# Preferential Bayesian optimisation with Skew Gaussian Processes

Alessio Benavoli
alessio.benavoli@tcd.ie
School of Computer Science and Statistics, Trinity College Dublin, Ireland

Dario Azzimonti
Dario Piga
dario.azzimonti@idsia.ch
dario.piga@idsia.ch
Dalle Molle Institute for Artificial Intelligence USI-SUPSI Lugano, Switzerland

## ABSTRACT

Preferential Bayesian optimisation (PBO) deals with optimisation problems where the objective function can only be accessed via preference judgments, such as "this is better than that" between two candidate solutions (like in A/B tests). The state-of-the-art approach to PBO uses a Gaussian process to model the preference function and a Bernoulli likelihood to model the observed pairwise comparisons. Laplace's method is then employed to compute posterior inferences and, in particular, to build an appropriate acquisition function. In this paper, we prove that the true posterior distribution of the preference function is a Skew Gaussian Process (SkewGP), with highly skewed pairwise marginals and, thus, show that Laplace's method usually provides a very poor approximation. We then derive an efficient method to compute the exact SkewGP posterior and use it as surrogate model for PBO employing standard acquisition functions (Upper-Credible-Bound, etc.). We illustrate the benefits of our exact PBO-SkewGP in a variety of experiments, by showing that it consistently outperforms PBO based on Laplace's approximation both in terms of convergence speed and computational time. We also show that our framework can be extended to deal with mixed preferential-categorical BO, where binary judgments (valid or non-valid) together with preference judgments are available.

## KEYWORDS

Bayesian Optimisation, Bayesian preferential optimisation, Skew Gaussian Processes

**ACM Reference Format:**
Alessio Benavoli, Dario Azzimonti, and Dario Piga. 2021. Preferential Bayesian optimisation with Skew Gaussian Processes. In *2021 Genetic and Evolutionary Computation Conference Companion (GECCO '21 Companion), July 10–14, 2021, Lille, France.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3449726.3463128

## 1 INTRODUCTION

Bayesian optimization (BO) is a powerful tool for global optimisation of expensive-to-evaluate black-box objective functions [5, 14]. However, in many realistic scenarios, the objective function to be optimized cannot be easily quantified. This happens for instance in optimizing chemical and manufacturing processes, in cases where judging the quality of the final product can be a difficult and costly task, or simply in situations where only human preferences are available, like in A/B tests [20]. In such situations, Preferential Bayesian optimization (PBO) [11] or more general algorithms for active preference learning should be adopted [3, 6, 18, 24]. These approaches require the users to simply compare the final outcomes of two different experiments and indicate which they prefer. Indeed, it is well known that humans are better at comparing two options rather than assessing the value of "goodness" of an option [7, 22].

This contribution is focused on PBO. As in the state-of-the-art approach for PBO [11], we use a Gaussian process (GP) as a prior distribution of the *latent preference function* and a probit likelihood to model the observed pairwise comparisons. However, our contribution differs from and improves [11] in several directions.

First, the state-of-the-art PBO methods usually approximate the posterior distribution of the preference function via Laplace's approach. On the other hand, we compute the exact posterior distribution, which we prove to be a Skew Gaussian Process (SkewGP) (recently introduced in [4] for binary classification).[1] Through several examples, we show that the posterior has a strong skewness, and thus any approximation of the posterior that relies on a symmetric distribution (such as Laplace's approximation) results in sub-optimal predictive performances and, thus, slower convergence in PBO.

Second, we propose computationally efficient methods to draw samples from the posterior that are then used to calculate the acquisition function.

Third, we extend standard acquisition functions used in BO to deal with preference observations and propose a new acquisition function for PBO obtained by combining the *dueling information gain* with the expected probability of improvement.

Fourth, we define an *affine probit likelihood* to model the observations. Such a likelihood allows us to handle, in a unified framework, mixed categorical and preference observations. These two different types of information are usually available in manufacturing, where some parameters may cause the process to fail and, therefore, produce no output. In standard BO, where the function evaluation is a scalar, a common way to address this problem is by penalizing the


*GECCO '21 Companion, July 10–14, 2021, Lille, France*

ACM ISBN 978-1-4503-8351-6/21/07.
https://doi.org/10.1145/3449726.3463128

[1] In particular, the present works extends the results in [4] by showing that SkewGPs are conjugate to *probit affine likelihoods*. This allows us to apply SkewGPs for preference learning.

objective function when no output is produced, but this approach is not suitable in PBO as the output is not a scalar.

The rest of the paper is organized as follows. Section 2 reviews skew-normal distributions and SkewGP. The main results of the paper are reported in Section 3, where we show that the posterior distribution of the latent preference function is a SkewGP under the proposed affine probit likelihood. The marginal likelihood is derived and maximized to chose the model's hyper-parameters. An illustrative example is presented in 4 to show the drawbacks of Laplace's approximation, thus highlighting the benefits of computing the exact SkewGP posterior distribution. PBO with SkewGP is discussed in Section 5, where extensive tests with different acquisition functions are reported and clearly show that PBO based on SkewGP consistently outperforms Laplace's approximation both in terms of convergence speed and computational time.

## 2 BACKGROUND ON SKEW-NORMAL DISTRIBUTIONS AND SKEW-GAUSSIAN PROCESSES

In this section we provide details on the skew-normal distribution. The skew-normal [2, 16] is a large class of probability distributions that generalize a normal by allowing for non-zero skewness. A univariate skew-normal distribution is defined by three parameters location $\xi \in \mathbb{R}$, scale $\sigma > 0$ and skew parameter $\alpha \in \mathbb{R}$ and has the following [16] Probability Density Function (PDF)

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

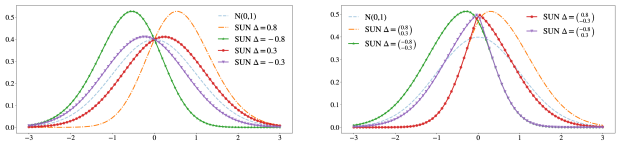
where $\phi$ and $\Phi$ are the PDF and Cumulative Distribution Function (CDF), respectively, of the standard univariate Normal distribution. Over the years many generalisations of this distribution were proposed, in particular [1] provided a unification of those generalizations in a single and tractable 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

A vector $\mathbf{z} \in \mathbb{R}^p$ is distributed as a multivariate Unified Skew-Normal distribution with latent skewness dimension $s$, $\mathbf{z} \sim \mathrm{SUN}_{p,s}(\boldsymbol{\xi}, \Omega, \Delta, \boldsymbol{\gamma}, \Gamma)$, if its probability density function [2, Ch.7] is:

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

where $\phi_p(\mathbf{z}-\boldsymbol{\xi};\Omega)$ represents the PDF of a multivariate Normal distribution with mean $\boldsymbol{\xi} \in \mathbb{R}^p$ and covariance $\Omega = D_\Omega\bar{\Omega}D_\Omega \in \mathbb{R}^{p\times p}$, with $\bar{\Omega}$ being 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 $N_s(0, M)$ evaluated at $\mathbf{a} \in \mathbb{R}^s$. The parameters $\boldsymbol{\gamma} \in \mathbb{R}^s, \Gamma \in \mathbb{R}^{s\times s}, \Delta^{p\times s}$ of the SUN distribution are related to a latent variable that controls the skewness, in particular $\Delta$ is called Skewness matrix.




