} (f - \xi); \Gamma - \Delta^T \bar{\Omega}^{-1} \Delta)^s, \]  

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

\[ \Phi_{m_u}(Z + Wf; \Sigma) = \Phi_{m_u}(Z + Wf; \Sigma) + (\bar{\Omega} D_\bar{\Omega} W^T)^s \bar{\Omega}^{-1} D_{\bar{\Omega}}^{-1} (f - \xi); (W \Omega W^T + \Sigma) - WD_\Omega \bar{\Omega} D_\Omega W^T \).
Therefore, we can write
\[
\Phi_{m_a}(Z + W|f) \Phi_f \left( \gamma + \Delta^T \bar{\Omega}^{-1} D_{\Omega}^{-1} (f - \xi); \Gamma - \Delta^T \bar{\Omega}^{-1} \Delta \right)
\]
\[= \Phi_{m_a} (Z + W \xi + (\bar{\Omega} D_{\Omega} W^T)^T \bar{\Omega}^{-1} D_{\Omega}^{-1} (f - \xi); (W \Omega W^T + \Sigma) - WD_{\Omega} \bar{\Omega} D_{\Omega} W^T) \cdot \Phi_f (\gamma + \Delta^T \bar{\Omega}^{-1} D_{\Omega}^{-1} (f - \xi); \Gamma - \Delta^T \bar{\Omega}^{-1} \Delta) \]
\[= \Phi_{s+m_a} (m; M) \tag{55}
\]
with
\[
m = \begin{bmatrix} \gamma + \Delta^T \bar{\Omega}^{-1} D_{\Omega}^{-1} (f - \xi) \\ Z + W \xi + (\bar{\Omega} D_{\Omega} W^T)^T \bar{\Omega}^{-1} D_{\Omega}^{-1} (f - \xi) \end{bmatrix},
\]
and
\[
M = \begin{bmatrix} \Gamma - \Delta^T \bar{\Omega}^{-1} \Delta & 0 \\ 0 & (W \Omega W^T + \Sigma) - WD_{\Omega} \bar{\Omega} D_{\Omega} W^T \end{bmatrix}.
\]

From (54)–(55) and the definition of the PDF of the SUN distribution (6), we can easily show that we can rewrite (54) as a SUN distribution with updated parameters:
\[
\begin{align*}
\tilde{\xi} &= \xi, \\
\tilde{\Omega} &= \Omega, \\
\tilde{\Delta} &= [\Delta, \bar{\Omega} D_{\Omega} W^T], \\
\tilde{\gamma} &= [\gamma, Z + W \xi]^T, \\
\tilde{\Gamma} &= \begin{bmatrix} \Gamma & \Delta^T D_{\Omega} W^T \\ WD_{\Omega} \Delta & (W \Omega W^T + \Sigma) \end{bmatrix}.
\end{align*}
\]

