3rd Symposium on Advances in Approximate Bayesian Inference, 2020 1–9

# Neural Linear Models with Functional Gaussian Process Priors

**Anonymous Authors**
*Anonymous Institution*

## Abstract

Neural linear models (NLM) and Gaussian processes (GP) are both examples of Bayesian linear regression on rich feature spaces. In contrast to the widespread use of nonparametric GPs for probabilistic nonlinear regression, NLMs remain an underused parametric alternative because standard type II maximum likelihood (ML) training leads to overconfidence outside of the data distribution. Therefore, we propose to augment this training procedure through functional variational inference (fVI) proposed by Sun et al. (2019), which is particularly well suited for NLMs due to their closed-form predictive distribution. Additionally, we investigate whether an appropriate functional prior can guide parametric NLMs to attain nonparametric GP performance, despite using fewer parameters. Results show that functional priors do improve performance of NLM over ML training, and that the NLM performs on par with weight space BNNs in this setting.

## 1. Introduction

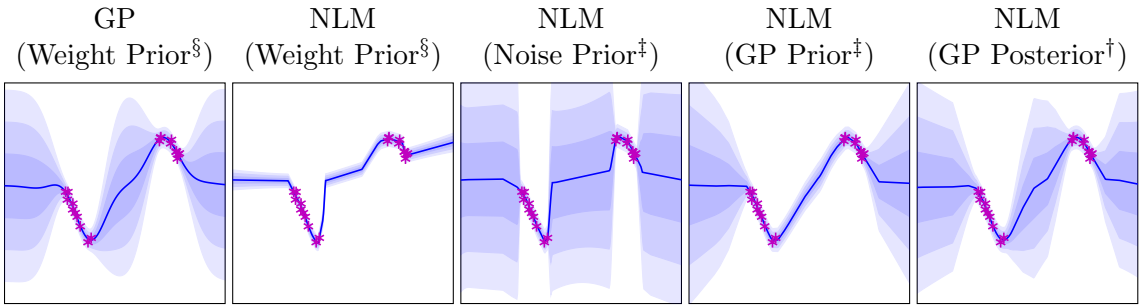


Figure 1: Predictive densities of Bayesian models —— of data *. The models are trained with type II maximum likelihood (§), functional ELBO (‡) and functional KL (†) objectives, and the NLMs use leaky ReLU activations. Functional priors demonstrate that the NLM is expressive enough to learn diverse feature spaces and approximate a GP, but type II ML and fELBO training can lead to overfitting.

The neural linear model (NLM) (Lázaro-Gredilla and Figueiras-Vidal (2010); Ober and Rasmussen (2019)) is a Bayesian neural network which combines deterministic neural network features with a Bayesian weight distibution for the final layer. As such, it is a form of Bayesian linear regression with a finite feature space, contrasting Gaussian processes (GP) with infinite features spaces induced by kernel functions (Rasmussen and Williams (2005)). These infinite feature spaces make GPs rich probabilistic models for nonlinear regression, which are capable of expressing well-calibrated uncertainty quantification (UQ) outside of

© A. Authors.

the data distribution (OOD). This UQ is crucial for tasks where sample efficiency and safety are critical, such as active learning, black-box optimization and model-based reinforcement learning. Bayesian neural networks (BNN) have been widely criticized for their poor UQ (e.g. Foong et al. (2019); Osband et al. (2018)), which limits their application to these tasks for which Bayesian methods are preferred.

The neural linear model has several attractive qualities compared to alternative BNN approaches, such as Hamiltonian Monte Carlo (Neal (1995)), stochastic variational inference (Blundell et al. (2015)), Monte Carlo dropout (Gal and Ghahramani (2016)) and deep ensembles (Lakshminarayanan et al. (2017)). The predictive distribution is closed-form, rather than an implicit distribution, reducing the computational complexity of prediction. As it is a Bayesian linear model, it also has much fewer parameters for equivalent architectures compared to models with weight-space priors. However, as this neural feature space is trained with type II maximum likelihood (ML), the features overfit to the data and struggle to express well-calibrated UQ when predicting OOD. The inductive bias of the GP's expressive kernels has so far demonstrated to be a superior approach. Despite this, as GPs struggle to scale as well as BNNs due to the nonparametric computation, we are motivated to enable NLMs that can approach GP performance. This question is then whether the neural feature space is *capable* of matching or surpassing the performance of Gaussian process kernels.