(a1) $s = 1, \Gamma = 1$ (a2) $s = 2, \Gamma_{1,2} = 0.8$

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

The PDF (1) is well-defined provided that the matrix

$$M := \begin{bmatrix} \Gamma & \Delta^T \\ \Delta & \bar{\Omega} \end{bmatrix} \in \mathbb{R}^{(s+p)\times(s+p)} > 0, \tag{2}$$

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

Figure 1 shows the density of a univariate SUN distribution with latent dimensions $s = 1$ (a1) and $s = 2$ (a2).

For what follows however it is important to know that [see, e.g., 2, Ch.7] the distribution is closed under marginalization and conditioning. We have reviewed these results in Appendix A together with an *additive representation* of the SUN that is useful for sampling from the posterior.

A SkewGP [4] is a generalization of a skew-normal distribution to a stochastic process. Its construction is based on a result derived by [9] for the parametric case, who showed that the skew-normal distribution and probit likelihood are conjugate.

To define a SkewGP, we consider here a location function $\xi : \mathbb{R}^d \to \mathbb{R}$, a scale (kernel) function $\Omega : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, a skewness vector function $\Delta : \mathbb{R}^d \to \mathbb{R}^s$ and the parameters $\boldsymbol{\gamma} \in \mathbb{R}^s, \Gamma \in \mathbb{R}^{s\times s}$. A real function $f : \mathbb{R}^d \to \mathbb{R}$ is a SkewGP with latent dimension $s$, if for any sequence of $n$ points $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^d$, the vector $[f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)] \in \mathbb{R}^n$ is skew-normal distributed with parameters $\boldsymbol{\gamma}, \Gamma$ and location, scale and skewness matrices, respectively, given by

$$\xi(X) := \begin{bmatrix} \xi(\mathbf{x}_1) \\ \xi(\mathbf{x}_2) \\ \vdots \\ \xi(\mathbf{x}_n) \end{bmatrix}, \quad \Omega(X,X) := \begin{bmatrix} \Omega(\mathbf{x}_1,\mathbf{x}_1) & \Omega(\mathbf{x}_1,\mathbf{x}_2) & \ldots & \Omega(\mathbf{x}_1,\mathbf{x}_n) \\ \Omega(\mathbf{x}_2,\mathbf{x}_1) & \Omega(\mathbf{x}_2,\mathbf{x}_2) & \ldots & \Omega(\mathbf{x}_2,\mathbf{x}_n) \\ \vdots & \vdots & \ldots & \vdots \\ \Omega(\mathbf{x}_n,\mathbf{x}_1) & \Omega(\mathbf{x}_n,\mathbf{x}_2) & \ldots & \Omega(\mathbf{x}_n,\mathbf{x}_n) \end{bmatrix},$$
$$\Delta(X) := \begin{bmatrix} \Delta(\mathbf{x}_1) & \Delta(\mathbf{x}_2) & \ldots & \Delta(\mathbf{x}_n) \end{bmatrix}. \tag{3}$$

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 $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^d$ and for any $n$. In that case we write $f \sim \mathrm{SkewGP}_s(\xi, \Omega, \Delta, \gamma, \Gamma)$. Benavoli et al. [4] shows that this is a well defined stochastic process. In the next section we connect this stochastic process to preference learning.

## 3 SKEWGP AND AFFINE PROBIT LIKELIHOOD

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

likelihood as

$$p(W \mid f(X)) = \Phi_m(Wf(X)), \tag{4}$$

where $\Phi_m(\mathbf{x}) := \Phi_m(\mathbf{x}; I_m)$ is the standard $m$-variate Gaussian CDF evaluated at $\mathbf{x} \in \mathbb{R}^m$ with identity covariance matrix. Note that this likelihood model includes the classic GP probit classification model [17] 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 $m = n$. Moreover, as we will show in Corollary 3.3, the likelihood in (4) is equal to the preference likelihood for a particular choice of $W$. This model however also allows to seamlessly mix classification and preference information as we will show below. Here we prove how the skew-normal distribution is connected to the affine probit likelihood, extending a result proved in [9, Th.1 and Co.4] for the parametric setting for a standard probit likelihood.[2]

Theorem 3.1. *Let us assume that $f(\mathbf{x})$ is GP distributed with mean function $\xi(\mathbf{x})$ and covariance function $\Omega(\mathbf{x}, \mathbf{x}')$, that is $f(\mathbf{x}) \sim GP(\xi(\mathbf{x}), \Omega(\mathbf{x}, \mathbf{x}'))$, and consider the likelihood $p(W \mid f(X)) = \Phi_m(Wf(X))$ where $W \in \mathbb{R}^{m \times n}$. The posterior distribution of $f(X)$ is a SUN:*

$$p(f(X)|W) = SUN_{n,m}(\tilde{\xi}, \tilde{\Omega}, \tilde{\Delta}, \tilde{\gamma}, \tilde{\Gamma}) \;\; with$$
$$\tilde{\xi} = \boldsymbol{\xi},\; \tilde{\Omega} = \Omega,\; \tilde{\Delta} = \bar{\Omega} D_\Omega W^T,\; \tilde{\gamma} = W\xi,\; \tilde{\Gamma} = W\Omega W^T + I_m, \tag{5}$$

*where, for simplicity of notation, we denoted $\xi(X), \Omega(X, X)$ as $\boldsymbol{\xi}, \Omega$ and $\Omega = D_\Omega \bar{\Omega} D_\Omega$.*

All the proofs are in Appendix B. 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 (4).

Theorem 3.2. *Let us assume a GP prior $f(\mathbf{x}) \sim GP(\xi(\mathbf{x}), \Omega(\mathbf{x}, \mathbf{x}'))$, the likelihood $p(W \mid f(X)) = \Phi_m(Wf(X))$ with $W \in \mathbb{R}^{m \times n}$, then a-posteriori $f$ is SkewGP with mean function $\xi(\mathbf{x})$, covariance function $\Omega(\mathbf{x}, \mathbf{x}')$, skewness function $\Delta(\mathbf{x}, X) = \Omega(\mathbf{x}, X)W^T$, and $\tilde{\gamma}, \tilde{\Gamma}$ as in* (5).