**Theorem 6** Consider the test point \( x \in \mathbb{R}^d \) and the vector \( \hat{f} = \begin{bmatrix} f(X) \\ f(x) \end{bmatrix} := [f \ f_x] \) we have
\[
p(f, f_x) = \text{SkewGP} \left( \begin{bmatrix} \xi(X) \\ \xi(x) \end{bmatrix}, \begin{bmatrix} \Omega(X, X) & \Omega(X, x) \\ \Omega(x, X) & \Omega(x, x) \end{bmatrix}, \begin{bmatrix} \Delta(X) \\ \Delta(x) \end{bmatrix}, \gamma, \Gamma \right)
\]
and the predictive distribution is by definition
\[
\begin{align*}
p(f_x | Z, W) &= \int p(f_x | f) p(f | Z, W) df \\
&= \int p(f_x | f) \frac{p(Z, W | f) p(f)}{p(Z, W)} df \\
&\propto \int p(f_x, f) p(Z, W | f) df.
\end{align*}
\]
We can then apply Lemma 4 with \( \hat{f} \) and the likelihood
\[
p(\begin{bmatrix} Z \\ W \end{bmatrix} | \hat{f}) = \Phi_{m_a+1} \left( \begin{bmatrix} Z \\ W \end{bmatrix} + \begin{bmatrix} f(X) \\ f(x) \end{bmatrix}; \begin{bmatrix} \Sigma & 0 \\ 0 & 1 \end{bmatrix} \right)
\]
which does not depend on \( f(x) \). This results in a posterior distribution
\[
p \left( \begin{bmatrix} f(X) \\ f(x) \end{bmatrix} | \begin{bmatrix} Z \\ W \end{bmatrix} \right) = \text{SUN}_{m_a+1}(\hat{\xi}, \hat{\Omega}, \hat{\Delta}, \hat{\gamma}, \hat{\Gamma})
\]
with
\[
\hat{\xi} = [\xi(X) \quad \xi(x)]^T
\]
\[
\hat{\Omega} = \begin{bmatrix}
\Omega(X,X) & \Omega(X,x) \\
\Omega(x,X) & \Omega(x,x)
\end{bmatrix}
\]
\[
\hat{\Delta} = \begin{bmatrix}
\Delta(X) \\
\Delta(x)
\end{bmatrix}
\begin{bmatrix} D_{\Omega(X,X)}^{-1} & \Omega(X,X)W^T \\
\Omega(x,X) & \Omega(x,x)
\end{bmatrix}
\begin{bmatrix} W_0^T \\
0
\end{bmatrix} = \begin{bmatrix}
\Delta(X) \\
\Delta(x)
\end{bmatrix}
\begin{bmatrix} D_{\Omega(X,X)}^{-1} & \Omega(X,X)W^T \\
\Omega(x,X) & \Omega(x,x)
\end{bmatrix}
\]
\[
\hat{\gamma} = \tilde{\gamma},
\hat{\Gamma} = \tilde{\Gamma},
\]

with \( \tilde{\gamma}, \tilde{\Gamma} \) from Lemma 4. By exploiting the marginalization properties of the SUN distribution, see Section 2.3, we obtain
\[
p(f(x) | W, f(X)) = SUN_{1,s+m_a} \left( \xi(x), \Omega(x,x), \Delta(x), D_{\Omega(x,x)}^{-1} \Omega(x,X)W^T, \tilde{\gamma}, \tilde{\Gamma} \right). \tag{56}
\]

References

R. B Arellano and Adelchi Azzalini. On the unification of families of skew-normal distributions. *Scandinavian Journal of Statistics*, 33(3):561–574, 2006.

Adelchi Azzalini. *The skew-normal and related families*, volume 3. Cambridge University Press, 2013.

Dario Azzimonti and David Ginsbourger. Estimating orthant probabilities of high-dimensional gaussian vectors with an application to set estimation. *Journal of Computational and Graphical Statistics*, 27(2):255–267, 2018.

Alberto Bemporad and Dario Piga. Active preference learning based on radial basis functions. *arXiv preprint arXiv:1909.13049*, 2019.

Alessio Benavoli, Dario Azzimonti, and Dario Piga. Skew Gaussian Processes for Classification. *Machine Learning*, 109:1877–1902, 2020a.

Alessio Benavoli, Dario Azzimonti, and Dario Piga. Preferential Bayesian optimisation with Skew Gaussian Processes. *arXiv preprint arXiv:2008.06677*, 2020b.

Felix Berkenkamp, Andreas Krause, and Angela P Schoellig. Bayesian optimization with safety constraints: safe and automatic parameter tuning in robotics. *arXiv preprint arXiv:1602.04450*, 2016a.

Felix Berkenkamp, Angela P Schoellig, and Andreas Krause. Safe controller optimization for quadrotors with gaussian processes. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 491–496. IEEE, 2016b.

Zdravko I Botev. 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, 2017.
E. Brochu, N. de Freitas, and A. Ghosh. Active preference learning with discrete choice data. In *Advances in neural information processing systems*, pages 409–416, 2008.

Jian Cao, Daniele Durante, and Marc G. Genton. Scalable computation of predictive probabilities in probit models with gaussian process priors. *arxiv.org/abs/2009.01471*, 2020.

Wei Chu and Zoubin Ghahramani. Preference learning with gaussian processes. New York, NY, USA, 2005a. Association for Computing Machinery. ISBN 1595931805.

