Monotonic Gaussian Process Flows: Supplementary material

Ivan Ustyuzhaninov*
University of Tübingen

Ieva Kazlauskaite*
University of Bath,
Electronic Arts

Carl Henrik Ek
University of Bristol

Neill D. F. Campbell
University of Bath,
Royal Society

Supplementary material

A.1 Numerical solution of the SDE

To ensure that the SDE solutions are monotonic functions of the initial values, we make assumptions about the Wiener process realisations $W(\cdot, \omega)$. To compute the SDE solutions under such assumptions, we draw a Wiener process realisation as well as the flow field drift and diffusion, and given these draws, we use the Euler-Maryama numerical solver (following [Hegde et al., 2019]). Specifically, starting with the initial state $(x = x_1, t = 0), \dots, (x = x_N, t = 0)$, we use (2) to compute the drift and diffusion at the current state, and the discretised version of (1) (*i.e.* with Δt and ΔW instead of dt and dW) to compute the state update Δx. This gives the new state $(x_1 + \Delta x_1, \Delta t), ..., (x_n + \Delta x_n, \Delta t)$, and repeating this procedure $(T/\Delta t)$ times, we arrive at the state $(S(T, \omega; x_1), T), ..., (S(T, \omega; x_N), T)$, corresponding to the approximate SDE solution at time T. The monotonic trajectories are recovered by the numerical SDE solver in the limit of the step size going to zero, $\Delta t \to 0$. Therefore, the step size must be sufficiently small w.r.t. the smoothness of the flow; since we use a GP to define the flow, the smoothness is determined by the lengthscale of the kernel.

A.2 Implementation details

Our model is implemented in Tensorflow [Abadi et al., 2015]. For the evaluations in Tables 1 and 2 we use 10000 iterations with the learning rate of 0.01 that gets reduced by a factor of $\sqrt{10}$ when the objective does not improve for more than 500 iterations. For numerical solutions of SDE, we use Euler-Maruyama solver with 20 time steps, as proposed in [Hegde et al., 2019].

A.3 Computational complexity

Computational complexity of drawing a sample from the monotonic flow model is $\mathcal{O}\big(N_{\text{steps}}(NM^2 + N^2)\big)$, where N_{steps} is the number of steps in numerical computation of the approximate SDE solution, NM^2 is the complexity of computing the GP posterior for N inputs based on M inducing points, and N^2 is the complexity of drawing a sample from this posterior. We typically draw fewer than 5 samples to limit the computational cost.

A.4 Non-Gaussian noise

The inference procedures for the monotonic flow and for the 2-layer model can be easily applied to arbitrary likelihoods, because they are based on stochastic variational inference and do not require the closed form integrals of the likelihood density.

A.5 Functions for evaluating the monotonic flow model

The functions we use for evaluations are the following:

$f_1(x) = 3,\ x \in (0, 10]$ (flat function)

$f_2(x) = 0.32\,(x + \sin(x)),\ x \in (0, 10]$ (sinusoidal function)

$f_3(x) = 3$ if $x \in (0, 8],\ f_3(x) = 6$ if $x \in (8, 10]$ (step function)

$f_4(x) = 0.3x,\ x \in (0, 10]$ (linear function)

$f_5(x) = 0.15\ \exp(0.6x - 3),\ x \in (0, 10]$ (exponential function)

$f_6(x) = 3\ /\ [1 + \exp(-2x + 10)],\ x \in (0, 10]$ (logistic function)

For the experiments shown in Fig. 3 we generate 50 data points using $y = \text{sinc}(\pi x) + \varepsilon, \varepsilon \sim \mathcal{N}(0, 0.02)$ for linearly spaced inputs $x \in [-1, 1]$.

A.6 Regression evaluation parameters

For the GP with monotonicity information we choose M virtual points and place them equidistantly in the range of the data; we provide the best RMSEs for

$M \in [10, 20, 50, 100]$. For the transformed GP we report the best results for the boundary conditions $L \in [10, 15, 20, 30]$ and the number of terms in the approximation $J \in [2, 3, 5, 10, 15, 20, 25, 30]$. For both models we use a squared exponential kernel. Our method depends on the time T, the kernel and the number of inducing points M. For this experiment, we consider $T \in [1, 5]$, $M = 40$ and two kernel options, squared exponential and ARD Matérn 3/2. The lowest RMSE are achieved using the flow and the transformed GP.