This is the main result of the paper and allows us to show that, in the case of preference learning, we can compute exactly the posterior and, therefore, Laplace's approximation is not necessary.

## 3.1 Exact preference learning

We now apply results of Theorem 3.1 and Theorem 3.2 to the case of preference learning. For two different inputs $\mathbf{v}_k, \mathbf{u}_k \in X$, a pairwise preference $\mathbf{v}_k \succ \mathbf{u}_k$ is observed, where $\mathbf{v}_k \succ \mathbf{u}_k$ expresses the preference of the instance $\mathbf{v}_k$ over $\mathbf{u}_k$. A set of $m$ pairwise preferences is given and denoted as $\mathcal{D} = \{\mathbf{v}_k \succ \mathbf{u}_k : k = 1, \ldots, m\}$.

*Likelihood.* We assume that there is an underlying hidden function $f : \mathbb{R}^d \to \mathbb{R}$ which is able to describe the observed set of pairwise preferences $\mathcal{D}$. Specifically, given a preference $\mathbf{v}_k \succ \mathbf{u}_k$, then the function $f$ is such that $f(\mathbf{v}_k) \geq f(\mathbf{u}_k)$. To allow tolerances to model noise, we assume that value of the hidden function $f$ is corrupted by a Gaussian noise with zero mean and variance $\sigma^2$.

We use the likelihood introduced in [8], which is the joint probability distribution of observing the preferences $\mathcal{D}$ given the values of the function $f$ at $X$, i.e.,

$$p(\mathcal{D} \mid f(X)) = \prod_{k=1}^{m} p(\mathbf{v}_k \succ \mathbf{u}_k | f(\mathbf{v}_k), f(\mathbf{u}_k)) = \prod_{k=1}^{m} p(f(\mathbf{v}_k) - f(\mathbf{u}_k) \geq 0)$$
$$= \prod_{k=1}^{m} \Phi\left(\frac{f(\mathbf{v}_k) - f(\mathbf{u}_k)}{\sqrt{2}\sigma}\right) = \Phi_m\left(\begin{pmatrix} \frac{f(\mathbf{v}_1) - f(\mathbf{u}_1)}{\sqrt{2}\sigma} \\ \vdots \\ \frac{f(\mathbf{v}_m) - f(\mathbf{u}_m)}{\sqrt{2}\sigma} \end{pmatrix}\right). \tag{6}$$

For identifiability reasons, without loss of generality, we set $\sigma^2 = \frac{1}{2}$.[3]

*Posterior.* The posterior distribution of the values of the hidden function $f$ at all $\mathbf{x} \in X$ given the observations $\mathcal{D}$ is then:

$$p(f(X) \mid \mathcal{D}) = \frac{p(f(X))}{p(\mathcal{D})} \Phi_m\left(\begin{pmatrix} f(\mathbf{v}_1) - f(\mathbf{u}_1) \\ \cdots \\ f(\mathbf{v}_m) - f(\mathbf{u}_m) \end{pmatrix}\right). \tag{7}$$

In state-of-art PBO [11], a Laplace's approximation of the posterior $p(f(X) \mid \mathcal{D})$ is used to construct the acquisition function. The following Corollary shows that the posterior $p(f(X) \mid \mathcal{D})$ is distributed as a SkewGP.

Corollary 3.3. *Consider $f(\mathbf{x}) \sim GP(\xi(\mathbf{x}), \Omega(\mathbf{x}, \mathbf{x}'))$ and the likelihood $p(\mathcal{D} \mid f(X))$ in* (6). *If we denote by $W \in \mathbb{R}^{m \times n}$ the matrix defined as $W_{i,j} = V_{i,j} - U_{i,j}$ where $V_{i,j} = 1$ if $\mathbf{v}_i = \mathbf{x}_j$ and 0 otherwise and $U_{i,j} = 1$ if $\mathbf{u}_i = \mathbf{x}_j$ and 0 otherwise. Then the posterior of $f(X)$ is given by* (5).

In PBO, in order to compute the acquisition functions, we must be able to draw efficiently independent samples from the posterior in Theorem 3.2.

Proposition 3.4. *Given a test point $\mathbf{x}$, posterior samples of $f(\mathbf{x})$ can be obtained as:*

$$f(\mathbf{x}) \sim \tilde{\xi}(\mathbf{x}) + D_{\Omega(\mathbf{x},\mathbf{x})}\left(U_0 + \Omega(\mathbf{x}, X)W^T(W\Omega(X, X)W^T + I_m)^{-1}U_1\right), \tag{8}$$
$$U_0 \sim \mathcal{N}(0; \bar{\Omega}(\mathbf{x}, \mathbf{x}) - \Omega(\mathbf{x}, X)W^T\tilde{\Gamma}^{-1}W\Omega(\mathbf{x}, X)^T),$$
$$U_1 \sim \mathcal{T}_{\tilde{\boldsymbol{\gamma}}}(0; W\Omega(X, X)W^T + I_m),$$

*where $\mathcal{T}_{\tilde{\boldsymbol{\gamma}}}(0; \tilde{\Gamma})$ is the pdf of a multivariate Gaussian distribution with zero mean and covariance $\tilde{\Gamma}$ truncated component-wise below $-\tilde{\boldsymbol{\gamma}} = -W\xi(X)$.*

Note that sampling $U_0$ can be achieved efficiently with standard methods, however using standard rejection sampling for the variable $U_1$ would incur in exponentially growing sampling time as the dimension $m$ increases. Here we use the recently introduced sampling technique *linear elliptical slice sampling* (*lin-ess*,Gessner et al. [10]) which improves Elliptical Slice Sampling (*ess*, Murray et al. [15]) 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. Since *lin-ess* is rejection-free,[4] we can compute exactly the computation complexity of (8): $O(n^3)$ with storage demands of $O(n^2)$. SkewGPs have

[2] [9, Th.1 and Co.4] assumes that the matrix $W$ is diagonal, but the same results can straightforwardly extend to generic $W$.

[3] Equivalently, we instead estimate the kernel variance.

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

similar bottleneck computational complexity of full GPs. Finally, observe that $U_1$ does not depend on $\mathbf{x}$ and, therefore, we do not need to re-sample $U_1$ to sample $f$ at another test point $\mathbf{x}'$. This is fundamental because acquisition functions are functions of $\mathbf{x}$ and, we need to optimize them in PBO.

