# Learning GPLVM with arbitrary kernels using the unscented transformation: Supplementary Materials

## A THE BAYESIAN GPLVM EVIDENCE LOWER BOUND

The unscented transformation and the competing approximations are applied to terms in the evidence lower bound of the Bayesian GPLVM model derived in Titsias and Lawrence (2010). This section will present a definition of the Bayesian GPLVM, the variational distributions, the evidence lower bound, and where each approximation is applied.

### A.1 Prior and variational distributions

The model is defined as:

$$\begin{aligned}
\boldsymbol{Z} &\in \mathbb{R}^{m \times D_x} \\
p(\boldsymbol{X}) &= \prod_{i=1}^{N} \mathcal{N}(\boldsymbol{x}_i \mid \boldsymbol{0}, \boldsymbol{I}), \\
p(\boldsymbol{U} \mid \boldsymbol{Z}) &= \prod_{d=1}^{D_y} \mathcal{N}(\boldsymbol{u}_{:d} \mid \boldsymbol{0}, \boldsymbol{K}_u), \\
p(\boldsymbol{F} \mid \boldsymbol{X}, \boldsymbol{U}, \boldsymbol{Z}) &= \prod_{d=1}^{D_y} \mathcal{N}\Big(\boldsymbol{f}_{:d} \mid \boldsymbol{K}_{fu}\boldsymbol{K}_u^{-1}\boldsymbol{u}_{:d}, \boldsymbol{K}_f - \boldsymbol{K}_{fu}\boldsymbol{K}_u^{-1}\boldsymbol{K}_{fu}^{\intercal}\Big), \\
p(\boldsymbol{Y} \mid \boldsymbol{F}) &= \prod_{d=1}^{D_y} \mathcal{N}\big(\boldsymbol{y}_{:d} \mid \boldsymbol{f}_{:d}, \sigma^2\boldsymbol{I}\big), \\
&\text{where:} \\
[\boldsymbol{K}_f]_{ij} &= k(\boldsymbol{x}_i, \boldsymbol{x}_j), \\
[\boldsymbol{K}_u]_{ij} &= k(\boldsymbol{z}_i, \boldsymbol{z}_j), \\
[\boldsymbol{K}_{fu}]_{ij} &= k(\boldsymbol{x}_i, \boldsymbol{z}_j).
\end{aligned}$$

The approximate posterior is defined as:

$$\begin{aligned}
q(\boldsymbol{F}, \boldsymbol{U}, \boldsymbol{X}) &= p(\boldsymbol{F} \mid \boldsymbol{U}, \boldsymbol{X}) q(\boldsymbol{U}) q(\boldsymbol{X}), \\
&\text{where:} \\
q(\boldsymbol{X}) &= \prod_{i=1}^{N} \mathcal{N}\big(\boldsymbol{x}_i \mid \check{\boldsymbol{\mu}}_i, \mathrm{Diag}(\check{\boldsymbol{\sigma}}_i^2)\big), \\
&\mathrm{Diag}(\boldsymbol{x}) \text{ maps a vector into a corresponding diagonal matrix,} \\
q(\boldsymbol{U}) &= \prod_{d=1}^{D_y} \mathcal{N}\Big(\boldsymbol{u}_{:d} \mid \boldsymbol{K}_u\big(\sigma^2\boldsymbol{K}_u + \boldsymbol{\Psi}_2\big)^{-1}\boldsymbol{\Psi}_1^{\intercal}\boldsymbol{y}_{:d},\ \sigma^2\boldsymbol{K}_u\big(\sigma^2\boldsymbol{K}_u + \boldsymbol{\Psi}_2\big)^{-1}\boldsymbol{K}_u\Big).
\end{aligned}$$

Therefore, the hyperparameters and variational parameters of this model are the output noise variance $\sigma^2$, the kernel hyperparameters, the pseudo-inputs $\boldsymbol{Z}$, and the variational parameters $\check{\boldsymbol{\mu}}_i$ and $\check{\boldsymbol{\sigma}}_i^2$ for each $i$ from 1 to $n$.

### A.2 Evidence lower bound

The parameters are learned through joint-optimization of the evidence lower bound (ELBO). The ELBO obtained in Titsias and Lawrence (2010) is:

$$\ln p(\boldsymbol{Y}) \geq \frac{1}{2}\sum_{d=1}^{D_y}\left[\ln|\boldsymbol{K}_u| - n\ln\left(2\pi\sigma^2\right) - \ln|\boldsymbol{W}| - \frac{\boldsymbol{y}_{:d}^\intercal \boldsymbol{y}_{:d}}{\sigma^2} + \frac{\boldsymbol{y}_{:d}^\intercal \boldsymbol{\Psi}_1 \boldsymbol{W}^{-1}\boldsymbol{\Psi}_1^\intercal \boldsymbol{y}_{:d}}{\sigma^2} - \frac{\psi_0}{\sigma^2} + \frac{\mathrm{tr}\left(\boldsymbol{K}_u^{-1}\boldsymbol{\Psi}_2\right)}{\sigma^2}\right], \quad (1)$$

