Deep Learning with differential Gaussian process flows

Aalto University
School of Science

Pashupati Hegde Markus Heinonen Harri Lähdesmäki Samuel Kaski

Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, Finland

contact: firstname.lastname@aalto.fi code: github.com/hegdepashupati/differential-dgp

HIIT

1. Motivation

- Modern deep learning methods involve discrete sequence of transformations including DNNs, state-space models among others.
- We propose **a paradigm of continuous-time learning** with probabilistic non-linear transformations using SDEs.
- The proposed model is an approximation to **infinitely deep Gaussian process with infinitesimal increments**.

2. Model

(1)

We warp observed inputs $\mathbf{X}$ through a stochastic differential system defined by

$$d\mathbf{x}_t = \boldsymbol{\mu}(\mathbf{x}_t)dt + \sqrt{\boldsymbol{\Sigma}(\mathbf{x}_t)}dW_t,$$

where $\boldsymbol{\mu}(\mathbf{x}_t)$ and $\boldsymbol{\Sigma}(\mathbf{x}_t)$ are the mean and covariance functions of GP prior on the differential function $\mathbf{f}$

$$\mathbf{f}(\mathbf{x}) \sim \mathcal{GP}(\mathbf{0}, K(\mathbf{x}, \mathbf{x}'))$$

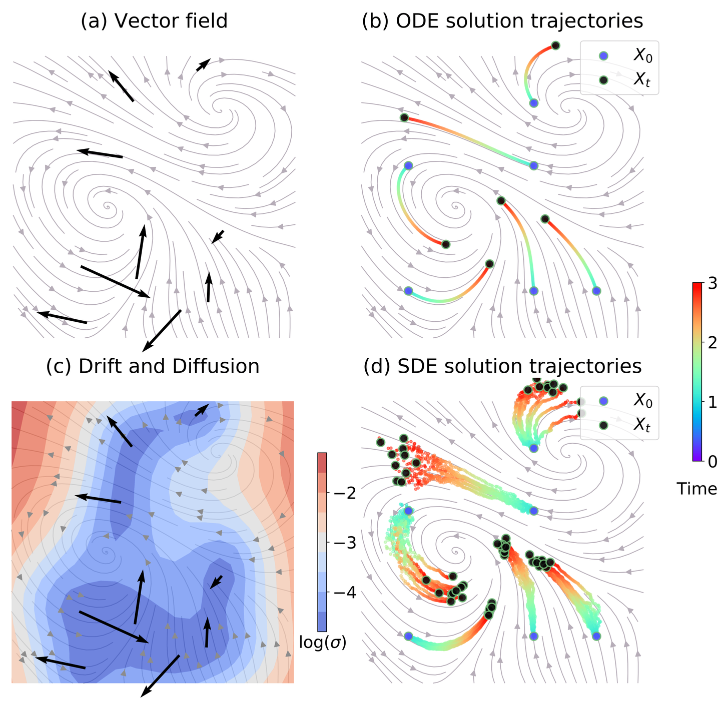
$$\mathbf{f}(\mathbf{x})|\mathbf{U_f}, \mathbf{Z_f} \sim \mathcal{N}(\boldsymbol{\mu}(\mathbf{x}), \boldsymbol{\Sigma}(\mathbf{x})).$$

(2)

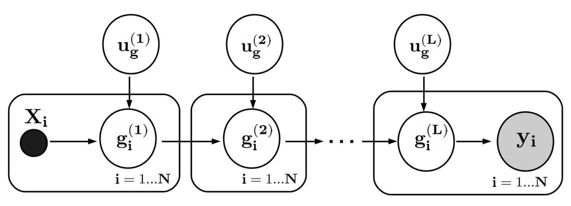
We then classify or regress the final data points $\mathbf{X}_T$ after T time of an SDE flow with a predictor Gaussian process

$$g(\mathbf{x}_T) \sim \mathcal{GP}(0, K(\mathbf{x}_T, \mathbf{x}_T'')).$$

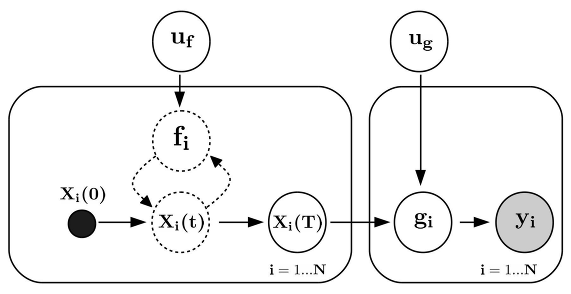
The framework reduces to a conventional Gaussian process with zero flow time $T = 0$.