*Marginal likelihood.* Here, we follow the usual GP literature ([17]) and we consider a zero mean function $\xi(\mathbf{x}) = 0$ and a parametric covariance kernel $\Omega(\mathbf{x}, \mathbf{x}')$ indexed by $\theta \in \Theta$. Typically, $\theta$ contains lengthscale parameters and a variance parameter. For instance, for the RBF kernel

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

we have that $\theta = [\ell, \sigma]$.[5] The parameters $\theta$ are chosen by maximizing the marginal likelihood, that for SkewGP is provided hereafter.

Corollary 3.5. *Consider a GP prior $f(\mathbf{x}) \sim GP(\xi(\mathbf{x}), \Omega(\mathbf{x}, \mathbf{x}'))$ and the likelihood $p(\mathcal{D} \mid f(X)) = \Phi_m(Wf(X))$, then the marginal likelihood of the observations $\mathcal{D}$ is*

$$p(\mathcal{D}) = \Phi_m(\tilde{\boldsymbol{\gamma}};\ \tilde{\Gamma})\ \left(\geq \sum_{i=1}^{b} \Phi_{|B_i|}(\tilde{\boldsymbol{\gamma}}_{B_i};\ \tilde{\Gamma}_{B_i}) - (b-1)\right), \quad (9)$$

*with $\tilde{\boldsymbol{\gamma}}, \tilde{\Gamma}$ defined in Theorem 3.2 (they depend on $\theta$).*

If the size of $W$ is too large the evaluation of $\Phi_m$ could become infeasible, therefore here we use the approximation introduced in [4], see inequality in (9) 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, $\tilde{\boldsymbol{\gamma}}_{B_i}$, $\tilde{\Gamma}_{B_i}$ are the parameters of the posterior computed using only the subset $B_i$ of the data (in the experiments $|B_i| = 30$). Details about the routine we use to compute $\Phi_{|B_i|}(\cdot)$ and the optimization method we employ to maximise the lower bound in (9) are in Appendix C.

## 3.2 Mixed classification and preference information

Consider now a problem where we have two types of information: whether a certain instance is preferable to another (preference-like observation) and whether a certain instance is attainable or not (classification-like observation). Such situation often comes up in industrial applications. For example imagine a machine that produces a product whose final quality depends on certain input parameters. Assume now that certain values of the input parameters produce no product. In this case we might want to evaluate the quality of the product with binary comparisons (preferences) along with a binary class that indicates whether the input configuration is valid. By using only a preference likelihood or a classification likelihood we would not be using all information. In this case, observations are in fact pairs and the space of possibility is $\mathcal{Z} = \{(\text{valid}, \mathbf{v}_k \succ \mathbf{u}_k),\ (\text{valid}, \mathbf{u}_k \succ \mathbf{v}_k),\ (\text{non-valid}, None)\}$, where $\mathbf{v}_k$ and $\mathbf{u}_k$ are respectively the current and reference input. Note that the reference input is always valid, while the current input could be valid or not. We propose a new likelihood function to model the above setting, which is defined as follows:

$$P(z_k | f(\mathbf{v}_k), f(\mathbf{u}_k)) \quad (10)$$
$$= \begin{cases} \Phi\left(f(\mathbf{v}_k)\right)\Phi\left(f(\mathbf{v}_k) - f(\mathbf{u}_k)\right), & z_k = (\text{valid}, \mathbf{v}_k \succ \mathbf{u}_k) \\ \Phi\left(f(\mathbf{v}_k)\right)\Phi\left(f(\mathbf{u}_k) - f(\mathbf{v}_k)\right), & z_k = (\text{valid}, \mathbf{u}_k \succ \mathbf{v}_k) \\ \Phi\left(-f(\mathbf{v}_k)\right), & z_k = (\text{non-valid}, None). \end{cases}$$

It is then immediate to write the above likelihood (10) in the form (4) and, therefore, use both sources of information. We associate to each point $\mathbf{x}_i$ a binary output $y_i \in \{0, 1\}$ where the class 0 denotes a non-valid output. In case of valid output, we assume that we conducted $m$ comparisons obtaining the couples $\mathcal{D} = \{\mathbf{v}_k \succ \mathbf{u}_k : k = 1, \ldots, m\}$ where $\mathbf{v}_k \succ \mathbf{u}_k$ expresses the preference of the instance $\mathbf{v}_k$ over $\mathbf{u}_k$. The likelihood is then a product of two independent probit likelihood functions

$$p_{\text{class}}(W_{\text{class}} \mid f(X)) = \Phi_n\left(\begin{bmatrix} 2y_1-1 & 0 & \cdots & 0 \\ 0 & 2y_2-1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 2y_n-1 \end{bmatrix} f(X)\right),$$
$$p_{\text{pref}}(W_{\text{pref}} \mid f(X)) = \Phi_m\left(\begin{bmatrix} f(\mathbf{v}_1)-f(\mathbf{u}_1) \\ \vdots \\ f(\mathbf{v}_m)-f(\mathbf{u}_m) \end{bmatrix}\right)$$

Since we assume that the two likelihood are independent we can compute the overall likelihood:

$$p(W_{\text{class}}, W_{\text{pref}} \mid f(X)) = p_{\text{class}}(W_{\text{class}} \mid f(X)) p_{\text{pref}}(W_{\text{pref}} \mid f(X)) = \Phi_{n+m}(Wf(X)), \quad (11)$$

with $W = \begin{bmatrix} W_{\text{class}} \\ W_{\text{pref}} \end{bmatrix} \in \mathbb{R}^{(n+m)\times n}$, where $W_{\text{pref}}$ is the matrix of preferences defined as in Corollary 3.3. Therefore, the results in Section 3 still holds in this mixed setting.

# 4 COMPARISON SKEWGP VS. LAPLACE'S APPROXIMATION

We provide a one-dimensional illustration of the difference between Gaussian Process with Laplace's approximation (GPL) and SkewGP.

Consider the non-linear function $g(x) = cos(5x) + e^{-\frac{x^2}{2}}$ which has a global maximum at $x = 0$. We assume we can only query this function through pairwise comparisons. We generate 7 pairwise random comparisons: the query point $x_i$ is preferred to $x_j$ (that is $x_i \succ x_j$) if $g(x_i) > g(x_j)$. Figure 2(top-left) shows $g(x)$ and the location of the queried points.[6] Figure 2(bottom-left) shows the predicted posterior preference function $f(x)$ (and relative 95% credible region) computed according to GPL and SkewGP. Both the methods have the same prior: a GP with zero mean and RBF covariance function (the hyperparameters are the same for both methods and have been set equal to the values that maximise Laplace's approximation to the marginal likelihood, $l = 0.35$ and $\sigma^2 = 0.02$). Therefore, the only difference between the two posteriors is due to the Laplace's approximation. The true posterior (SkewGP) of the preference function is skewed, this can be seen from the density plot for $f(-0.51)$ in Figure 2(top-right). Figure 2(bottom-right) shows an example, $f(0.19)$, where SkewGP and Laplace's approximation differ significantly: Laplace's approximation heavily underestimates the mean and the

[5] In the numerical experiments, we use a RBF kernel with ARD and so we have a lengthscale parameter for each component of $\mathbf{x}$.

