Deep Gaussian Processes for Multi-fidelity Modelling

Paper Summary, Analysis, Discussion

Sergey S. Tambovskiy [1]

[1] Ericsson Research

ITN WindMill Reading Group

Webinar, 15.04.2020

Multi-fidelity

Setting

We need data to successfully apply machine learning to:

- Data-driven modelling
- Experimental design
- Engineering tasks

Limitations:

- Expensive computer experiments (e.g. weather simulator)
- Hardware limitations (e.g. robotics)
- Emulators (e.g. reinforcement learning)

Design a model to sample new data points.

Multi-fidelity

Definition and Problem

Simulator samples $\rightarrow$ Data types $\leftarrow$ Real observations

$\downarrow$

Combine

1. Naive combination:
   - $\hookrightarrow$ Biased predictions
   - $\hookrightarrow$ Misspecified model
2. Combine in a principled manner (Multi-fidelity):

True observations $\downarrow$ High-fidelity

Sampled approximations $\downarrow$ Low-fidelity

Bayesian inference $\rightarrow$ Gaussian Process $\rightarrow$ address: Nonlinearities, Uncertainty, Overfitting.

Prior works

1. Kennedy & O'Hagan – GP, where kernel models linear correlations between data at T ordered fidelity levels, $\mathcal{O}\left(\left(\sum_{t=1}^{T} N_t\right)^3\right)$.

2. Le Gratiet & Garnier – Recursive model, fidelities are independent GPs, $\mathcal{O}\left(\sum_{t=1}^{T} N_t^3\right)$.

3. Perdikaris et al. – NAR-GP, note on similarity between nested GP models for multi-fidelity and Deep GPs.

4. Raissi & Karniadakis – Deep-MF, DNN combined with GP.

5. Other works:
   - ↪ Zaytsev & Burnaev – Scalability.
   - ↪ Liu et. al. – Mismatched training and target distributions.
   - ↪ Lam et. al. & Poloczek et. al. – Non-hierarchical ordering of fidelities.
   - ↪ Sen et. al. & Kandasamy et al. – Bayesian Optimisation, Bandit algorithms.

6. ...

Motivation

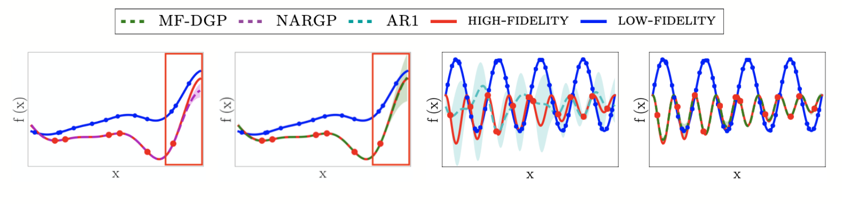
Autoregressive Model (AR1) – $f_t(x) = \rho f_{t-1}(x) + \delta_t(x)$ – scalar ρ cannot capture nonlinear transitions between fidelities.

Nonlinear Autoregressive Model (NAR-GP) – $f_t(x) = \rho\left(f_{t-1}(x)\right) + \delta_t(x)$

replacing the Gaussian process (GP) prior f_{t-1} with the GP posterior, $f^{\star}_{t-1}$

the additive structure and independence assumption allows to rewrite this as a composition of GPs $f_t(x) = g_t\left(f^{\star}_{t-1}(x), x\right)$

Disjointed nesting of GPs results in overfitting.



Deep Gaussian Processes

Supervised learning of a mapping between vectors and labels:

- Inputs: $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\top, \mathbf{x}_i \in \mathbb{R}^{D_{\text{in}}}$
- Labels: $\mathbf{y} = [y_1, \ldots, y_N]^\top, y_i \in \mathbb{R}$

Observations: noisy realisations of function values drawn from a GP.

- $\mathbf{f} = [f_1, \ldots, f_N]^\top$
- $p(\mathbf{y}|\mathbf{f})$

RBF Kernel:

- $k(\mathbf{x}_i, \mathbf{x}_j|\boldsymbol{\theta}) = \sigma^2 \exp\left[-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^\top \Lambda^{-1} (\mathbf{x}_i - \mathbf{x}_j)\right]$
- $\Lambda = \mathrm{diag}\left(l_1^2, \ldots, l_{D_{\text{in}}}^2\right)$

Deep GP:

- Multi-fidelity $\rightarrow$ interpretable latent functions.
- High computational complexity $\rightarrow$ use variational inference.