3. Inference



(a) Deep GP



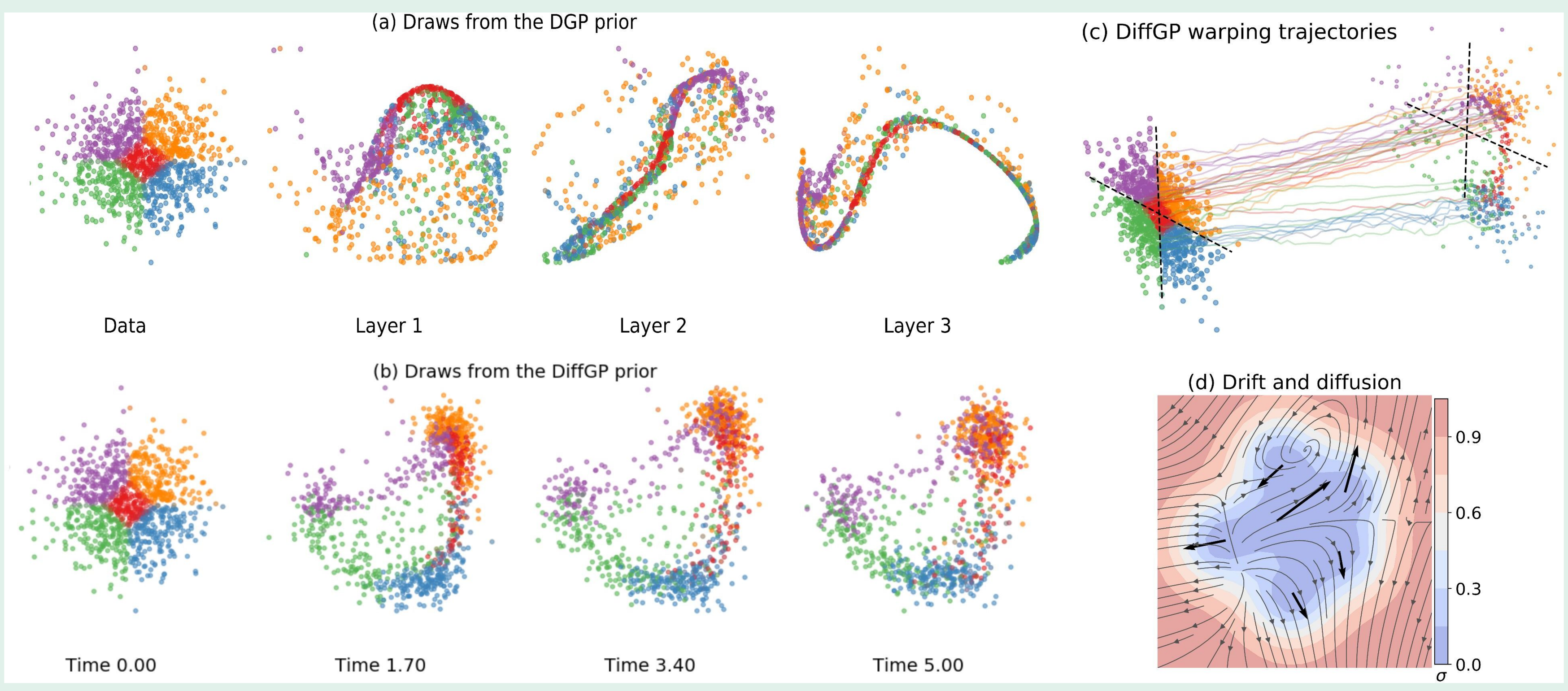
(b) DiffGP (our method)

- We follow the SVI framework for GPs [1]
- Model is fully parameterized by two sets of inducing points for $\mathbf{f}(\cdot)$ and $g(\cdot)$ respectively, as well as, kernel and likelihood parameters.
- We integrate out the state distributions using Euler-Maruyma solver for the posterior SDE

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \boldsymbol{\mu}_q(\mathbf{x}_k)\Delta t + \sqrt{\boldsymbol{\Sigma}_q(\mathbf{x}_k)}\Delta W_k,$$

where, drift $\boldsymbol{\mu}_q$ and diffusion $\boldsymbol{\Sigma}_q$ are defined by the posterior parameters of latent process $\mathbf{f}$.

