Deep Gaussian Processes for Multi-fidelity Modeling

Kurt Cutajar[*]
EURECOM
Sophia Antipolis, France

Mark Pullin
Amazon
Cambridge, UK

Andreas Damianou
Amazon
Cambridge, UK

Neil Lawrence
Amazon
Cambridge, UK

Javier González
Amazon
Cambridge, UK

Abstract

Multi-fidelity models are prominently used in various science and engineering applications where cheaply-obtained, but possibly biased and noisy observations must be effectively combined with limited or expensive true data in order to construct reliable models. The notion of applying deep Gaussian processes (DGPs) to this setting has recently shown great promise by capturing complex nonlinear correlations across fidelities. However, the architectures explored thus far are burdened by structural assumptions and constraints which deter such models from performing to the best of their expected capabilities. In this paper we propose a novel approach for DGP multi-fidelity modeling which treats DGP layers as fidelity levels and uses a variational inference scheme to propagate uncertainty across them. In our experiments, we show that this approach makes substantial improvements in quantifying and propagating uncertainty in multi-fidelity set-ups, which in turn improves their effectiveness in decision-making pipelines.

1 Introduction

Multi-fidelity models [4, 7] are designed to fuse limited true observations (high-fidelity) with cheaply-obtained lower granularity representations (low-fidelity). Gaussian processes [GPs; 9] are well-suited to multi-fidelity problems due to their ability to encode prior beliefs about how fidelities are related, yielding predictions accompanied by uncertainty estimates. GPs formed the basis of seminal autoregressive models (AR1) investigated by [4] and [6], and are suitable when the mapping between fidelities is linear, i.e. the high-fidelity function f_t can be modeled as:

$$f_t(x) = \rho f_{t-1}(x) + \delta_t(x), \tag{1}$$

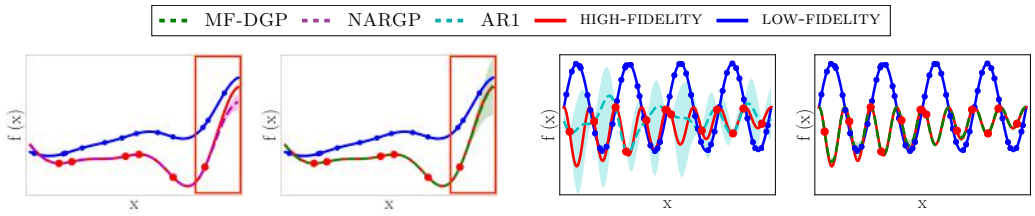
where ρ is a constant scaling the contribution of samples f_{t-1} drawn from the GP modeling the data at the preceding fidelity, and $\delta_t(x)$ models the bias between fidelities. However, this is insufficient when the mapping is nonlinear, i.e. ρ is now a nonlinear transformation such that:

$$f_t(x) = \rho_t(f_{t-1}(x)) + \delta_t(x). \tag{2}$$

The additive structure and independence assumption between the GPs for modeling $\rho_t(f_{t-1}(x))$ and $\delta_t(x)$ permits us to combine these as a single GP that takes as inputs both x and $f^*_{t-1}(x)$, which here denotes a sample from the posterior of the GP modeling the preceding fidelity evaluated at x. This can be expressed as $f_t(x) = g_t\left(f^*_{t-1}(x), x\right)$.

Third workshop on Bayesian Deep Learning (NeurIPS 2018), Montréal, Canada.

[*]Work carried out during an internship at Amazon, Cambridge.



(a) *Left:* Overfitting in the NARGP model. *Right:* Well-calibrated fit using proposed MF-DGP model.

(b) *Left:* AR1 cannot capture nonlinear mappings. *Right:* Fixed by compositional structure of MF-DGP.

Figure 1: Limitations addressed and resolved jointly by MF-DGP. Blue and red markers denote low and high-fidelity observations respectively. Shaded regions indicate the 95% confidence interval.

Deep Gaussian processes [DGPs; 2] are a natural candidate for handling such relationships, allowing for uncertainty propagation in a nested structure of GPs where each GP models the transition from one fidelity to the next. However, DGPs are cumbersome to develop and approximations are necessary for enabling tractable inference. While motivated by the structure of DGPs, the nonlinear multi-fidelity model (NARGP) proposed in [8] amounts to a disjointed architecture whereby each GP is fitted in an isolated hierarchical manner, preventing GPs at lower fidelities from being updated once they have been fit. Consider the example given in Figure 1a. In the boxed area, we would expect the model to return high uncertainty to reflect the lack of data available, but overfitting in NARGP results in predicting an incorrect result with reasonably high confidence.