To approach this, we look at functional variational inference (fVI), which seeks to apply approximate inference over function distributions. With the functional variational BNN (fBNN), Sun et al. (2019) used a GP posterior as a functional prior for BNNs, which was then refined on the data using the functional evidence lower bound objective (fELBO). While this approach is essentially a form of empirical Bayes that requires two Bayesian models to be learned, we seek to use fVI to assess the representional power of NLMs in the context of GPs and other priors. If NLMs are capable of matching or surpassing the performance of these nonparametric models, then this motivates research into learning objectives for NLMs that imitates the qualities of these rich kernel feature spaces.

## 2. Neural Linear Model

Intuitively, NLMs can be viewed as Bayesian linear regression in a projected feature space, where the projection is learned by a neural network. Note that while the NLM can be easily deployed for multivariate regression, the following derivations (and later experiments) in this work focus on univariate targets for simplicity.

Let $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be the observed data, $\mathbf{x}_i \in \mathbb{R}^k$, $y_i \in \mathbb{R}$, such that $\mathbf{X} \in \mathbb{R}^{n \times k}$ and $\mathbf{y} \in \mathbb{R}^n$. Additionally, let $\boldsymbol{\phi}(\cdot; \boldsymbol{\theta}) \colon \mathbb{R}^k \to \mathbb{R}^m$ be a feature space projection with parameters $\boldsymbol{\theta}$, $\boldsymbol{\phi}_i = \boldsymbol{\phi}(\mathbf{x}_i; \boldsymbol{\theta})$ and $\boldsymbol{\Phi} = \left[\boldsymbol{\phi}_1^\top \dots \boldsymbol{\phi}_n^\top\right] \in \mathbb{R}^{n \times m}$, the matrix of vertically stacked row vectors. It is common to add constant or linear terms to $\boldsymbol{\Phi}$ to implicitly represent bias or identity terms. However, for notational clarity, we ignore these terms and denote the projected feature space as $\mathbb{R}^m$.

A latent function $f$ is modeled using Bayesian linear regression (Box and Tiao (1973); Bishop (2006); Murphy (2012)) with weights $\boldsymbol{\beta}$ and additive, zero-mean, Gaussian noise $\epsilon$

with variance $\sigma^2$, where

$$y_i = f(\mathbf{x}_i; \boldsymbol{\theta}) = \boldsymbol{\phi}_i^\top \boldsymbol{\beta} + \epsilon_i. \tag{1}$$

Placing a conjugate Gaussian prior $\mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0^{-1})$ over $\boldsymbol{\beta}$ results in a Gaussian posterior $\mathcal{N}(\boldsymbol{\mu}_n, \boldsymbol{\Lambda}_n^{-1})$ with an explicit Gaussian predictive distribution for query $\mathbf{x}$,

$$y \mid \mathbf{x}, \mathcal{D}, \boldsymbol{\theta} \sim \mathcal{N}(\cdot \mid \boldsymbol{\phi}_\mathbf{x}^\top \boldsymbol{\mu}_n, \sigma^2 + \boldsymbol{\phi}_\mathbf{x}^\top \boldsymbol{\Lambda}_n^{-1} \boldsymbol{\phi}_\mathbf{x}), \tag{2}$$

where $\boldsymbol{\mu}_n$ and $\boldsymbol{\Lambda}_n$ are the mean vector and precision matrix of the posterior weight distribution. The exact posteriors are detailed in Section A. The observation noise $\sigma^2$, prior weight parameters $\boldsymbol{\mu}_0$ and $\boldsymbol{\Lambda}_0$, and $\boldsymbol{\theta}$ can be either set to constants or optimized jointly by maximizing the marginal likelihood. With $\boldsymbol{\mu}_0 = \mathbf{0}$, this model is equivalent to a Gaussian process with kernel $k(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta}) = \boldsymbol{\phi}(\mathbf{x}; \boldsymbol{\theta})^\top \boldsymbol{\Lambda}_0^{-1} \boldsymbol{\phi}(\mathbf{x}'; \boldsymbol{\theta})$ (Rasmussen and Williams (2005)).