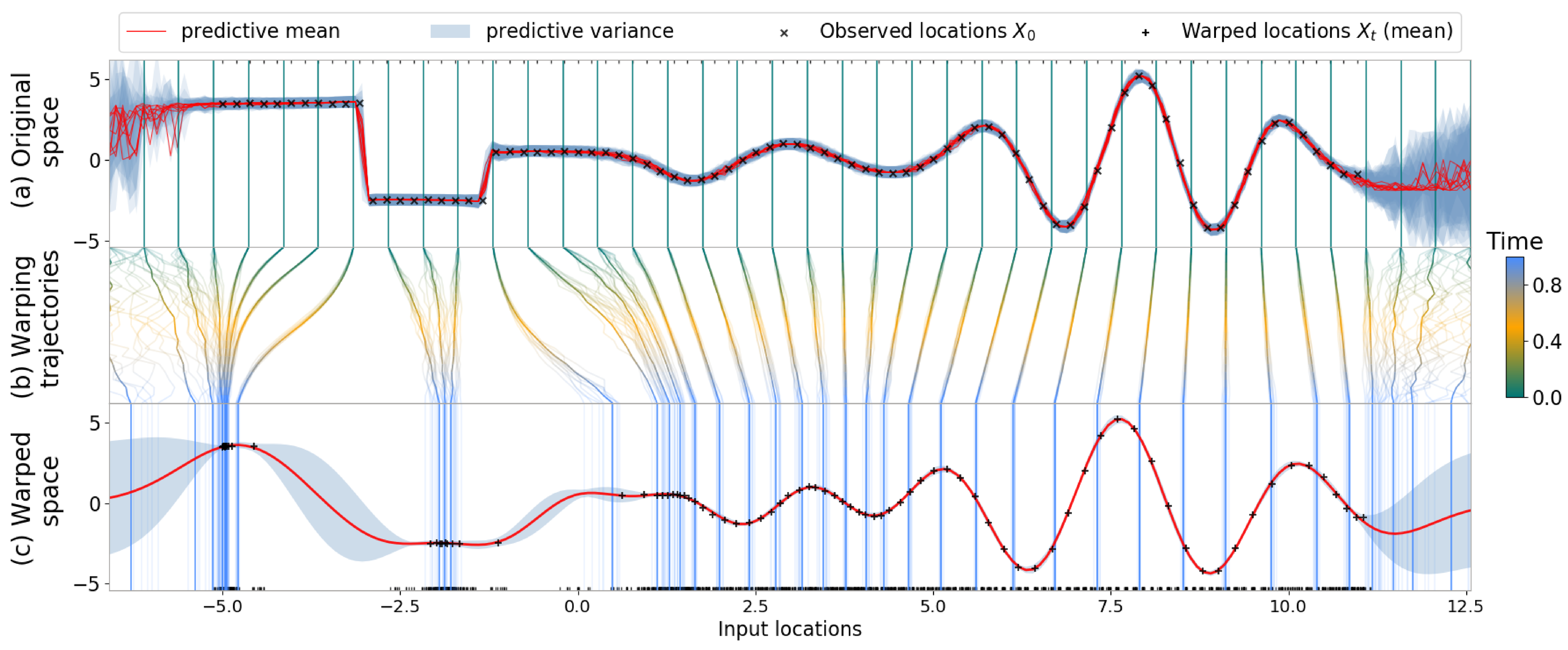
4. Characterizing DiffGP prior



(a) Samples from a 2D deep GP prior exhibit a pathology wherein representations in deeper layers concentrate on low-rank manifolds. **(b)** Samples from a diffGP prior result in rank-preserving representations. **(c)** Continuous trajectories are formed with smooth drift and structured diffusion **(d)**.

