A BACKGROUND ON THE SKEW-NORMAL DISTRIBUTION

In this section we provide more details on the skew-normal distribution. See, e.g. [2] for a complete reference on skew-normal distributions and their parametrizations.

A.1 Additive representations

The role of the latent dimension s can be briefly explained as follows. Consider a random vector $\begin{bmatrix}\mathbf{x}_0\\ \mathbf{x}_1\end{bmatrix} \sim N_{s+p}(0, M)$ with M as in (2) and define $\mathbf{y}$ as the vector with distribution $(\mathbf{x}_1 \mid \mathbf{x}_0 + \boldsymbol{\gamma} > 0)$. The density of y can be written as

$$f(\mathbf{y}) = \frac{\int_{\mathbf{x}_0+\boldsymbol{\gamma}>0} \varphi_{s+p}((\mathbf{t}_0, \mathbf{y}); M) d\mathbf{t_0}}{\int_{\mathbf{x}_0+\boldsymbol{\gamma}>0} \varphi_s(\mathbf{t}; \Gamma) d\mathbf{t}} = \varphi_p(\mathbf{y}; \bar{\Omega}) \frac{P(\mathbf{x}_0 + \boldsymbol{\gamma} > 0 \mid \mathbf{x}_1 = \mathbf{y})}{\Phi_s(\boldsymbol{\gamma}; \Gamma)}$$
$$= \varphi_p(\mathbf{y}; \bar{\Omega}) \frac{\Phi_s\left(\boldsymbol{\gamma} + \Delta^T \bar{\Omega}^{-1} \mathbf{y}; \Gamma - \Delta^T \bar{\Omega}^{-1} \Delta\right)}{\Phi_s(\boldsymbol{\gamma}; \Gamma)},$$

where the first equality comes from a basic property of conditional distributions, see, e.g.[2, Ch. 1.3], and the second equality is a consequence of the multivariate normal conditioning properties. Then we have that $\mathbf{z} = \boldsymbol{\xi} + D_\Omega \mathbf{y} \sim \mathrm{SUN}_{p,s}(\boldsymbol{\xi}, \Omega, \Delta, \boldsymbol{\gamma}, \Gamma)$.

The previous representation provides an interesting point of view on the skew-Gaussian random vector, however the following representation turns out to be more practical for sampling from this distribution. Consider the independent random vectors $\mathbf{u}_0 \sim N_p(0, \bar{\Omega} - \Delta\Gamma^{-1}\Delta^T)$ and $\mathbf{u}_{1,-\boldsymbol{\gamma}}$, the truncation below $\boldsymbol{\gamma}$ of $\mathbf{u}_1 \sim N_s(0, \Gamma)$. Then the random variable

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

is distributed as (1).

Proof. We can show that the representation $\mathbf{x}_0, \mathbf{x}_1$ is equivalent to $\mathbf{u}_0, \mathbf{u}_1$. Define $\mathbf{u}_1 = \mathbf{x}_0$ and $\mathbf{u}_0 = \mathbf{x}_1 - \mathbb{E}[\mathbf{x}_1 \mid \mathbf{x}_0]$, where $\begin{bmatrix}\mathbf{x}_0\\ \mathbf{x}_1\end{bmatrix} \sim N_{s+p}(0, M)$. Note that $\mathbb{E}[\mathbf{x}_1 \mid \mathbf{x}_0] = \Delta\Gamma^{-1}\mathbf{x}_0$ and $\mathbf{u}_0 = \mathbf{x}_1 - \Delta\Gamma^{-1}\mathbf{x}_0 \sim N(0, \bar{\Omega} - \Delta\Gamma^{-1}\Delta^T)$. Then we have that $\mathbf{u}_0$ and $\mathbf{u}_1$ are independent. This can be verified by the fact that $\mathbf{u}_0$ and $\mathbf{u}_1$ are normally distributed with covariance $\mathrm{Cov}(\mathbf{u}_0, \mathbf{u}_1) = 0$ which can be verified with algebraic computations. Finally note that $(\mathbf{u}_0 + \Delta\Gamma^{-1}\mathbf{u}_{1,-\boldsymbol{\gamma}})$ is distributed as $(\mathbf{x}_1 \mid \mathbf{x}_0\boldsymbol{\gamma} > 0)$. □

The additive representation introduced above is used in Section 2.4 to draw samples from the distribution.