[6] The preferences between the queried points are $1.25 \succ -1.8$, $-1.23 \succ 1.25$, $0.18 \succ -1.23$, $0.18 \succ -2.52$, $-2.52 \succ 2.18$, $-1.8 \succ -0.5$, $-1.8 \succ 0.67$.

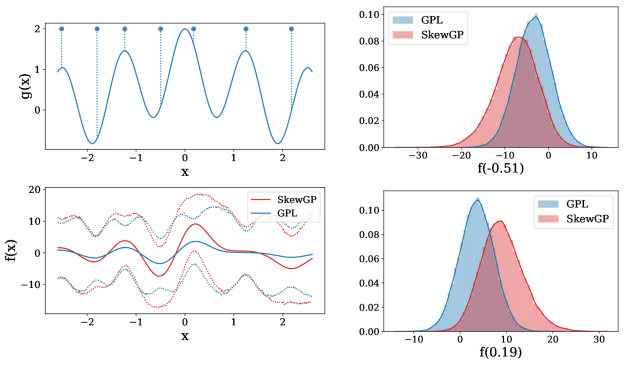



**Figure 2: Comparison between Laplace's approximation (GPL) and the exact posterior (SkewGP). Top-left shows $g(x)$. Bottom-left shows the predicted posterior preference function $f(x)$ (continuous lines) and the relative 95% credible region (dashed lines) for GPL and SkewGP. Right-column reports the density plots for $f(0.19)$ (bottom) and $f(-0.51)$ (top) for both models.**

support of the true posterior (SkewGP) also evident from Figure 2(bottom-left). These differences determine the poor performance of PBO based on GPL as we will see in the next sections.

## 5 DUELING ACQUISITION FUNCTIONS

In sequential BO, our objective is to seek a new data point $\mathbf{x}$ which will allow us to get closer to the maximum of the target function $g$. Since $g$ can only be queried via preferences, this is obtained 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, that is the point that has the highest probability of winning most of the duels (given the observed data $\mathcal{D}$) and, therefore, it is the most likely point maximizing $g$.[7] We consider three pairwise modifications of standard acquisition functions: (i) Upper Credible Bound (UCB); (ii) Thompson sampling (TH); (iii) 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 $f(\mathbf{x}) - f(\mathbf{x}_r)$.

**TH:** The dueling Thompson acquisition function is $f_j(\mathbf{x}) - f_j(\mathbf{x}_r)$, where $f_j$ is a sampled function from the posterior.

**EIIG:** We now propose the dueling EIIG that is the combination of the expected probability of improvement (in log-scale) and the dueling information gain:

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

where $IG(\mathbf{x}, \mathbf{x}_r) = h\left(E_{f\sim p(f|\mathcal{D})}\left(\Phi\left(\frac{f(\mathbf{x})-f(\mathbf{x}_r)}{\sqrt{2}\sigma}\right)\right)\right)$
$-E_{f\sim p(f|\mathcal{D})}\left(h\left(\Phi\left(\frac{f(\mathbf{x})-f(\mathbf{x}_r)}{\sqrt{2}\sigma}\right)\right)\right)$, with $h(p) = -p\log(p) - (1-p)\log(1-p)$ being the binary entropy function of $p$. This definition of dueling information gain is an extension to preferences of the information gain, formulated for GP classifiers in [13]. This last acquisition function allows us to balance exploration-exploitation by means of the nonnegative scalar $k$ (in the experiments we use $k = 0.1$ (more exploration) and $k = 0.5$). To compute these acquisition functions, we make explicitly use of the generative capabilities of our SkewGP surrogated model as well as the efficiency of sampling the learned preference function.

Note that, [11] use different acquisition functions based on the Copland's score (to increase exploration). Moreover, they optimize $\alpha(\mathbf{x}_a, \mathbf{x}_b)$ with respect to both $\mathbf{x}_a, \mathbf{x}_b$, while $\mathbf{x}_b$ is fixed and equal to $\mathbf{x}_r$ in our setting. SkewGP can easily be employed as surrogate model in [11] PBO setting (and very likely improve their performance due to the limits of the Laplace's approximation). We have focused on the above acquisition functions (UCB, TH, EIIG) because they can easily be computed – instead the Copland's score requires to numerically compute an integral with respect to $\mathbf{x}$.

## 6 NUMERICAL EXPERIMENTS

In this section we present numerical experiments to validate our PBO-SkewGP and compare it with PBO based on the Laplace's approximation (PBO-GPL).[8]

First, we consider again the maximization of $g(x) = cos(5x) + e^{-\frac{x^2}{2}}$ and the same 7 initial preferences used in Section 4.

We run PBO for 4 iterations and, at each step, we query the point that maximises the UCB of GPL. We also compute the true posterior and the true maximum of UCB using SkewGP for comparison. Both the methods have the same prior: a GP with zero mean and RBF covariance function (the hyperparameters are fixed to the same values for both methods, that is the values that maximise Laplace's approximation to the marginal likelihood, $l = 0.35$ and $\sigma^2 = 0.02$). Therefore, the only difference between the two posteriors is due to the Laplace's approximation.

Figure 3 shows, for each iteration, a row with three plots. The left plot reports $g(x)$ and the queried points ($x_r$ is in orange). The central plot shows the GPL and SkewGP posterior predictive means of $f(x) - f(x_r)$ and the relative 95% credible intervals. The maximum of each UCB is showed with a star marker. The right plot shows the skewness statistics for the SkewGP predictive posterior distribution of $f(x) - f(x_r)$ as a function of $x$, defined as: $SS(f(x) - f(x_r)) := \frac{\mathrm{E}\left[(f(x)-\mu)^3\right]}{(\mathrm{E}\left[(f(x)-\mu)^2\right])^{3/2}}$, with $\mu := \mathrm{E}\left[f(x)\right]$, and computed via Monte Carlo sampling from the posterior.

Figure 3 shows that the Laplace approximation for $f(x) - f(x_r)$ is much worse than SkewGP. Moreover, the posterior of $f(x) - f(x_r)$ is heavily skewed. The maximum magnitude of $SS(f(x) - f(x_r))$ is $-1.3$ (see the relative marginal posteriors in Figure 4(left)).