Contribution: In this work, we propose the first complete interpretation of multi-fidelity modeling using DGPs, which we refer to as MF-DGP. In particular, we leverage the sparse DGP approximation proposed in [10] for constructing a multi-fidelity DGP model which can be trained end-to-end, overcoming the constraints that hinder existing attempts at using DGP structure for this purpose. Returning to the example given in Figure 1a, we see that our model fits the true function properly while also returning sensibly conservative uncertainty estimates. Additionally, our model also inherits the compositional structure of NARGP, alleviating a crucial limitation of AR1 (Figure 1b).

2 Multi-fidelity Deep Gaussian Process (MF-DGP)

The application of DGPs to the multi-fidelity setting is particularly appealing because if we assume that each layer corresponds to a fidelity level, then the latent functions at the intermediate layers are given a meaningful interpretation which is not always available in standard DGP models. The first attempt at using compositions of GPs in a multi-fidelity setting [8] relied on structural assumptions on the data to circumvent the intractability of DGPs, but this heavily impairs their expected flexibility. Recent advances in the DGP literature [1, 10] have leveraged traditional GP approximations to construct scalable DGP models which are easier to specify and train; we build our extension atop the model presented in [10] to avoid the constraints imposed on selecting kernel functions in [1].

2.1 Model Specification

Let us assume a dataset $\mathcal{D}$ having observations at T fidelities, where $\mathbf{X}^t$ and $\mathbf{y}^t$ denote the n_t inputs and corresponding outputs observed with fidelity level t:

$$\mathcal{D} = \left\{ \left(\mathbf{X}^1, \mathbf{y}^1\right), \ldots, \left(\mathbf{X}^t, \mathbf{y}^t\right), \ldots, \left(\mathbf{X}^T, \mathbf{y}^T\right) \right\}.$$

For enhanced interpretability, we assume that each layer of our MF-DGP model corresponds to the process modeling the observations available at fidelity level t, and that the bias or deviation from the true function decreases from one level to the next. We use the notation $\mathbf{F}_l^t$ to denote the evaluation at layer l for inputs observed with fidelity t; for example, the evaluation of the process at layer '1' for the inputs observed with fidelity '3' is denoted as $\mathbf{F}_1^3$. A conceptual illustration of the proposed MF-DGP architecture is given in Figure 2 (left) for a dataset with three fidelities. Note that the GP at each layer is conditioned on the data belonging to that level, as well as the evaluation of that same input data at the preceding fidelity. This gives greater purpose to the notion of feeding forward the original inputs at each layer, as originally suggested in [3] for avoiding pathologies in deep architectures.

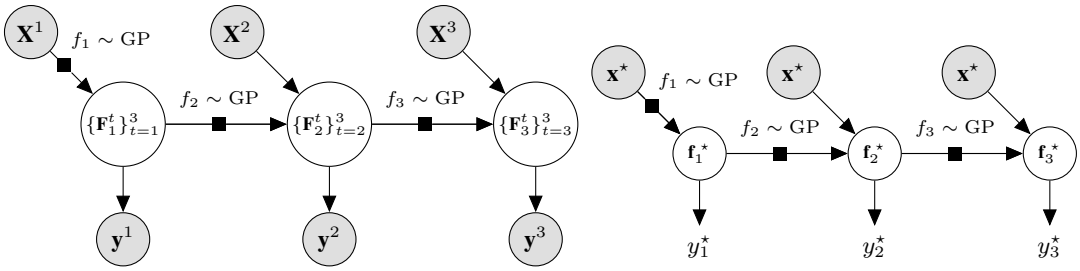


Figure 2: *Left:* MF-DGP architecture with 3 fidelity levels. *Right:* Predictions using same MF-DGP.