A.2 Closure properties

The Skew-Normal family has several interesting properties, see Azzalini [2, Ch.7] for details. Most notably, it is closed under marginalization and affine transformations. Specifically, if we 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

$$z_1 \sim SUN_{p_1,s}(\boldsymbol{\xi}_1, \Omega_{11}, \Delta_1, \boldsymbol{\gamma}, \Gamma),$$
$$\text{with } \boldsymbol{\xi} = \begin{bmatrix}\boldsymbol{\xi}_1\\ \boldsymbol{\xi}_2\end{bmatrix}, \ \Delta = \begin{bmatrix}\Delta_1\\ \Delta_2\end{bmatrix}, \ \Omega = \begin{bmatrix}\Omega_{11} & \Omega_{12}\\ \Omega_{21} & \Omega_{22}\end{bmatrix}. \tag{12}$$

Moreover, [2, Ch.7] the conditional distribution is a unified skew-Normal, i.e.,

$$(Z_2|Z_1 = z_1) \sim SUN_{p_2,s}(\boldsymbol{\xi}_{2|1}, \Omega_{2|1}, \Delta_{2|1}, \boldsymbol{\gamma}_{2|1}, \Gamma_{2|1}),$$

where

$$\boldsymbol{\xi}_{2|1} := \boldsymbol{\xi}_2 + \Omega_{21}\Omega_{11}^{-1}(z_1 - \boldsymbol{\xi}_1), \quad \Omega_{2|1} := \Omega_{22} - \Omega_{21}\Omega_{11}^{-1}\Omega_{12},$$
$$\Delta_{2|1} := \Delta_2 - \bar{\Omega}_{21}\bar{\Omega}_{11}^{-1}\Delta_1,$$
$$\boldsymbol{\gamma}_{2|1} := \boldsymbol{\gamma} + \Delta_1^T\Omega_{11}^{-1}(z_1 - \boldsymbol{\xi}_1), \quad \Gamma_{2|1} := \Gamma - \Delta_1^T\bar{\Omega}_{11}^{-1}\Delta_1,$$

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

In section 2.4 we exploit this property to obtain samples from the predictive posterior distribution at a new input $\mathbf{x}^*$ given samples of the posterior at the training inputs.

A.3 Sampling from the posterior predictive distribution

Consider a test point $\mathbf{x}^*$ and assume we have a sample from the posterior distribution $f(X) \mid \mathcal{D}$. Consider the vector $\hat{\mathbf{f}} = [f(X)\, f(\mathbf{x}^*)]^T$, which is distributed as $\mathrm{SUN}_{n+1,s}(\hat{\boldsymbol{\xi}}, \hat{\Omega}, \hat{\Delta}, \boldsymbol{\gamma}, \Gamma)$, where

$$\hat{\boldsymbol{\xi}} = \begin{bmatrix}\xi(X)\\ \xi(\mathbf{x}^*)\end{bmatrix}, \ \hat{\Delta} = \begin{bmatrix}\Delta(X)\\ \Delta(\mathbf{x}^*)\end{bmatrix}, \ \hat{\Omega} = \begin{bmatrix}\Omega(X,X) & \Omega(X,\mathbf{x}^*)\\ \Omega(\mathbf{x}^*,X) & \Omega(\mathbf{x}^*,\mathbf{x}^*)\end{bmatrix}$$

Then by using the marginalization property introduced above we obtain the formula in (5).

B PROOFS OF THE RESULTS IN THE PAPER

Theorem 3.1. This proof is based on the proofs in [9, Th.1 and Co.4]. We aim to derive the posterior of $f(X)$. The joint distribution of $f(X), \mathcal{D}$ is

$$p(\mathcal{D}|f(X))p(f(X)) = \Phi_m(Wf)\ \phi_n(f - \xi; \Omega) \tag{13}$$

where we have omitted the dependence on X for easier notation. We note that

$$\Phi_m(Wf) = \Phi_m\left(W\xi + (\bar{\Omega}D_\Omega W^T)^T\bar{\Omega}^{-1}D_\Omega^{-1}(f - \xi); (W\Omega W^T + I_m) - (W\Omega W^T)\right)$$