A.7 Uncertainty in alignment model

To further illustrate the advantages of capturing the uncertainty about the warpings, we wish to find the possibly bi-modal warpings for each sequence. We use a Gaussian mixture model (instead of a single Gaussian) as the distribution of both, the warpings and the latent variables $\mathbf{Z}$ in the GP-LVM. In particular, the inducing points of the flow for each sequence are defined to be distributed as a mixture of two multivariate Gaussians. Then, given a draw from the Categorical distribution of this mixture, we defines the clusters assignments for each sample, and assign the resulting aligned sequences $\mathbf{s}_j$ to one of the coherent mixture component in the distribution of the latent points of the GP-LVM. Fig. A1 illustrates this behaviour, and gives an example where the uncertainty in the warps results in ambiguity in cluster assignments. A full discussion of the importance of correlations in the variational parameters for compositional uncertainty is available in [Ustyuzhaninov et al., 2019] which provides further details of the inference scheme used.

A.8 Quantitative results

The expected log posterior predictive density is an evaluation metric defined as:

$$\begin{aligned} \text{ELPD} &= \log \int p(y^* \mid f^*) p(f^* \mid \mathbf{y}) \mathrm{d}f^* \\ &\approx \log \int p(y^* \mid f^*) q(f^* \mid \mathbf{y}) \mathrm{d}f^*. \end{aligned} \tag{1}$$

The results on the data described in Sec. 5 (with 100 data points) for the GP with derivatives [Riihimäki and Vehtari, 2010], the transformed GP [Andersen et al., 2018] and the monotonic flow are given in Table 3.

References

[Abadi et al., 2015] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Software available from tensorflow.org.

[Andersen et al., 2018] Andersen, M. R., Siivola, E., Riutort-Mayol, G., and Vehtari, A. (2018). A nonparametric probabilistic model for monotonic functions. *“All Of Bayesian Nonparametrics” Workshop at NeurIPS.*

[Hegde et al., 2019] Hegde, P., Heinonen, M., Lähdesmäki, H., and Kaski, S. (2019). Deep learning with differential gaussian process flows. In *International Conference on Artificial Intelligence and Statistics (AISTATS).*

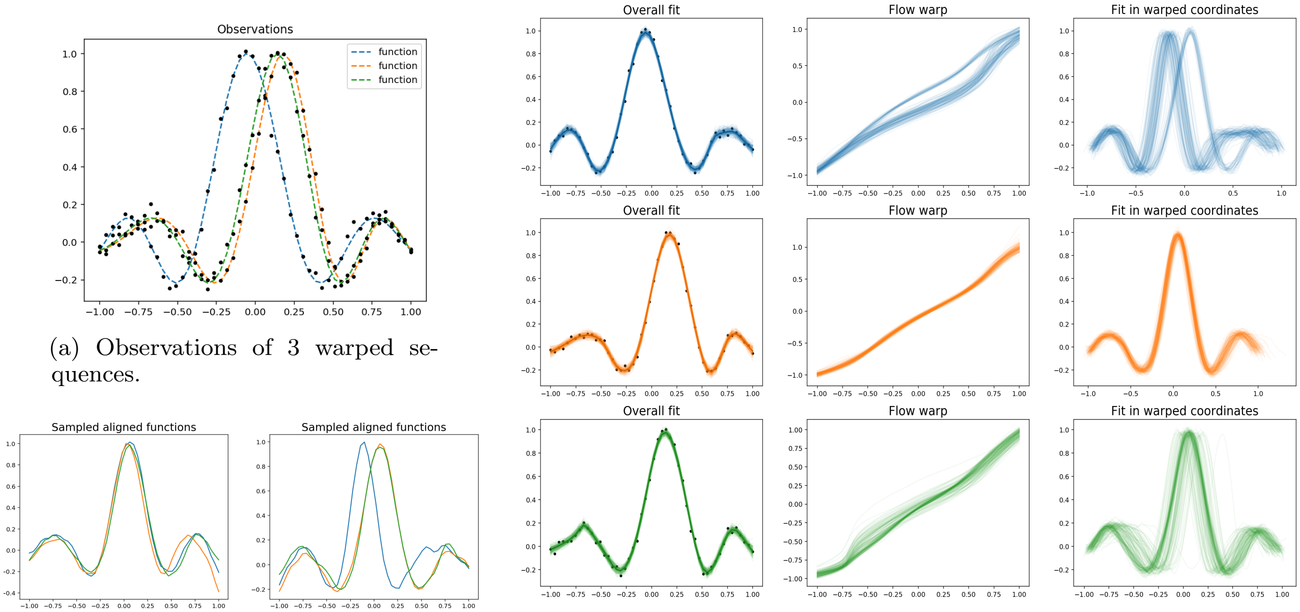
[Lin and Dunson, 2014] Lin, L. and Dunson, D. (2014). Bayesian monotone regression using gaussian process projection. *Biometrika*, 101(2):303–317.