At each layer we rely on the sparse variational approximation of a GP for inference, thus obtaining the following variational posterior distribution:

$$q\left(\mathbf{F}_l^t|\mathbf{U}_l\right) = p\left(\mathbf{F}_l^t|\mathbf{U}_l; \{\mathbf{F}_{l-1}^t, \mathbf{X}^t\}, \mathbf{Z}_{l-1}\right) q\left(\mathbf{U}_l\right), \tag{3}$$

where $\mathbf{Z}_{l-1}$ denotes the inducing inputs for l, $\mathbf{U}_l$ their corresponding function evaluation, and $q\left(\mathbf{U}_l\right) = \mathcal{N}\left(\mathbf{U}_l|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l\right)$ is the variational approximation of the inducing points. The mean and variance defining this variational approximation, i.e. $\boldsymbol{\mu}_l$ and $\boldsymbol{\Sigma}_l$, are optimized during training. Furthermore, if $\mathbf{U}_l$ is marginalized out from Equation 3, the resulting variational posterior is once again Gaussian and fully defined by its mean, $\widetilde{\mathbf{m}}_l$, and variance, $\widetilde{\mathbf{S}}_l$:

$$q\left(\mathbf{F}_l^t|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l; \{\mathbf{F}_{l-1}^t, \mathbf{X}^t\}, \mathbf{Z}_{l-1}\right) = \mathcal{N}\left(\mathbf{F}_l^t \mid \widetilde{\mathbf{m}}_l^t, \widetilde{\mathbf{S}}_l^t\right), \tag{4}$$

which can be derived analytically. The likelihood noise at lower fidelity levels is encoded as additive white noise in the kernel function of the GP at that layer.

We can then formulate the variational lower bound on the marginal likelihood as follows:

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

where we assume that the likelihood is factorized across fidelities and observations, and D_{KL} denotes the Kullback-Leibler divergence. Samples from the model are obtained recursively using the reparameterization trick [5] to draw samples from the variational posterior.

Model predictions with different fidelities are also obtained recursively by propagating the input through the model up to the chosen fidelity. At all intermediate layers, the output from the preceding layer is augmented with the original input, as will be made evident by the choice of kernel explained in the next section. The output of a test point $\mathbf{x}^\star$ can then be predicted with fidelity level t as follows:

$$q\left(\mathbf{f}_t^\star\right) \approx \frac{1}{S}\sum_{s=1}^{S} q\left(\mathbf{f}_t^{s,\star}|\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t; \{\mathbf{f}_{t-1}^{s,\star}, \mathbf{x}^\star\}, \mathbf{Z}_{t-1}\right), \tag{5}$$

where S denotes the number of Monte Carlo samples and t replaces l as the layer indicator. This procedure is illustrated in Figure 2 (*right*).

2.2 Multi-fidelity Covariance

For every GP at an intermediate layer, we opt for the multi-fidelity kernel function proposed in [8], since this captures both the potentially nonlinear mapping between outputs as well as the correlation in the original input space:

$$k_l = k_l^\rho\left(\mathbf{x}^i, \mathbf{x}^j; \boldsymbol{\theta}_l^\rho\right) k_l^{f-1}\left(f_{l-1}^*(\mathbf{x}^i), f_{l-1}^*(\mathbf{x}^j); \boldsymbol{\theta}_l^{f-1}\right) + k_l^\delta\left(\mathbf{x}^i, \mathbf{x}^j; \boldsymbol{\theta}_l^\delta\right), \tag{6}$$