Therefore, we can write

$$\begin{aligned}\Phi_m(Wf)\ \phi_n(f - \xi; \Omega) &= \Phi_m\left(W\xi + (\bar{\Omega}D_\Omega W^T)^T\bar{\Omega}^{-1}D_\Omega^{-1}(f - \xi); (W\Omega W^T + I_m) - (W\Omega W^T)\right)\\ &\cdot \phi_n(f - \xi; \Omega)\\ &= \Phi_m(m; M)\phi_n(f - \xi; \Omega)\end{aligned} \tag{14}$$

with

$$m = W\xi + (\bar{\Omega}D_\Omega W^T)^T\bar{\Omega}^{-1}D_\Omega^{-1}(f - \xi)$$

and

$$M = (W\Omega W^T + I_m) - WD_\Omega\bar{\Omega}D_\Omega W^T$$

From (13)–(14) and the definition of the PDF of the SUN distribution, we can easily show that we can rewrite (13) as a SUN distribution with updated parameters:

$$\tilde{\xi} = \xi, \qquad \tilde{\Omega} = \Omega,$$
$$\tilde{\Delta} = \bar{\Omega}D_\Omega W^T,$$
$$\tilde{\gamma} = W\xi, \qquad \tilde{\Gamma} = (W\Omega W^T + I_m).$$

Theorem 3.2. Consider the test point $\mathbf{x} \in \mathbb{R}^d$ and the vector $\hat{f} = \begin{bmatrix} f(X) \\ f(\mathbf{x}) \end{bmatrix} := [\mathbf{f}\ f_*]$ we have

$$p(\mathbf{f}, f_*) = N\left(\begin{pmatrix} \xi(X) \\ \xi(\mathbf{x}) \end{pmatrix}, \begin{pmatrix} \Omega(X,X) & \Omega(X,\mathbf{x}) \\ \Omega(\mathbf{x},X) & \Omega(\mathbf{x},\mathbf{x}) \end{pmatrix}\right)$$

and the predictive distribution is by definition

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

We can then apply Lemma 3.1 with $\hat{f}$ and the likelihood $p([W \mid \mathbf{0}] \mid \hat{f}) = \Phi_{m+1}([W \mid \mathbf{0}] \begin{bmatrix} f(X) \\ f(\mathbf{x}) \end{bmatrix}; I_m)$ which results in a posterior distribution

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

with

$$\begin{aligned} \hat{\boldsymbol{\xi}} &= [\xi(X)\ \xi(\mathbf{x})]^T \\ \hat{\Omega} &= \begin{bmatrix} \Omega(X,X) & \Omega(X,\mathbf{x}) \\ \Omega(\mathbf{x},X) & \Omega(\mathbf{x},\mathbf{x}) \end{bmatrix} \\ \hat{\Delta} &= \begin{bmatrix} \Omega(X,X) & \Omega(X,\mathbf{x}) \\ \Omega(\mathbf{x},X) & \Omega(\mathbf{x},\mathbf{x}) \end{bmatrix} [W \mid \mathbf{0}]^T = \begin{bmatrix} \Omega(X,X)W^T \\ \Omega(\mathbf{x},X)W^T \end{bmatrix} \\ \hat{\gamma} &= [\xi(X)^T\ \xi(\mathbf{x})] \begin{bmatrix} W^T \\ \mathbf{0} \end{bmatrix} = \xi(X)^T W^T \\ \hat{\Gamma} &= [W \mid \mathbf{0}] \begin{bmatrix} \Omega(X,X) & \Omega(X,\mathbf{x}) \\ \Omega(\mathbf{x},X) & \Omega(\mathbf{x},\mathbf{x}) \end{bmatrix} \begin{bmatrix} W^T \\ \mathbf{0} \end{bmatrix} + I_m \\ &= W\Omega(X,X)W^T + I_m \end{aligned}$$

By exploiting the marginalization properties of the SUN distribution, see Section A.2, we obtain

$$\begin{aligned} &p\left(f(\mathbf{x}) \mid W, f(X)\right) \\ &= SUN_{1,m}\left(\xi(\mathbf{x}), \Omega(\mathbf{x},\mathbf{x}), \Omega(\mathbf{x},X)W^T, \xi(X)^T W^T, W\Omega(X,X)W^T + I_m\right). \end{aligned} \tag{15}$$