## 3. Functional Variational Inference

Functional variational inference considers a functional evidence lower bound objective (fELBO) through use of a functional KL divergence (fKL) and functional prior $p(f)$,

$$\mathcal{L}(\mathcal{D}) := \mathbb{E}_{f \sim q(\cdot)}[\log p(\mathcal{D} \mid f)] - D_{\mathrm{KL}}(q(f) \mid\mid p(f)). \tag{3}$$

Here, $\mathcal{L}(\mathcal{D})$ is a lower bound of the log-marginal likelihood $\log p(\mathcal{D})$, such that maximizing $\mathcal{L}(\mathcal{D})$ translates to maximizing $p(\mathcal{D})$. In general, $p(f)$ and $q(f)$ can be arbitrary stochastic processes, i.e. distributions over functions, and the fKL between two stochastic processes (de G. Matthews et al. (2016)) is a generalization of the regular KL divergence between two finite-dimensional probability distributions. However, the generalization to function space introduces obstacles in terms of tractability. Roughly speaking, evaluation of the fKL requires integration in an infinite-dimensional vector space, which is intractable because there is no infinite-dimensional Lebesgue measure (Hunt (1992)). Nonetheless, it is possible to approximate this quantity by averaging over a finite number of evaluations at $M$ locations $\mathbf{x}_m$ (Sun et al. (2019)),

$$D_{\mathrm{KL}}(q(f) \mid\mid p(f)) \approx \tfrac{1}{M} \textstyle\sum_{m=1}^{M} D_{\mathrm{KL}}(f_q(\mathbf{x}_m) \mid\mid f_p(\mathbf{x}_m)), \tag{4}$$

such that the quantity is concrete.

A key factor for a successful fKL approximation is the choice of $\{\mathbf{x}_m\}$. Ideally, $\{\mathbf{x}_m\}$ should be representative of the domain of $f$. At first sight, this may seem very feasible for $f : \mathbb{R}^k \to \mathbb{R}$. One could simply draw samples uniformly at random from $\mathbb{R}^k$. However, the infiniteness of the problem quickly manifests itself: It is impossible to entirely cover $\mathbb{R}^k$ with any finite number of samples. Therefore, it is necessary to select a bounded subset of the original domain as 'domain of interest', which can be challenging if the underlying data distribution is unknown.

## 4. Functional Priors in Practice

While the variational stochastic process $q(f)$ can be modeled using any probabilistic function approximator, such as a NLM or any other Bayesian neural network, a beneficial yet

practicable choice for the prior stochastic process $p(f)$ is less apparent. White Gaussian noise (WGN), more specifically a Gaussian process with zero mean and isotropic covariance,

$$p(f) := \mathcal{GP}(\cdot \mid 0, \sigma_p^2), \qquad f_p(\mathbf{x}) := \mathcal{N}(\cdot \mid 0, \sigma_p^2), \tag{5}$$

is a simple candidate akin to commonly used weight space priors. To attain scale invariance across datasets, it is sensible to define this prior in whitened space.

Alternatively, it is possible to use a more informative prior, such as a Gaussian process posterior $\mathcal{GP}(f \mid 0, \mathbf{K})$ whose kernel parameters were calibrated using the available training data (Sun et al. (2019); Shi et al. (2019)). The scalar $k$, vector $\mathbf{k}$ and matrix $\mathbf{K}$ are computed using the kernel, training data and function input $\mathbf{x}$, so

$$f_p(\mathbf{x}) := \mathcal{N}(\cdot \mid \mathbb{E}(\mathbf{x}), \mathbb{V}(\mathbf{x})), \qquad \mathbb{E}(\mathbf{x}) = \mathbf{k}^\top \mathbf{K}^{-1}\mathbf{y}, \qquad \mathbb{V}(\mathbf{x}) = k - \mathbf{k}^\top \mathbf{K}^{-1}\mathbf{k}. \tag{6}$$

Note that this choice of functional prior $p(f)$ involves the training data $\mathcal{D}$. Therefore, it is not a pure prior in a Bayesian sense, but could be interpreted as a form of empirical Bayes (Morris (1983)), which may result in overfitting.