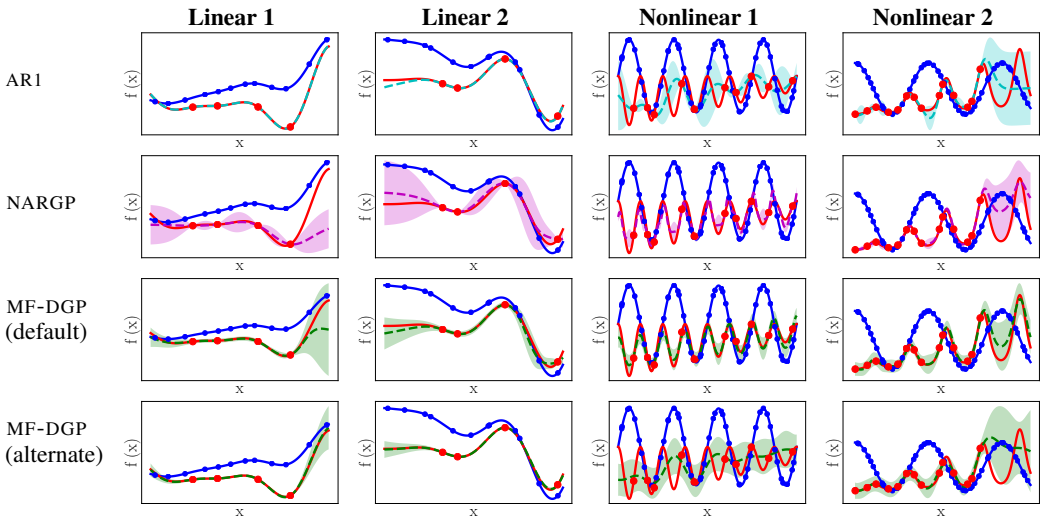


Figure 3: Comparison across methods and benchmarks for challenging multi-fidelity scenarios. The importance of choosing an appropriate kernel for MF-DGP is also reinforced here.

where k_l^{f-1} denotes the covariance between outputs obtained from the preceding fidelity level, k_l^ρ is a space-dependent scaling factor, and k_l^δ captures the bias at that fidelity level. At the first layer this reduces to $k_1 = k_1^\delta\left(\mathbf{x}^i, \mathbf{x}^j; \boldsymbol{\theta}_1^\delta\right)$.

In [8], it was assumed that each individual component of the composite kernel function is an RBF kernel, and we shall also assume this to be the default setting for MF-DGP. However, this may not be appropriate when the mapping between fidelities is linear. In such instances, we propose to replace k_l^{f-1} with an alternate linear kernel such that the composite intermediate layer covariance becomes:

$$k_l = k_l^\rho\left(\mathbf{x}^i, \mathbf{x}^j; \boldsymbol{\theta}_l^\rho\right) f_{l-1}^*(\mathbf{x}^i)^\top f_{l-1}^*(\mathbf{x}^j) + k_l^\delta\left(\mathbf{x}^i, \mathbf{x}^j; \boldsymbol{\theta}_l^\delta\right). \tag{7}$$

3 Experimental Evaluation

In the preceding sections, we demonstrated how the formulation of state-of-the-art DGP models can be adapted to the multi-fidelity setting. Through a series of experiments, we validate that beyond its novelty and theoretic appeal, the proposed MF-DGP model also works well in practice.

Improved UQ: We empirically validate MF-DGP's well-calibrated uncertainty quantification by considering experimental set-ups where the available data is generally insufficient to yield confident predictions, and higher uncertainty is prized. In Figure 3, we consider multi-fidelity scenarios where the allocation of high-fidelity data is limited or constrained to lie in one area of the input domain. In all of the examples, our model yields appropriately conservative estimates in regions where insufficient observations are available. As evidenced by the overfitting exhibited by AR1 for the LINEAR 2 example, deep models can also be useful for problems having linear mappings.

Table 1: Model comparison on multi-fidelity benchmark examples. 'Default' indicates use of the kernel listed in Equation 6, while 'alternate' indicates that the covariance in Equation 7 was used.

			Mean Squared Error		
Benchmark	n_{low}	n_{high}	**AR1**	**NARGP**	**MF-DGP**
Linear 1	12	6	0.0074	2.0711	0.0156 (alternate)
Linear 2	12	5	0.1071	0.5937	0.1157 (alternate)
Nonlinear 1	50	14	0.1385	7.1270e-5	0.0004 (default)
Nonlinear 2	50	14	0.2307	0.0331	0.0205 (default)

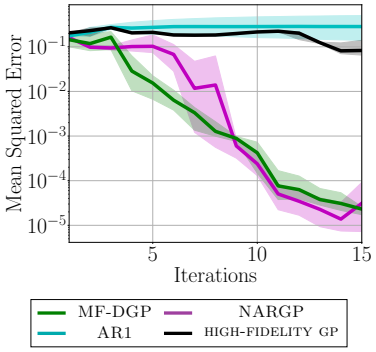


Figure 4: Experimental design loop.