- Each DGP layer is a fidelity level.
- Cov. function. captures nonlinear mapping between levels and correlations in original input space.
- Stochastic variational inference propagates uncertainty across fidelity levels.

Default Kernel: $k_l = k_l^\rho\left(\mathbf{x}^i, \mathbf{x}^j\right) \times k_l^{f-1}\left(f_{l-1}^*\left(\mathbf{x}^i\right), f_{l-1}^*\left(\mathbf{x}^j\right)\right) + k_l^\delta\left(\mathbf{x}^i, \mathbf{x}^j\right)$

Alternate Kernel: $k_l = k_l^\rho\left(\mathbf{x}^i, \mathbf{x}^j\right) \times f_{l-1}^*\left(\mathbf{x}^i\right)^\top f_{l-1}^*\left(\mathbf{x}^j\right) + k_l^\delta\left(\mathbf{x}^i, \mathbf{x}^j\right)$

$$q\left(\mathbf{F}_l^t|\mathbf{U}_l\right) = p\left(\mathbf{F}_l^t|\mathbf{U}_l; \left\{\mathbf{F}_{l-1}^t, \mathbf{X}^t\right\}, \mathbf{Z}_{l-1}\right) q\left(\mathbf{U}_l\right)$$

$$\mathcal{L}_{\mathrm{MF-DGP}} = \sum_{t=1}^T \sum_{i=1}^{n_t} \mathbb{E}_{q\left(\mathbf{f}_t^{i,t}\right)}\left[\log p\left(y^{i,t}|\mathbf{f}_t^{i,t}\right)\right] + \sum_{l=1}^L D_{\mathrm{KL}}\left(q\left(\mathbf{U}_l\right) \| p\left(\mathbf{U}_l; \mathbf{Z}_{l-1}\right)\right)$$

Multi-fidelity Deep Gaussian Process

2

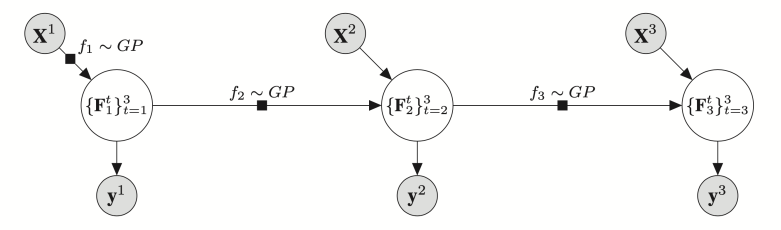


Figure 1: MF-DGP architecture with three fidelity levels.

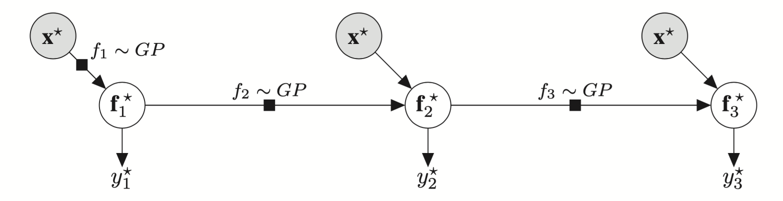


Figure 2: Making predictions with MF-DGP.

Complexity scales with:

$$\mathcal{O}\left(SNM^2\left(D_{\text{out, }1} + \dots + D_{\text{out. }L}\right)\right)$$

or $\mathcal{O}\left(SNM^2 L\right)$

Experiments

Uncertainty Quantification

EXAMPLE	FIDELITY	FUNCTION
LINEAR-A	LOW	$y_l(x) = \frac{1}{2} y_h(x) + 10\left(x - \frac{1}{2}\right) + 5$
	HIGH	$y_h(x) = (6x - 2)^2 \sin(12x - 4)$
LINEAR-B	LOW	$y_l(x) = 2 y_h(x) + \left(x^3 - \frac{1}{2}\right) \sin\left(3x - \frac{1}{2}\right) + 4\cos(2x)$
	HIGH	$y_h(x) = 5x^2 \sin(12x)$
NONLINEAR-A	LOW	$y_l(x) = \sin(8\pi x)$
	HIGH	$y_h(x) = (x - \sqrt{2})\,(y_l(x))^2$
NONLINEAR-B	LOW	$y_l(x) = \cos(15x)$
	HIGH	$y_h(x) = x e^{y_l(2x - .2)} - 1$

Figure 3: Synthetic benchmarks.