Lastly, it is also possible to use the fKL directly so that $q(f)$ approximates another stochastic process $p(f)$, e.g. a Gaussian process posterior. If $p(f)$ adequately represents the observed data then $q(f)$ will naturally fit the data by minimizing the fKL while also inheriting other properties, such as extrapolation behavior and uncertainty quantification. In particular, this functional inference can transfer stochastic processes which are modeled by arbitrary probabilistic function approximators, e.g. GPs and NLMs, even if both models have considerably different architectures. This could enable model compression and distillation, which may be practical for nonparametric models like Gaussian processes.

## 5. Experimental Results

This section discusses our nonlinear regression experiment, which aims to demonstrate the viability of combining NLMs with functional inference. To this end, we replicated the experiments presented in Sun et al. (2019) and evaluated the NLM on several UCI regression datasets. The goal is to assess how the NLM compares to the mean-field variational inference (MFVI) model in this setting, while also comparing the fKL and fELBO objective, given the NLM's tendency to overfit.

All models were implemented using PyTorch (Paszke et al. (2019)). For the GPs we additionally used GPyTorch (Gardner et al. (2018)). The NLM was implemented with $\boldsymbol{\mu}_0 = \mathbf{0}$ and a diagonal matrix with $m$ distinct parameters to represent $\boldsymbol{\Lambda}_0$. The observation noise parameter $\sigma^2$, neural network weights $\boldsymbol{\theta}$ and $\boldsymbol{\Lambda}_0$ were learned via backpropagation. To encode positive value constraints, $\sigma^2$ and $\boldsymbol{\Lambda}_0$ were learned in log-space. Following Sun et al. (2019), NLMs used a single hidden layer with 50 units. For optimization parameters, please refer to Appendix B.

Table 1 displays the results of our UCI regression benchmark experiments, comparing NLMs to GPs and previous work. In general, the NLM achieves competitive performance. Although using the WGN prior results in underfitting for data with low observation noise (energy, yacht), it is mostly beneficial for the data with high observation noise (boston, concrete, wine). Applying fELBO inference with the GP prior, the NLM mostly outperforms

| MODEL | | | BOSTON | CONCRETE | ENERGY | WINE | YACHT |
|---|---|---|---|---|---|---|---|
| GP | § | RBF | $-2.516 \pm 0.114$ | $-3.512 \pm 0.011$ | $\mathbf{-0.659 \pm 0.050}$ | $\mathbf{-0.963 \pm 0.021}$ | $\mathbf{-0.150 \pm 0.037}$ |
| NLM | § | LRELU | $-2.563 \pm 0.048$ | $-3.105 \pm 0.026$ | $-0.905 \pm 0.058$ | $-1.000 \pm 0.013$ | $-1.325 \pm 0.087$ |
| | | TANH | $-2.634 \pm 0.043$ | $-3.107 \pm 0.019$ | $-1.067 \pm 0.077$ | $-0.991 \pm 0.011$ | $-0.954 \pm 0.082$ |
| | WGN ‡ | LRELU | $-2.569 \pm 0.069$ | $-3.167 \pm 0.014$ | $-1.126 \pm 0.017$ | $\mathbf{-0.981 \pm 0.018}$ | $-1.592 \pm 0.075$ |
| | | TANH | $-2.570 \pm 0.066$ | $-3.093 \pm 0.015$ | $-1.381 \pm 0.024$ | $\mathbf{-0.989 \pm 0.017}$ | $-1.100 \pm 0.109$ |
| | GP ‡ | LRELU | $-2.538 \pm 0.087$ | $\mathbf{-3.010 \pm 0.024}$ | $\mathbf{-0.627 \pm 0.063}$ | $\mathbf{-0.976 \pm 0.030}$ | $-0.785 \pm 0.099$ |
| | | TANH | $-2.594 \pm 0.121$ | $\mathbf{-3.015 \pm 0.037}$ | $-0.748 \pm 0.046$ | $\mathbf{-0.994 \pm 0.022}$ | $-0.697 \pm 0.077$ |
| | GP † | LRELU | $-2.604 \pm 0.140$ | $-3.449 \pm 0.005$ | $\mathbf{-0.660 \pm 0.036}$ | $\mathbf{-0.959 \pm 0.021}$ | $-0.886 \pm 0.083$ |
| | | TANH | $-2.533 \pm 0.126$ | $-3.452 \pm 0.007$ | $\mathbf{-0.633 \pm 0.041}$ | $\mathbf{-0.991 \pm 0.019}$ | $-0.723 \pm 0.072$ |
| BBB[1] | $*$ | RELU | $-2.602 \pm 0.031$ | $-3.149 \pm 0.018$ | $-1.500 \pm 0.006$ | $\mathbf{-0.977 \pm 0.017}$ | $-2.408 \pm 0.007$ |
| NNG[2] | $*$ | RELU | $-2.446 \pm 0.029$ | $\mathbf{-3.039 \pm 0.025}$ | $-1.421 \pm 0.005$ | $\mathbf{-0.969 \pm 0.014}$ | $-2.316 \pm 0.006$ |
| fBNN[3] | GP ‡ | RELU | $\mathbf{-2.301 \pm 0.038}$ | $-3.096 \pm 0.016$ | $\mathbf{-0.684 \pm 0.020}$ | $-1.040 \pm 0.013$ | $-1.033 \pm 0.033$ |