[Maatouk, 2017] Maatouk, H. (2017). Finite-dimensional approximation of gaussian processes with inequality constraints. *arXiv:1706.02178.*

[Riihimäki and Vehtari, 2010] Riihimäki, J. and Vehtari, A. (2010). Gaussian processes with monotonicity information. *International Conference on Artificial Intelligence and Statistics (AISTATS).*

[Shively et al., 2009] Shively, T. S., Sager, T. W., and Walker, S. G. (2009). A bayesian approach to nonparametric monotone function estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(1):159–175.

[Ustyuzhaninov et al., 2019] Ustyuzhaninov, I., Kazlauskaite, I., Kaiser, M., Bodin, E., Campbell, N. D. F., and Ek, C. H. (2019). Compositional uncertainty in deep gaussian processes. In *Bayesian Deep Learning Workshop at NeurIPS.*



(a) Observations of 3 warped sequences.

(b) Examples of sampled aligned functions **s**.

(c) Fitted sequences (left), estimated warps (middle) and fits in the warped coordinates (right) for the 3 sequences.

Figure A1: Illustration of uncertainty in warps and cluster assignments. When the warps and the cluster assignment are allowed to be bi-modal, and model captures two possible solutions, one that assigns all sequences to a single cluster and aligns them within the cluster, and another solution that favours the model with two separate clusters. This can be see in the fit in warped coordinates figure for the blue curve where the majority of the samples are assigned to one cluster (which corresponds to now aligning the blue function to the other, as shown on the right in Fig. A1b) while a small subset is assigned to a new cluster (which corresponds to all sequences being aligned together, as shown on the left in Fig. A1b).

	flat	sinusoidal	step	linear	exponential	logistic
GP	15.1	21.9	27.1	16.7	19.7	25.5
GP projection [Lin and Dunson, 2014]	11.3	21.1	25.3	16.3	19.1	22.4
Regression splines [Shively et al., 2009]	9.7	22.9	28.5	24.0	21.3	19.4
GP approximation [Maatouk, 2017]	8.2	20.6	41.1	15.8	20.8	21.0
GP with derivatives [Riihimäki and Vehtari, 2010]	16.5 ± 5.1	19.9 ± 2.9	68.6 ± 5.5	16.3 ± 7.6	27.4 ± 6.5	30.1 ± 5.7
Transformed GP [Andersen et al., 2018] (VI-full)	**6.4** ± 4.5	20.6 ± 5.9	52.5 ± 3.6	**11.6** ± 5.8	17.5 ± 7.3	**17.1** ± 6.2
Monotonic Flow (ours)	6.8 ± 3.2	**17.9** ± 4.2	**20.5** ± 5.0	13.2 ± 6.7	**14.4** ± 4.8	18.1 ± 5.0