Note from Figure 3(1st-row, central) that, while the maximum of UCB for GPL and SkewGP almost coincides in the initial iteration, they significantly differ in the second iteration, Figure 3(2nd-row, central): SkewGP's UCB selects a point very close to the global maximum, while GPL explores the border of the search space. This

[7] By optimizing the acquisition function $\alpha(\mathbf{x}, \mathbf{x}_r)$, we aim to find a point that is better than $\mathbf{x}_r$ (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 $\mathbf{x}_n \overset{?}{\succ} \mathbf{x}_r$. If $\mathbf{x}_n \succ \mathbf{x}_r$ then $\mathbf{x}_n$ becomes the new reference point ($\mathbf{x}_r$) for the next iteration.

[8] A notebook to partially reproduce these results is available at https://github.com/benavoli/SkewGP.

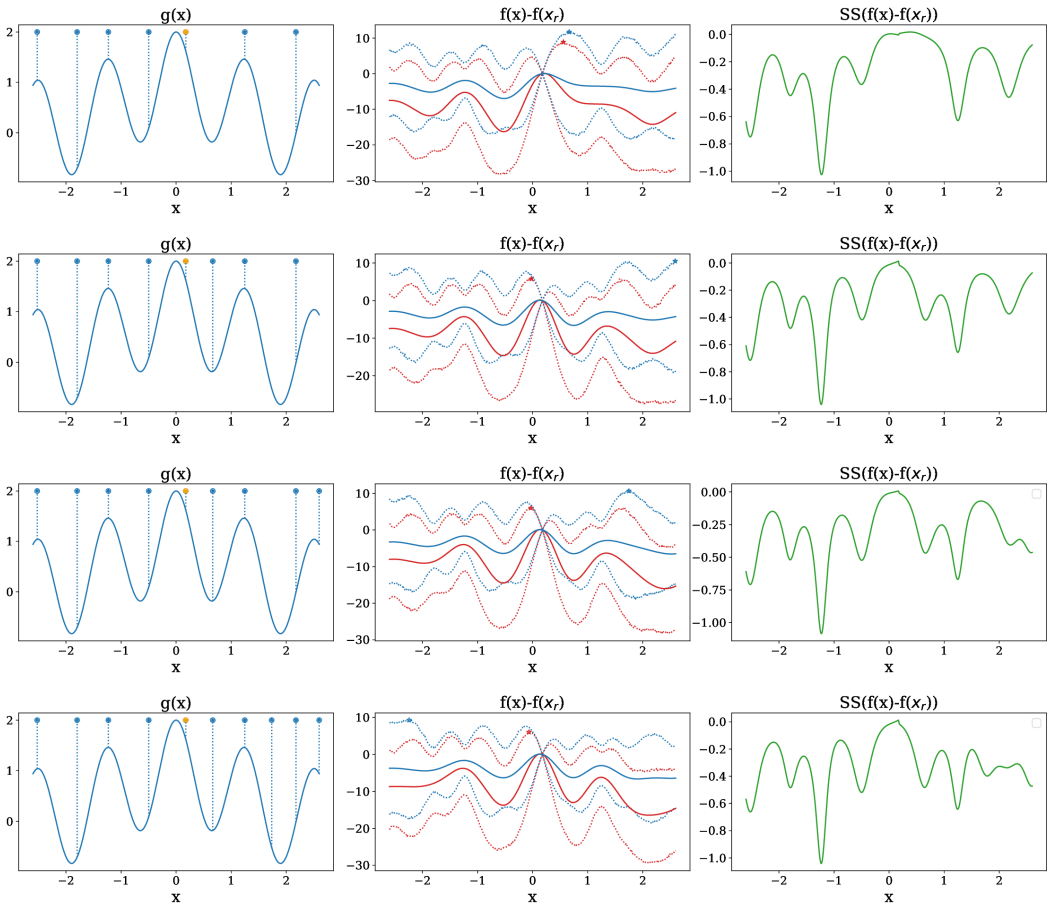


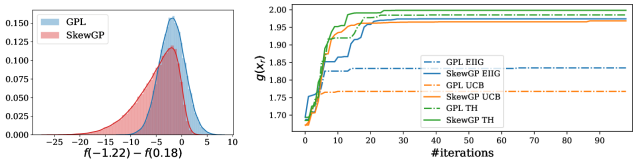

**Figure 3: PBO run for 4 iterations. In each iteration, a row with three plots is showed. The left plot shows the objective function and the queried points ($x_r$ is in orange). The central plot shows the GPL (blue) and SkewGP (red) posterior predictive means of $f(x) - f(x_r)$ and the relative 95% credible intervals. The maximum of each UCB is showed with a star marker. The right plot shows the skewness statistics for the SkewGP predictive posterior distribution of $f(x) - f(x_r)$. At each step, we query the point that maximises the UCB of GPL.**

again happens in the subsequent two iterations, see 3(3nd-row and 4th-row, central).

This behavior is neither special to this trial nor to the UCB acquisition function, we repeat this experiment 20 times starting with 10 initial (randomly selected) duels and using all the three acquisition functions. We compare GPL versus SkewGP with fixed kernel hyperparameters (same as above) so that the only difference between the two algorithms is in the computation of the posterior.[9] We report the average (over the 20 trials) performance, defined as the value of $g$ evaluated at the current optimal point $\mathbf{x}_r$, considered to be the best by each method at each one of the 100 iterations.

The results are showed in Figure 4(right): SkewGP always outperforms GPL. This is only due to Laplace's approximation (hyperparameters are the same). In 1D, the differences are smaller for the Thompson (TH) acquisition function (due to the "noise" from the random sampling step). However, SkewGP-TH converges faster.

[9]In our implementation we compute the acquisition functions via Monte Carlo sampling (2000 samples).




(a) (b)

**Figure 4: Left: skewed marginal. Right: convergence speed (EIIG $k = 0.1$).**

## 6.1 Optimization benchmarks

We have considered for $g(\mathbf{x})$ the same four benchmark functions used in [11]: 'Forrester' (1D), 'Six-Hump Camel' (2D), 'Gold-Stein'

(2D) and 'Levy' (2D), and additionally 'Rosenbrock5' (5D) and 'Hartman6' (6D). These are minimization problems.[10]

Each experiment starts with 10 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 [11]. We compare PBO based on GPL versus SkewGP using the three different acquisition functions described in Section 5: UCB, EIIG ($k = 0.1$ and 0.5) and Thompson (denoted as TH). As before, we show plots of #iterations versus $g(\mathbf{x}_r)$. In these experiments we optimize the kernel hyperparameters by maximising the marginal likelihood for both GPL and SkewGP.