where:

$$\begin{aligned}
\boldsymbol{W} &= \sigma^2\boldsymbol{K}_u + \boldsymbol{\Psi}_2,\\
\psi_0 &= \sum_{i=1}^{N}\left\langle k(\boldsymbol{x}_i,\boldsymbol{x}_i)\right\rangle_{q(\boldsymbol{x}_i)},\\
[\boldsymbol{\Psi}_1]_{ij} &= \left\langle k(\boldsymbol{x}_i,\boldsymbol{z}_j)\right\rangle_{q(\boldsymbol{x}_i)},\\
[\boldsymbol{\Psi}_2]_{jm} &= \sum_{i=1}^{N}\left\langle k(\boldsymbol{x}_i,\boldsymbol{z}_j)k(\boldsymbol{x}_i,\boldsymbol{z}_m)\right\rangle_{q(\boldsymbol{x}_i)}.
\end{aligned}$$

The Kullback–Leibler divergence of the approximate posterior with the true posterior is minimized by maximizing the ELBO. Under certain circumstances, when an RBF kernel is used in this model, the Ψ-statistics can be computed analytically. However, for arbitrary kernels, we must apply approximations to compute the expected values.

## B UNSCENTED TRANSFORMATION, GAUSS-HERMITE QUADRATURE, AND MONTE CARLO INTEGRATION IN CONTEXT

The only terms that can not be computed analytically on the ELBO are the Ψ-statistics. This section presents how each of the discussed approximation methods can compute an estimate of these terms.

### B.1 Unscented transformation

Starting with the unscented transformation, each term is approximated as:

$$\begin{aligned}
\psi_0 &\approx \frac{1}{2D_x}\sum_{i=1}^{N}\sum_{t=1}^{2D_x} k(\boldsymbol{s}_i^{(t)},\boldsymbol{s}_i^{(t)}),\\
[\boldsymbol{\Psi}_1]_{ij} &\approx \frac{1}{2D_x}\sum_{t=1}^{2D_x} k(\boldsymbol{s}_i^{(t)},\boldsymbol{z}_j),\\
[\boldsymbol{\Psi}_2]_{jm} &\approx \frac{1}{2D_x}\sum_{i=1}^{N}\sum_{t=1}^{2D_x} k(\boldsymbol{s}_i^{(t)},\boldsymbol{z}_j)k(\boldsymbol{s}_i^{(t)},\boldsymbol{z}_m),
\end{aligned}$$

where:

$$\boldsymbol{s}_i^{(t)} = \begin{cases}\check{\boldsymbol{\mu}}_i + \mathrm{Diag}\left(\sqrt{D_x}\check{\boldsymbol{\sigma}}_i\right)_{:t} \text{ if } t \in [1, D_x],\\ \check{\boldsymbol{\mu}}_i - \mathrm{Diag}\left(\sqrt{D_x}\check{\boldsymbol{\sigma}}_i\right)_{:t} \text{ otherwise.}\end{cases}.$$

These expressions are exactly the empirical mean of the distribution induced by the transformed sigma points.

### B.2 Gauss-Hermite quadrature

A simple extension of Gauss-Hermite quadrature can be used to estimate expected values of mappings of multidimensional Gaussian distributed variables. First the number $H$ of evaluation points on each dimension is chosen, then the base $H$ evaluation points and weights are computed from the roots of the Hermite polynomial

$\mathcal{H}_H(x)$. These base evaluation points and weights are organized into a $H^{D_x}$ dimensional grid $\boldsymbol{r}_t$ and $w_t$. Finally, the approximation is computed as:

$$\begin{aligned}
\psi_0 &\approx \sum_{i=1}^{N}\sum_{t=1}^{H^{D_x}} w_t \cdot k(\boldsymbol{s}_i^{(t)}, \boldsymbol{s}_i^{(t)}),\\
[\boldsymbol{\Psi}_1]_{ij} &\approx \sum_{t=1}^{H^{D_x}} w_t \cdot k(\boldsymbol{s}_i^{(t)}, \boldsymbol{z}_j),\\
[\boldsymbol{\Psi}_2]_{jm} &\approx \sum_{i=1}^{N}\sum_{t=1}^{H^{D_x}} w_t \cdot k(\boldsymbol{s}_i^{(t)}, \boldsymbol{z}_j) k(\boldsymbol{s}_i^{(t)}, \boldsymbol{z}_m),
\end{aligned}$$

where:

$$\boldsymbol{s}_i^{(t)} = \check{\boldsymbol{\mu}}_i + \mathrm{Diag}\Big(\sqrt{2}\check{\boldsymbol{\sigma}}_i\Big)\boldsymbol{r}_t.$$