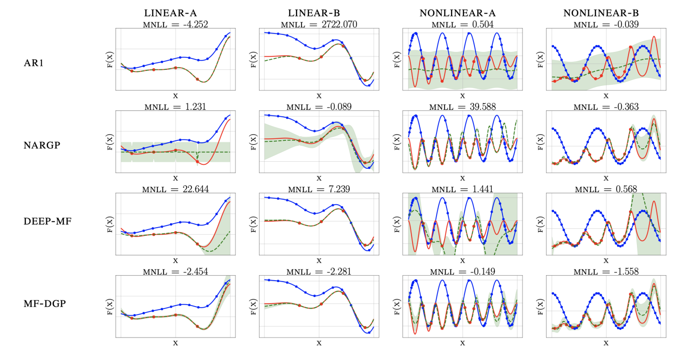


Figure 4: Methods and benchmarks for multi-fidelity scenarios.

Experiments

Benchmark Comparison

BENCHMARK	D_{in}	FIDELITY ALLOCATION	AR1 R^2	AR1 RMSE	AR1 MNLL	NARGP R^2	NARGP RMSE	NARGP MNLL	MF-DGP R^2	MF-DGP RMSE	MF-DGP MNLL
CURRIN	2	12-5	0.913	0.677	20.105	0.903	0.740	20.817	**0.935**	**0.601**	**0.763**
PARK	4	30-5	**0.985**	0.575	465.377	0.954	0.928	743.119	**0.985**	**0.565**	**1.383**
BOREHOLE	8	60-5	**1.000**	**0.005**	**-3.946**	0.973	0.063	-1.054	0.999	0.015	-2.031
BRANIN	2	80-30-10	0.891	0.044	-1.740	0.929	0.053	-1.223	**0.965**	**0.030**	**-2.572**
HARTMANN-3D	3	80-40-20	**0.998**	**0.043**	0.440	0.305	0.755	0.637	0.994	0.075	**-0.731**

Figure 5: Model comparison on multi-fidelity benchmark examples.

Experiments

Multi-fidelity in the Loop

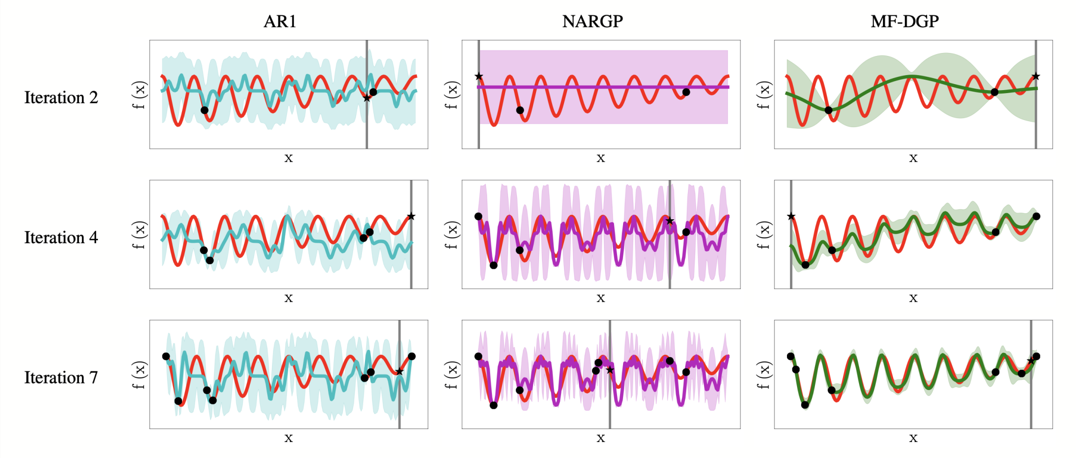


Figure 6: Comparison of fits after 2, 4 and 7 steps in the experimental design loop.

Key points and Discussion

Key points:

- An interpretation of DGPs whereby each layer models data with an increasing level of fidelity;
- Superior quantification of uncertainty given limited high-fidelity observations;
- Application to procedures such as experimental deisgn

Discussion:

1. How would You deal with inducing points in deeper layers?
   ↪ Suggestions?
2. What alternative multi-fidelity kernels would You choose?
3. have You seen other suitable robust multi-step optimisation schemes?
4. What alternatives to proposed modelling scheme exist?
   ↪ Scenarios?
5. How can problem become more "complex" (from theoretical perspective)?

Reference links

1. Deep Gaussian Processes for Multi-fidelity Modeling – Paper
2. Deep Gaussian Processes for Multi-fidelity Modeling – Poster
3. Deep Gaussian Processes for Multi-fidelity Modeling – Video

Thank You for Your attention!