Figure 5 reports the performance of the different methods. Consistently across all benchmarks PBO-SkewGP outperforms PBO-GPL no matter the acquisition function. PBO-SkewGP has also a lower computational burden as showed in the Table at the bottom of Figure 5 that compares the median (over 80 trials, that is 20 trials times 4 acquisition functions) computational time per 100 iterations in seconds (on a standard laptop).

## 6.2 Mixed preferential-categorical BO

We examine now situations where certain unknown values of the inputs produce no-output preference and address it as a mixed preferential-categorical BO as described in Section 3.2.

First, we consider again $g(x) = cos(5x) + e^{-\frac{x^2}{2}}$ and assume that any input $x \leq -0.2$ produces a non-valid output. Figure 6 shows the predicted posterior preference function $f(x)$ (and relative 95% credible region) computed according to SkewGP.[11] We can see how SkewGP learns the non-valid region: the posterior mean is negative for $x \lesssim -0.2$ and positive otherwise (the oscillations of the mean for $x \gtrsim -0.2$ capture the preferences). This is consistent with the likelihood in (11).

We consider now the following benchmark optimization problem proposed in [19]:

$$\min 2+\tfrac{1}{100}(x_2-x_1^2)^2+(1-x_1)^2+2(2-x_2)^2+7\sin(\tfrac{1}{2}x_1)\sin(\tfrac{7}{10}x_1x_2)$$

with $0 \leq x_1, x_2 \leq 5$ and we assume that the input is *valid* if $h(\mathbf{x}) = -\sin(x_1 - x_2 - \frac{\pi}{8}) \leq 0$. We also assume that both the objective function and $h(\mathbf{x})$ are unknown. The goal is to find the minimum: $x^\star = [2.7450\ 2.3523]'$ with optimal cost $-1.1743$. Figure 7(left) shows the level curves of the objective function, the non-valid zone (grey bands) and the location of the minimizer (red star). We compare two approaches. The first approach uses PBO-SkewGP that minimizes the objective plus a penalty term, $10^8 \max(0, h(\mathbf{x}))^2$, that is non-zero in the *non-valid* region. Adding a penalty for non-valid inputs is the most common approach to deal with this type of problems. The second approach uses a SkewGP based mixed preferential-categorical BO (SkewGP-mixed) that accounts for the *valid*/*non-valid* points as in Section 3.2. Figure 7(right) shows the performance of the two compared methods, which is consistent across the three different acquisition functions: SkewGP-mixed converges more quickly to the optimum. This confirms that modelling directly this type of problems via the mixed preferential-categorical likelihood in (11) enables the model to fully exploit the available information. Also in this case, SkewGP allows us to compute the corresponding posterior exactly.

[10] We converted them into maximizations so that the acquisition functions in Section 5 are well-defined.

[11] The preferences are $0.18 \succ 1.25$, $2.18 \succ 0.67$, $0.18 \succ 2.18$, $1.25 \succ 0.67$, $0.18 \succ 0.67$.

## 7 CONCLUSIONS

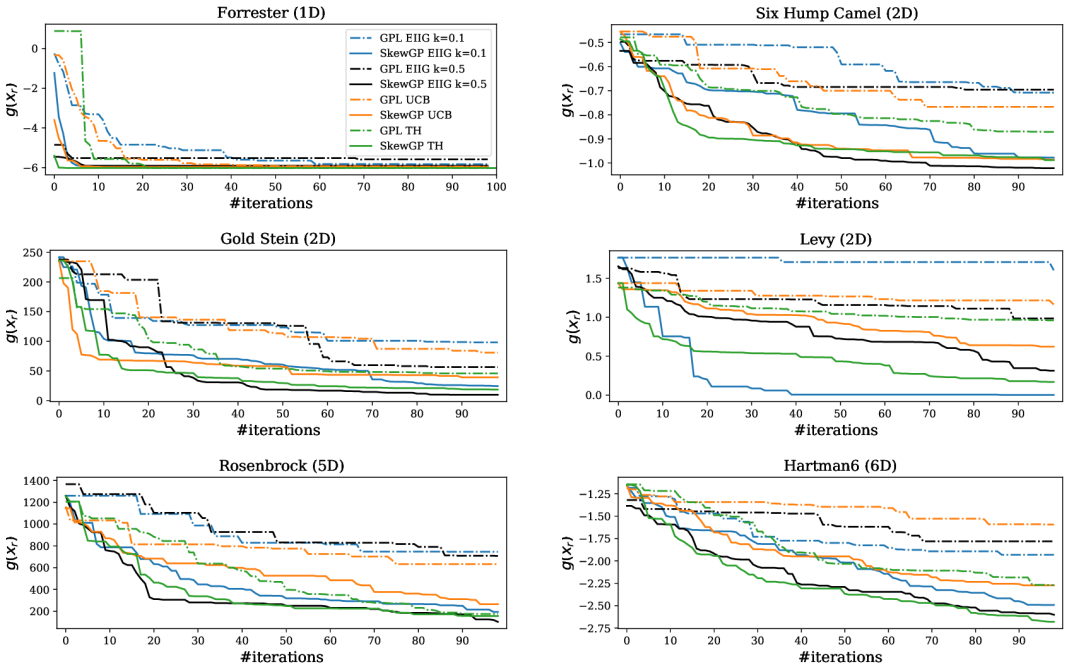
In this work we have shown that is possible to perform exact preferential Bayesian optimization by using Skew Gaussian processes. We have demonstrated that in the setting of preferential BO: (i) the Laplace's approximation is very poor; (ii) given the strong skewness of the posterior, any approximation of the posterior that relies on a symmetric distribution will result to sub-optimal predictive performances and, therefore, slower convergence in PBO. We have also shown that we can extend this model to deal with mixed preferential-categorical Bayesian optimisation, while still providing the exact posterior. We envisage an immediate extension of our current approach. Many optimisation applications are subject to safety constraints, so that inputs cannot be freely chosen from the entire input space. This leads to so-called safe Bayesian optimization [12], that has been extended to safe preferential BO in [21]. We plan to solve this problem exactly using SkewGP.

## ACKNOWLEDGMENTS

D. Azzimonti gratefully acknowledges support from the Swiss National Research Programme 75 "Big Data" Grant No. 407540_167199 / 1.