5. Experiments

Step function estimation



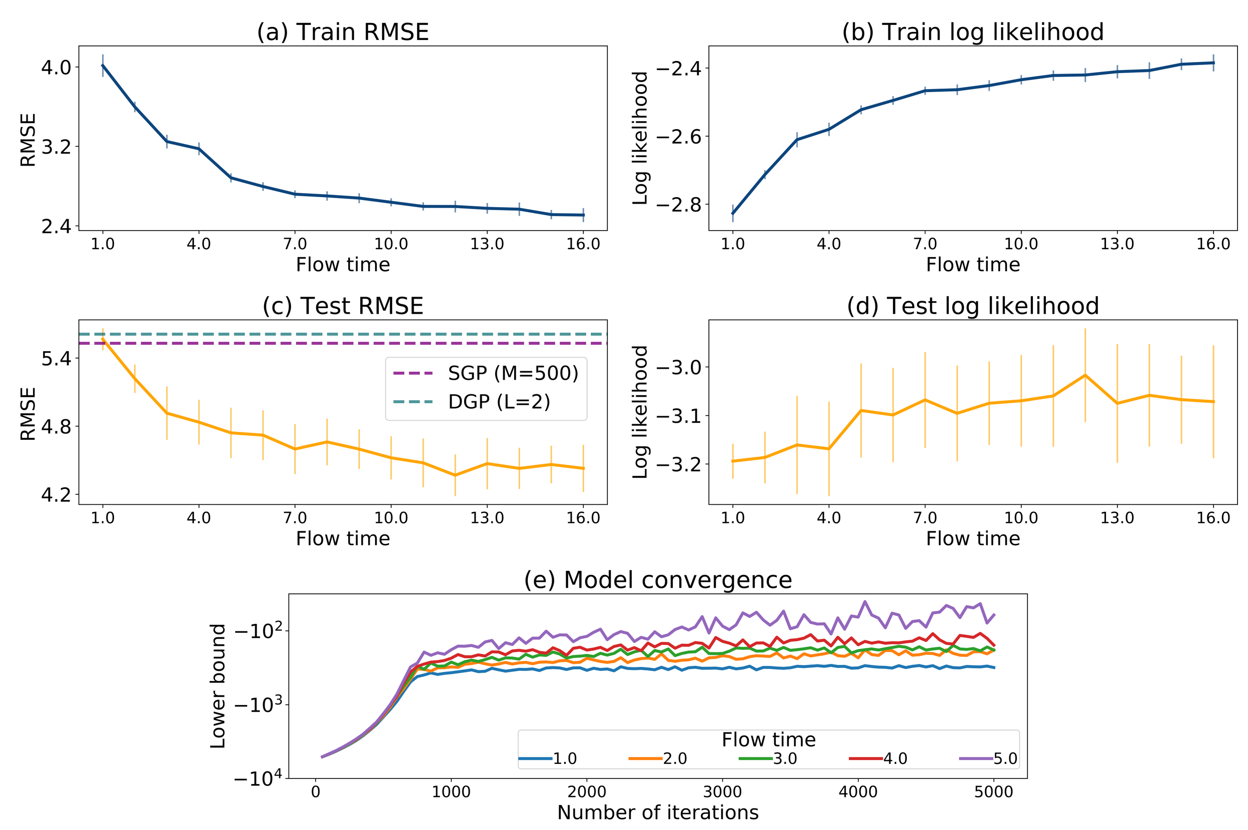
Observed input space **(a)** is transformed through stochastic continuous-time mappings **(b)** into a warped space **(c)**. The stationary Gaussian process in the warped space gives a smooth predictive distribution corresponding to highly non-stationary predictions in the original observed space.

UCI regression benchmarks

		boston	energy	concrete	wine_red	kin8mn	power	naval	protein
	N	506	768	1,030	1,599	8,192	9,568	11,934	45,730
	D	13	8	8	22	8	4	26	9
Linear		4.24	2.88	10.54	0.65	0.20	4.51	0.01	5.21
BNN	$L=2$	3.01	1.80	5.67	0.64	0.10	4.12	0.01	4.73
Sparse GP	$M=100$	2.87	0.78	5.97	0.63	0.09	3.91	**0.00**	4.43
	$M=500$	2.73	**0.47**	5.53	**0.62**	0.08	3.79	**0.00**	4.10
Deep GP $M=100$	$L=2$	2.90	**0.47**	5.61	0.63	**0.06**	3.79	**0.00**	4.00
	$L=3$	2.93	0.48	5.64	0.63	**0.06**	3.73	**0.00**	**3.81**
	$L=4$	2.90	0.48	5.68	0.63	**0.06**	3.71	**0.00**	**3.74**
	$L=5$	2.92	**0.47**	5.65	0.63	**0.06**	3.68	**0.00**	**3.72**
DiffGP $M=100$	$T=1.0$	2.80	0.49	**5.32**	0.63	**0.06**	3.76	**0.00**	4.04
	$T=2.0$	**2.68**	0.48	**4.96**	0.63	**0.06**	3.72	**0.00**	4.00
	$T=3.0$	**2.69**	**0.47**	**4.76**	0.63	**0.06**	3.68	**0.00**	3.92
	$T=4.0$	**2.67**	0.49	**4.65**	0.63	**0.06**	**3.66**	**0.00**	3.89
	$T=5.0$	**2.58**	0.50	**4.56**	0.63	**0.06**	**3.65**	**0.00**	3.87

The results are comparable with the other popular Bayesian approaches including BNNs and DGPs. The above table shows test RMSE values of 8 benchmark datasets (reproduced from [2]). Our method performs equal to very deep Gaussian process with a much simpler inference scheme.

6. Flow time



Increasing the flow time T improves the train and test errors **(a,c)**, likelihoods **(b,d)** and model convergence **(e)**.

- Increasing time can lead to an increase in the model capacity without over-fitting.
- Diffusion acts as regularization.

7. Contributions and conclusions

- We propose replacing discrete composition of **'layers' with a continuous-time composition of 'flows'**.
- We propose differentially deep Gaussian processes, a novel Bayesian deep learning model with a simple variational inference scheme.
- We empirically show excellent results in various regression tasks.

References

[1] J. Hensman, A. Matthews, and Z. Ghahramani. Scalable variational Gaussian process classification. In *Artificial Intelligence and Statistics*, pages 351–360, 2015.

[2] Hugh Salimbeni and Marc Deisenroth. Doubly stochastic variational inference for deep gaussian processes. In *Advances in Neural Information Processing Systems*, pages 4591–4602, 2017.