($a$) Log-Likelihood

| MODEL | | | BOSTON | CONCRETE | ENERGY | WINE | YACHT |
|---|---|---|---|---|---|---|---|
| GP | § | RBF | $2.977 \pm 0.249$ | $5.540 \pm 0.162$ | $0.472 \pm 0.021$ | $\mathbf{0.628 \pm 0.019}$ | $\mathbf{0.363 \pm 0.028}$ |
| NLM | § | LRELU | $3.055 \pm 0.178$ | $\mathbf{5.007 \pm 0.124}$ | $0.459 \pm 0.015$ | $\mathbf{0.649 \pm 0.010}$ | $0.922 \pm 0.094$ |
| | | TANH | $3.195 \pm 0.154$ | $\mathbf{5.127 \pm 0.133}$ | $0.523 \pm 0.017$ | $\mathbf{0.639 \pm 0.010}$ | $0.576 \pm 0.067$ |
| | WGN ‡ | LRELU | $3.010 \pm 0.228$ | $\mathbf{5.043 \pm 0.156}$ | $0.458 \pm 0.021$ | $\mathbf{0.641 \pm 0.013}$ | $1.189 \pm 0.137$ |
| | | TANH | $2.984 \pm 0.194$ | $\mathbf{4.897 \pm 0.156}$ | $0.544 \pm 0.026$ | $\mathbf{0.646 \pm 0.013}$ | $0.633 \pm 0.120$ |
| | GP ‡ | LRELU | $2.982 \pm 0.226$ | $\mathbf{4.833 \pm 0.140}$ | $\mathbf{0.437 \pm 0.020}$ | $\mathbf{0.638 \pm 0.017}$ | $0.502 \pm 0.053$ |
| | | TANH | $3.026 \pm 0.223$ | $\mathbf{4.828 \pm 0.211}$ | $0.503 \pm 0.019$ | $\mathbf{0.652 \pm 0.015}$ | $0.480 \pm 0.047$ |
| | GP † | LRELU | $2.962 \pm 0.246$ | $5.536 \pm 0.130$ | $0.463 \pm 0.018$ | $\mathbf{0.627 \pm 0.015}$ | $0.471 \pm 0.049$ |
| | | TANH | $2.830 \pm 0.255$ | $5.654 \pm 0.144$ | $0.453 \pm 0.018$ | $\mathbf{0.648 \pm 0.014}$ | $0.453 \pm 0.047$ |
| BBB[1] | $*$ | RELU | $3.171 \pm 0.149$ | $5.678 \pm 0.087$ | $0.565 \pm 0.018$ | $\mathbf{0.643 \pm 0.012}$ | $1.174 \pm 0.086$ |
| NNG[2] | $*$ | RELU | $2.742 \pm 0.125$ | $\mathbf{5.019 \pm 0.127}$ | $0.485 \pm 0.023$ | $\mathbf{0.637 \pm 0.011}$ | $0.979 \pm 0.077$ |
| fBNN[3] | GP ‡ | RELU | $\mathbf{2.378 \pm 0.104}$ | $\mathbf{4.935 \pm 0.180}$ | $\mathbf{0.412 \pm 0.017}$ | $0.673 \pm 0.014$ | $0.607 \pm 0.068$ |