## REFERENCES

[1] R. B Arellano and Adelchi Azzalini. 2006. On the unification of families of skew-normal distributions. *Scandinavian Journal of Statistics* 33, 3 (2006), 561–574.
[2] Adelchi Azzalini. 2013. *The skew-normal and related families*. Vol. 3. Cambridge University Press.
[3] Alberto Bemporad and Dario Piga. 2019. Active preference learning based on radial basis functions. *arXiv preprint arXiv:1909.13049* (2019).
[4] Alessio Benavoli, Dario Azzimonti, and Dario Piga. 2020. Skew Gaussian Processes for Classification. *accepted to Machine Learning, arXiv preprint arXiv:2005.12987* (2020).
[5] E. Brochu, V.M. Cora, and N. De Freitas. 2010. A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599* (2010).
[6] E. Brochu, N. de Freitas, and A. Ghosh. 2008. Active preference learning with discrete choice data. In *Advances in neural information processing systems*. 409–416.
[7] A. Chernev, U. Böckenholt, and J. Goodman. 2015. Choice overload: A conceptual review and meta-analysis. *Journal of Consumer Psychology* 25, 2 (2015), 333–358.
[8] Wei Chu and Zoubin Ghahramani. 2005. Preference Learning with Gaussian Processes. In *Proceedings of the 22nd International Conference on Machine Learning* (Bonn, Germany) *(ICML '05)*. Association for Computing Machinery, New York, NY, USA, 137–144. https://doi.org/10.1145/1102351.1102369
[9] Daniele Durante. 2019. Conjugate Bayes for probit regression via unified skew-normal distributions. *Biometrika* 106, 4 (08 2019), 765–779.
[10] Alexandra Gessner, Oindrila Kanjilal, and Philipp Hennig. 2019. Integrals over Gaussians under Linear Domain Constraints. *arXiv preprint arXiv:1910.09328* (2019).
[11] Javier González, Zhenwen Dai, Andreas Damianou, and Neil D Lawrence. 2017. Preferential bayesian optimization. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 1282–1291.
[12] Alkis Gotovos, ETHZ CH, Joel W Burdick, and CALTECH EDU. 2015. Safe Exploration for Optimization with Gaussian Processes. In *International Conference on Machine Learning (ICML).[Sui, Yue, and Burdick 2017] Sui, Y.*
[13] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745* (2011).
[14] Jonas Mockus. 2012. *Bayesian approach to global optimization: theory and applications*. Vol. 37. Springer Science & Business Media.
[15] Iain Murray, Ryan Adams, and David MacKay. 2010. Elliptical slice sampling. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*. PMLR, Chia, Italy, 541–548.




| | Forrester | Six Hump Camel | Gold Stein | Levy | Rosenbrock5 | Hartman6 |
|---|---|---|---|---|---|---|
| GPL PBO | 557 | 2986 | 3346 | 4704 | 4314 | 5801 |
| SkewGP PBO | 102 | 2276 | 1430 | 1211 | 2615 | 2178 |

**Figure 5: Averaged results over 20 trials for GPL versus SkewGP on the 6 benchmark functions considering 3 different acquisition functions. The x-axis represents the number of evaluation and the y-axis represents the value of the true objective function at the current optimum $x_r$. The table reports the median computational time per 100 iterations in seconds.**

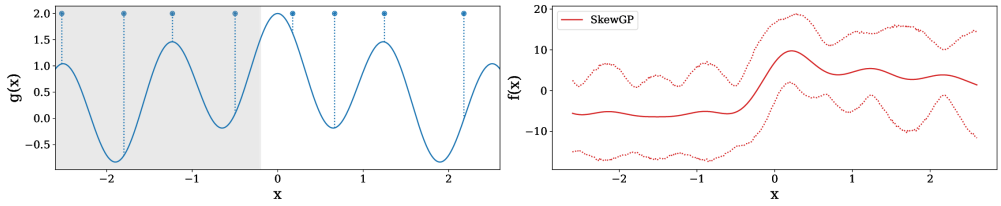



**Figure 6: Mixed preference-classification BO**

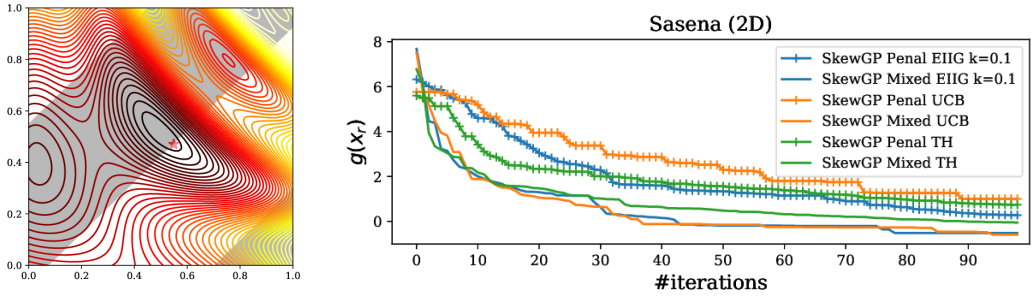



**Figure 7: Left: level sets of the benchmark function, non-valid domain (grey bands) and location of the minimum (red star). Right: averaged results across 20 trials for SkewGP penalised PBO vs. SkewGP mixed preference-classification BO using 3 different acquisition functions.**

[16] A O'Hagan and Tom Leonard. 1976. Bayes estimation subject to uncertainty about parameter constraints. *Biometrika* 63, 1 (1976), 201–203.
[17] Carl Edward Rasmussen and Christopher KI Williams. 2006. *Gaussian processes for machine learning*. MIT press Cambridge, MA.
[18] Dorsa Sadigh, Anca D Dragan, Shankar Sastry, and Sanjit A Seshia. 2017. Active Preference-Based Learning of Reward Functions. In *Robotics: Science and Systems*.
[19] Michael J Sasena, Panos Papalambros, and Pierre Goovaerts. 2002. Exploration of metamodeling sampling criteria for constrained global optimization. *Engineering optimization* 34, 3 (2002), 263–278.
[20] Dan Siroker and Pete Koomen. 2013. *A/B testing: The most powerful way to turn clicks into customers*. John Wiley & Sons.
[21] Yanan Sui, vincent Zhuang, Joel Burdick, and Yisong Yue. 2018. Stagewise Safe Bayesian Optimization with Gaussian Processes. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, Stockholmsmässan, Stockholm Sweden, 4781–4789. http://proceedings.mlr.press/v80/sui18a.html
[22] L.L. Thurstone. 1927. A law of comparative judgment. *Psychological review* 34, 4 (1927), 273.
[23] Giang Trinh and Alan Genz. 2015. Bivariate conditioning approximations for multivariate normal probabilities. *Statistics and Computing* 25, 5 (2015), 989–996.
[24] Masrour Zoghi, Shimon Whiteson, Remi Munos, and Maarten Rijke. 2014. Relative Upper Confidence Bound for the K-Armed Dueling Bandit Problem. In *Proceedings of the 31st International Conference on Machine Learning*. Bejing, China, 10–18.