Wei Chu and Zoubin Ghahramani. Gaussian processes for ordinal regression. *Journal of machine learning research*, 6(Jul):1019–1041, 2005b.

Daniele Durante. Conjugate Bayes for probit regression via unified skew-normals. *Biometrika*, 106(4):765–779, 08 2019.

Marc G Genton, David E Keyes, and George Turkiyyah. Hierarchical decompositions for the computation of high-dimensional multivariate normal probabilities. *Journal of Computational and Graphical Statistics*, 27(2):268–277, 2018.

Alan Genz. Numerical computation of multivariate normal probabilities. *Journal of computational and graphical statistics*, 1(2):141–149, 1992.

Alan Genz and Frank Bretz. *Computation of multivariate normal and t probabilities*, volume 195. Springer Science & Business Media, 2009.

Alexandra Gessner, Oindrila Kanjilal, and Philipp Hennig. Integrals over Gaussians under Linear Domain Constraints. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108, pages 2764–2774. PMLR, 26–28 Aug 2020.

Mark N Gibbs and David JC MacKay. Variational gaussian process classifiers. *IEEE Transactions on Neural Networks*, 11(6):1458–1464, 2000.

Javier González, Zhenwen Dai, Andreas Damianou, and Neil D Lawrence. Preferential bayesian optimization. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1282–1291. JMLR. org, 2017.

GPy, GPy: A gaussian process framework in python. http://github.com/SheffieldML/GPy, since 2012.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Malte Kuss and Carl Edward Rasmussen. Assessing approximate inference for binary Gaussian process classification. *Journal of machine learning research*, 6(Oct):1679–1704, 2005.

David JC MacKay. Bayesian methods for backpropagation networks. In *Models of neural networks III*, pages 211–254. Springer, 1996.

Thomas Peter Minka. A family of algorithms for approximate bayesian inference. In *In UAI*, *Morgan Kaufmann*, pages 362–369, 2001.
Iain Murray, Ryan Adams, and David MacKay. Elliptical slice sampling. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, pages 541–548, Chia, Italy, 13–15 May 2010. PMLR.

Hannes Nickisch and Carl Edward Rasmussen. Approximations for binary gaussian process classification. *Journal of Machine Learning Research*, 9(Oct):2035–2078, 2008.

A O’Hagan and Tom Leonard. Bayes estimation subject to uncertainty about parameter constraints. *Biometrika*, 63(1):201–203, 1976.

Anthony O’Hagan. Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society: Series B (Methodological)*, 40(1):1–24, 1978.

Manfred Opper and Cédric Archambeau. The variational gaussian approximation revisited. *Neural Computation*, 21(3):786–792, 2009.

Ioannis Phinikettos and Axel Gandy. Fast computation of high-dimensional multivariate normal probabilities. *Computational Statistics & Data Analysis*, 55(4):1521–1529, 2011.

Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006.

Dorsa Sadigh, Anca D Dragan, Shankar Sastry, and Sanjit A Seshia. Active preference-based learning of reward functions. In *Robotics: Science and Systems*, 2017.

Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P Adams, and Nando De Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175, 2015.

Yanan Sui, Alkis Gotovos, Joel Burdick, and Andreas Krause. Safe exploration for optimization with gaussian processes. In *International Conference on Machine Learning*, pages 997–1005. PMLR, 2015.

Giang Trinh and Alan Genz. Bivariate conditioning approximations for multivariate normal probabilities. *Statistics and Computing*, 25(5):989–996, 2015.

Carlos Villacampa-Calvo, Bryan Zaldivar, Eduardo C Garrido-Merchán, and Daniel Hernández-Lobato. Multi-class gaussian process classification with noisy inputs. *arXiv preprint arXiv:2001.10523*, 2020.

Christopher KI Williams and David Barber. Bayesian classification with gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(12):1342–1351, 1998.

Masrour Zoghi, Shimon Whiteson, Remi Munos, and Maarten Rijke. Relative upper confidence bound for the k-armed dueling bandit problem. In *Proceedings of the 31st International Conference on Machine Learning*, pages 10–18, Beijing, China, 2014.