($b$) RMSE

Table 1: Regression results on the UCI dataset. Models are trained with a type II maximum likelihood (§), ELBO ($*$), functional ELBO (‡) and functional KL (†) objective. Results obtained from 1, Blundell et al. (2015); 2, Zhang et al. (2018); 3, Sun et al. (2019).

the other Bayesian neural networks. The fKL objective with the GP posterior has performance that does indeed match the GP across most datasets. The NLM with fELBO training also generally outperforms the fBNN, apart from the boston dataset. Given that the fBNN has approximately twice as many parameters, a more elaborate training procedure, and an implicit predictive distribution, this indicates that the NLM is a better function class for functional prior research.

## 6. Related Work

The study of functional priors for Bayesian inference is a burgeoning field. Our work builds off Sun et al. (2019), who train a weight-space BNN with a GP posterior using a KL gradient estimator and mean-field stochastic variational inference. Shi et al. (2019) in turn build off this, taking a mirror descent interpretation to allow for incremental training with minibatches and the Gaussian process prior. As Gaussian processes are exact distributions over functions, sparse GPs may be viewed as approximate inference over functions

(de G. Matthews et al. (2016)), minimizing the fKL from its exact posterior via inducing points. The noise contrastive prior (NCP) of Hafner et al. (2019) is a similar idea where the training data is perturbed by random noise to serve as a 'data prior' for a BNN, in order to increase uncertainty estimation OOD. While effective empirically, the data prior is again akin to empirical Bayes as the prior is defined by the data. The practice of combining kernels in GPs has been translated to BNN architectures and activation functions, producing periodic and mixing phenomena in the network's feature space for more expressive models (Pearce et al. (2019)). Variational implicit priors (Ma et al. (2019)) use variational inference to deal with functional priors which you can only sample from, e.g. simulators, which provides the flexibility of a broad range of complex processes to be adopted as priors.

## 7. Outlook

The goal of this work is to motivate further research into the NLM by demonstrating its potential in combination with functional variational inference. Experiments have shown that this model is just as capable of representing a GP posterior as networks with weight priors, while being simpler and having closed-form predictive distributions. Our results suggest that an appropriate functional prior could enable GP-like performance and uncertainty quantification from NLMs without requiring explicit empirical Bayes.

## References

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, Berlin, Heidelberg, 2006.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, 2015.

G. E. P. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis*. John Wiley & Sons, New York, 1973.

Alexander G. de G. Matthews, James Hensman, Richard Turner, and Zoubin Ghahramani. On sparse variational methods and the kullback-leibler divergence between stochastic processes. In *International Conference on Artificial Intelligence and Statistics*, 2016.

Andrew Foong, Yingzhen Li, Jos Hernndez-Lobato, and Richard Turner. 'in-between' uncertainty in bayesian neural networks. In *ICML Workshop on Uncertainty and Robustness in Deep Learning*, 2019.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, 2016.

Jacob R. Gardner, Geoff Pleiss, David Bindel, Kilian Q. Weinberger, and Andrew Gordon Wilson. Gpytorch: Blackbox matrix-matrix gaussian process inference with gpu acceleration. In *Advances in Neural Information Processing Systems*, 2018.

Danijar Hafner, Dustin Tran, Alex Irpan, Timothy Lillicrap, and James Davidson. Noise contrastive priors for functional uncertainty. In *Uncertainty in Artificial Intelligence*, 2019.

B. Hunt. Prevalence: a translation-invariant almost every on infinite-dimensional spaces. *Bulletin of the American Mathematical Society*, 1992.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

Miguel Lázaro-Gredilla and Aníbal R. Figueiras-Vidal. Marginalized neural network mixtures for large-scale regression. *Transactions on Neural Networks*, 21(8), 2010.

Chao Ma, Yingzhen Li, and Jose Miguel Hernandez-Lobato. Variational implicit processes. In *International Conference on Machine Learning*, 2019.