As an concrete example, suppose that $H = 2$ and $D_x = 2$, then, the Hermite polynomial is $\mathcal{H}_2(x) = 4x^2 - 2$, its roots are $\pm\frac{\sqrt{2}}{2}$ with weights $\frac{\sqrt{\pi}}{2}$. Thus, $\boldsymbol{r}_t$ and $w_t$ are:

$$\begin{aligned}
\boldsymbol{r}_1 &= \left[-\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}\right], & w_1 &= \frac{\sqrt{\pi}}{2}\cdot\frac{\sqrt{\pi}}{2},\\
\boldsymbol{r}_2 &= \left[-\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right], & w_2 &= \frac{\sqrt{\pi}}{2}\cdot\frac{\sqrt{\pi}}{2},\\
\boldsymbol{r}_3 &= \left[\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}\right], & w_3 &= \frac{\sqrt{\pi}}{2}\cdot\frac{\sqrt{\pi}}{2},\\
\boldsymbol{r}_4 &= \left[\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right], & w_4 &= \frac{\sqrt{\pi}}{2}\cdot\frac{\sqrt{\pi}}{2}.
\end{aligned}$$

### B.3 Monte Carlo integration

Finally, we have Monte Carlo integration. First a number of samples $T$ is chosen. Then, the expected values are computed by the empirical mean of each trasformed sample:

$$\begin{aligned}
\psi_0 &\approx \frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T} k(\boldsymbol{s}_i^{(t)}, \boldsymbol{s}_i^{(t)}),\\
[\boldsymbol{\Psi}_1]_{ij} &\approx \frac{1}{T}\sum_{t=1}^{T} k(\boldsymbol{s}_i^{(t)}, \boldsymbol{z}_j),\\
[\boldsymbol{\Psi}_2]_{jm} &\approx \frac{1}{T}\sum_{i=1}^{N}\sum_{t=1}^{T} k(\boldsymbol{s}_i^{(t)}, \boldsymbol{z}_j) k(\boldsymbol{s}_i^{(t)}, \boldsymbol{z}_m),
\end{aligned}$$

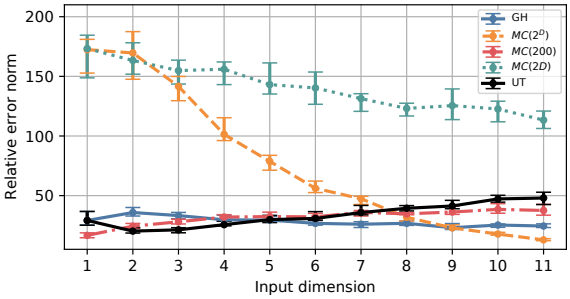
where:

$$\boldsymbol{s}_i^{(t)} = \check{\boldsymbol{\mu}}_i + \mathrm{Diag}(\check{\boldsymbol{\sigma}}_i)\boldsymbol{r}_i^{(t)},$$

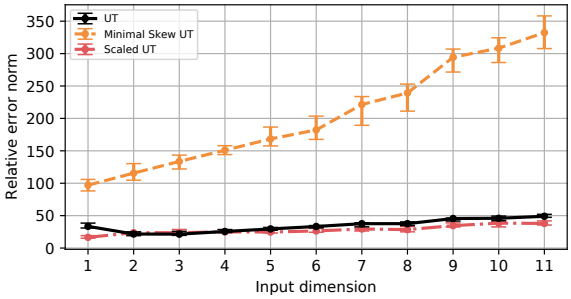
$\boldsymbol{r}_i^{(t)}$was sampled from a standard normal distribution.

## C PERFORMANCE COMPARISON OF APPROXIMATION METHODS

Alongside the timing experiments from Section 3.3, we also conducted an approximation performance experiment. This experiment's goal is to evaluate how well each approximation evaluates a single $\Psi$-statistic and how alternatives of the simple symmetric UT might compare to it.




(a) $L_1$-norm of the differences between approximations and an MC(10000) estimate. UT maintains a steady accuracy as dimensions increase. Other methods with comparable performance require many more samples.




(b) Uniform UT versus other selections of sigma points and weights. Best parameters for scaled and minimal skew UT were chosen through grid-search and displayed here. Note that our hyper-parameterless UT has a comparable accuracy to best selected scaled UT.

Figure 1: Results of the approximation performance experiment

The tested approximations are the unscented transformation presented in the main paper, Gauss-Hermite with 2 points on each grid for a total of $2^D$ points, Monte Carlo integration with $2D$, $2^D$ and 200 samples, and finally, the minimal skew UT and scaled UT as defined by Menegaz et al. (2015).

First, we computed $\mathbf{\Psi}_1$ for a Matérn 3/2 kernel using each approximation on random data ($N = 40, M = 20$) of varying dimensions. Then, we compare the results with an MC approximation that uses 10,000 samples as a proxy for the real value.

In Figure 1a, despite the increasing dimensionality, the UT performs on par with other methods that use many more samples, in contrast with the only method that uses the same amount of samples as it, MC($2D$). As for the comparison between UTs in Figure 1b, our choice of sigma-points outperforms minimal skew UT and does as well as scaled UT. It is worth highlighting that, contrary to its counterparts, this method is hyper-parameter-free.