Corollary 3.3. We can write the likelihood function as

$$p(\mathcal{D} \mid f(\mathcal{X})) = \Phi_m(Uf(\mathcal{X}) - Vf(\mathcal{X}); I_m),$$

where $V \in \mathbb{R}^{m \times n}$ with $V_{i,j} = 1$ if $v_i = x_j$ and 0 otherwise and $U \in \mathbb{R}^{m \times n}$ with $U_{i,j} = 1$ if $u_i = x_j$ and 0 otherwise. Then we can apply Lemma 3.1 for the posterior distribution of $f(X)$ and Theorem 3.2 for the posterior distribution of f at an unobserved point.

Proposition 3.4. As described in Section A.1, if we consider a random vector $\begin{bmatrix} \mathbf{x}_0 \\ \mathbf{x}_1 \end{bmatrix} \sim N_{s+p}(0, M)$ with $M = \left[\begin{smallmatrix} \Gamma & \Delta \\ \Delta^T & \Omega \end{smallmatrix}\right]$ and define $\mathbf{y}$ as the vector with distribution $(\mathbf{x}_1 \mid \mathbf{x}_0 + \boldsymbol{\gamma} > 0)$, then it can be shown [2, Ch. 7] that $\mathbf{z} = \boldsymbol{\xi} + D_\Omega \mathbf{y} \sim \mathrm{SUN}_{p,s}(\boldsymbol{\xi}, \Omega, \Delta, \boldsymbol{\gamma}, \Gamma)$. This allows one to derive the following sampling scheme:

$$\begin{gathered} f \sim \boldsymbol{\xi} + D_\Omega \left(U_0 + \Delta\Gamma^{-1} U_1\right), \qquad (16) \\ U_0 \sim \mathcal{N}(0; \bar{\Omega} - \Delta\Gamma^{-1}\Delta^T), \qquad U_1 \sim \mathcal{T}_{\boldsymbol{\gamma}}(0; \Gamma), \end{gathered}$$

where $\mathcal{T}_{\boldsymbol{\gamma}}(0; \Gamma)$ is the pdf of a multivariate Gaussian distribution truncated component-wise below $-\boldsymbol{\gamma}$. Then from the marginal (15) and the above sampling scheme (see Section A.3), we obtain the formulae in (8), main text.

Corollary 3.5. This follows from the Theorem 3.2: $\Phi_m(\tilde{\boldsymbol{\gamma}}, \tilde{\Gamma})$ is the normalization constant of the posterior and, therefore, the marginal likelihood is $\Phi_m(\tilde{\boldsymbol{\gamma}}, \tilde{\Gamma})$. The lower bound was proven in [4, Prop.2]

C IMPLEMENTATION

C.1 Laplace’s approximation

The Laplace’s approximation for preference learning was implemented as described in [8]. We use standard Bayesian optimisation to optimise the hyper-parameters of the kernel by maximising the Laplace’s approximation of the marginal likelihood.

C.2 Skew Gaussian Process

To compute $\Phi_{|B_i|}(\cdot)$ in (9), we use the routine proposed in [23], that computes multivariate normal probabilities using bivariate conditioning. This is very fast. We optimise the hyper-parameters of the kernel by maximising the lower bound in (9) and we use simulated annealing.

C.3 Acquisition function optimisation

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 $\mathbf{x}_r$ 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.

For both the models (Laplace’s approximation (GPL) and SkewGP) we compute the acquisition functions $\alpha(\mathbf{x}, \mathbf{x}_r)$ via Monte Carlo sampling from the posterior. In fact, although for GPL some analytical formulas are available (for instance for UCB), by using random sampling (2000 samples with fixed seed) for both GPL and SkewGP removes any advantage of SkewGP over GPL due to the additional exploration effect of Monte Carlo sampling. Computing $\alpha(\mathbf{x}, \mathbf{x}_r)$ in this way is very fast for both the models (for SkewGP this is due to lin-ess). We then optimize $\alpha(\mathbf{x}, \mathbf{x}_r)$: (i) by computing $\alpha(\mathbf{x}, \mathbf{x}_r)$ for 5000 random generated value of $\mathbf{x}$ (every time we use the same random points for both SkewGP and GPL); (ii) the best random instance is then used as initial guess for L-BFGS-B.