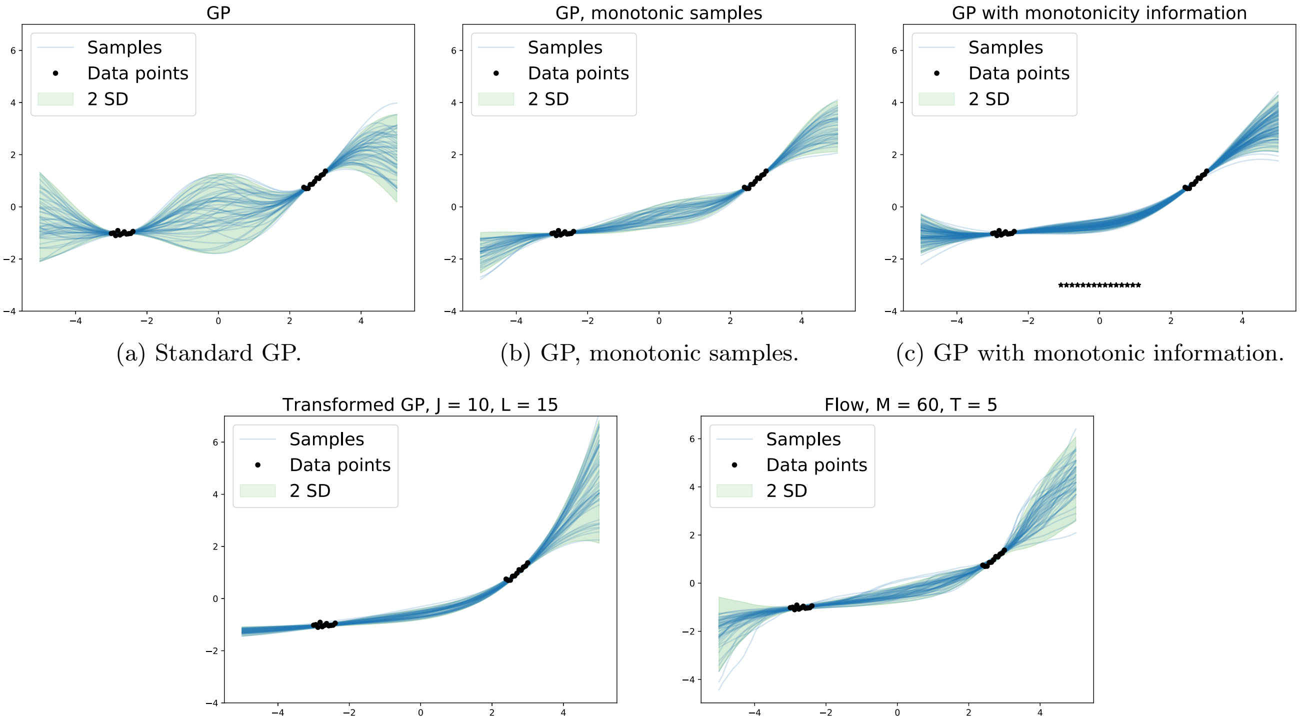
Table 1: Root-mean-square error ± SD (×100) of 20 trials for data of size $N = 100$

	flat	sinusoidal	step	linear	exponential	logistic
Transformed GP [Andersen et al., 2018] (VI-full)	**18.5** ± 14.4	40.0 ± 17.5	101.9 ± 11.4	37.4 ± 22.8	52.9 ± 11.9	51.7 ± 19.6
Monotonic Flow (ours)	21.7 ± 15.0	**39.1** ± 13.0	**64.5** ± 10.7	**30.8** ± 12.0	**32.8** ± 17.9	**43.2** ± 15.2

Table 2: Root-mean-square error ± SD (×100) of 20 trials for data of size $N = 15$

	flat	sinusoidal	step	linear	exponential	logistic
GP with derivatives [Riihimäki and Vehtari, 2010]	-1.43 ± 0.08	-1.41 ± 0.06	-1.69 ± 0.15	-1.36 ± 0.04	-1.45 ± 0.08	-1.45 ± 0.11
Transformed GP [Andersen et al., 2018] (VI-full)	-1.44 ± 0.03	-1.39 ± 0.02	-1.51 ± 0.06	-1.40 ± 0.03	-1.41 ± 0.02	-1.41 ± 0.02
Monotonic Flow (ours)	-1.39 ± 0.05	-1.42 ± 0.05	-1.41 ± 0.08	-1.39 ± 0.05	-1.40 ± 0.07	-1.43 ± 0.07

Table 3: Expected log posterior predictive density estimate (± SD) of 20 trials for data of size $N = 100$



(a) Standard GP.

(b) GP, monotonic samples.

(c) GP with monotonic information.

(d) Transformed GP (VI-full).

(e) Flow (ours).

Figure A2: Comparison of the confidence intervals for standard GP, and monotonic regression methods (GP with monotonic information from [Riihimäki and Vehtari, 2010] and Transformed GP from [Andersen et al., 2018]). The samples from the fitted models are shown in blue and the 2 standard deviations from the mean are shown in green.