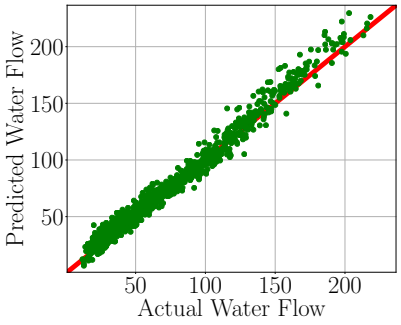


Figure 5: MF-DGP fit to Borehole function.

Benchmark Comparison: We also compare the predictive performance of MF-DGP to AR1 and NARGP on the same selection of benchmark examples. Twenty randomly-generated training sets are prepared for each example function, following the allocation of low and high-fidelity points listed in Table 1. The results denote the average mean squared error obtained using each model over a fixed test set covering the entire input domain. The obtained results give credence to our intuition that MF-DGP balances out issues in the two modeling approaches; it performs as well as NARGP on the nonlinear examples where AR1 falters, and outperforms the former on linear examples.

Multi-fidelity in the Loop: We further assess MF-DGP using an expository experimental design loop whereby points are sequentially chosen to reduce uncertainty about a function of interest. Starting with 20 low-fidelity and 3 high-fidelity observations, we learn the NONLINEAR 1 function by selecting to observe points where the variance of the predictive distribution at the high fidelity is largest. Figure 4 shows how the mean squared error against a constant test set evolves as more points are collected, averaged over 5 runs with different initial training data. Here we also compare against a standard GP trained on the high-fidelity observations only. As expected, NARGP and MF-DGP perform best as the model structure better represents the underlying data. Although NARGP and MF-DGP both converge to a similar solution once enough points are sampled, the benefit of using MF-DGP is evidenced in the initial steps of the procedure, whereby it fits the data sensibly after only few iterations.

Real-world Simulation: We fit MF-DGP to a two-level function that simulates stochastic water flow through a borehole [11] and depends on eight input parameters, for which a dataset of 150 low and 40 high-fidelity points was generated. Figure 5 illustrates the performance of MF-DGP for a test set containing 1000 high-fidelity points, where it achieves an R^2 of 0.98.

4 Conclusion

Reliable decision making under uncertainty is a core requirement in multi-fidelity scenarios where unbiased observations are scarce or difficult to obtain. In this paper, we proposed the first complete specification of a multi-fidelity model as a DGP that is capable of capturing nonlinear relationships between fidelities with reduced overfitting. By providing end-to-end training across all fidelity levels, MF-DGP yields superior quantification and propagation of uncertainty that is crucial in iterative methods such as experimental design. In spite of being prevalent in engineering applications, we believe that multi-fidelity modeling has been under-explored by the machine learning community, and hope that this work can reignite further interest in this direction.

References

[1] K. Cutajar, E. V. Bonilla, P. Michiardi, and M. Filippone. Random feature expansions for deep Gaussian processes. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 884–893, 2017.

[2] A. C. Damianou and N. D. Lawrence. Deep Gaussian processes. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2013, Scotts-*

dale, AZ, USA, April 29 - May 1, 2013, pages 207–215, 2013.

[3] D. K. Duvenaud, O. Rippel, R. P. Adams, and Z. Ghahramani. Avoiding pathologies in very deep networks. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, AISTATS 2014, Reykjavik, Iceland, April 22-25, 2014*, pages 202–210, 2014.

[4] M. C. Kennedy and A. O'Hagan. Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1):1–13, 2000.

[5] D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *Proceedings of the Second International Conference on Learning Representations, ICLR 2014, Banff, Canada, April 14-16, 2014*, 2014.

[6] L. Le Gratiet and J. Garnier. Recursive co-kriging model for design of computer experiments with multiple levels of fidelity. *International Journal for Uncertainty Quantification*, 4(5), 2014.

[7] B. Peherstorfer, K. Willcox, and M. Gunzburger. Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *SIAM Review*, 60(3):550–591, 2018.

[8] P. Perdikaris, M. Raissi, A. Damianou, N. D. Lawrence, and G. E. Karniadakis. Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473(2198):20160751, 2017.

[9] C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, 2006.

[10] H. Salimbeni and M. P. Deisenroth. Doubly stochastic variational inference for deep Gaussian processes. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 4591–4602, 2017.

[11] S. Xiong, P. Z. G. Qian, and C. F. J. Wu. Sequential design and analysis of high-accuracy and low-accuracy computer codes. *Technometrics*, 55(1):37–46, 2013.