Carl N Morris. Parametric empirical bayes inference: theory and applications. *Journal of the American statistical Association*, 78(381), 1983.

Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

Radford M. Neal. *Bayesian Learning for Neural Networks*. PhD thesis, University of Toronto, CAN, 1995.

Sebastian W. Ober and Carl Edward Rasmussen. Benchmarking the neural linear model for regression. In *Symposium on Advances in Approximate Bayesian Inference*, 2019.

Ian Osband, John Aslanides, and Albin Cassirer. Randomized prior functions for deep reinforcement learning. In *Advances in Neural Information Processing Systems*, 2018.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 2019.

Tim Pearce, Russell Tsuchida, Mohamed Zaki, Alexandra Brintrup, and Andy Neely. Expressive priors in bayesian neural networks: Kernel combinations and periodic functions. In *Uncertainty in Artificial Intelligence*, 2019.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2005.

Jiaxin Shi, Mohammad Emtiyaz Khan, and Jun Zhu. Scalable training of inference networks for Gaussian-process models. In *International Conference on Machine Learning*, 2019.

Shengyang Sun, Guodong Zhang, Jiaxin Shi, and Roger Grosse. Functional variational bayesian neural networks. In *International Conference on Learning Representations*, 2019.

Guodong Zhang, Shengyang Sun, David Duvenaud, and Roger Grosse. Noisy natural gradient as variational inference. In *International Conference on Machine Learning*, 2018.

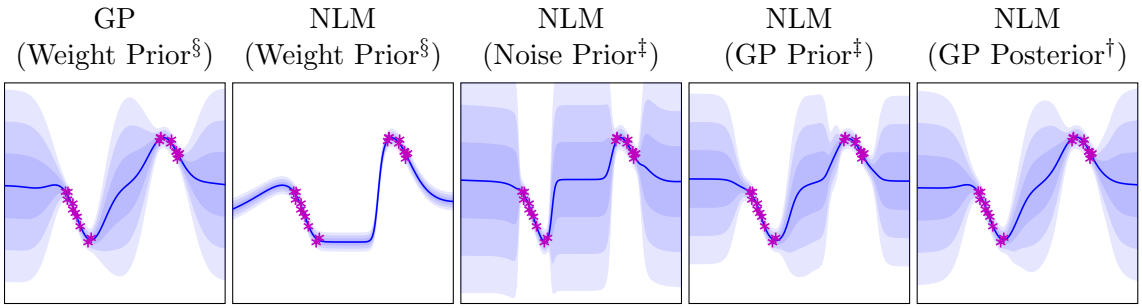


Figure 2: Reproduction of Figure 1 with tanh activation for the NLM.

## Appendix A. Bayesian Last Layer Equations

In this section, we collectively state the equations for Bayesian linear regression (Box and Tiao (1973); Bishop (2006); Murphy (2012)) with weights $\boldsymbol{\beta}$ and additive noise $\epsilon$, where

$$y_i = f(\mathbf{x}_i; \boldsymbol{\theta}) = \boldsymbol{\phi}_i^\top \boldsymbol{\beta} + \epsilon_i, \tag{7}$$

$$\epsilon_i \sim \mathcal{N}(\cdot \mid 0, \sigma^2), \tag{8}$$

$$\mathbf{y} \sim \mathcal{N}(\cdot \mid \boldsymbol{\Phi}\boldsymbol{\beta}, \sigma^2 \mathbf{I}). \tag{9}$$

Assuming known aleatoric uncertainty $\sigma^2$ (Bishop (2006)), a conjugate Gaussian prior over $\boldsymbol{\beta}$ results in a Gaussian posterior,

$$\boldsymbol{\beta} \sim \mathcal{N}(\cdot \mid \boldsymbol{\mu}_0, \boldsymbol{\Lambda}_0^{-1}), \qquad \boldsymbol{\mu}_n = \boldsymbol{\Lambda}_n^{-1}(\boldsymbol{\Lambda}_0 \boldsymbol{\mu}_0 + \sigma^{-2} \boldsymbol{\Phi}^\top \mathbf{y}), \tag{10}$$

$$\boldsymbol{\beta} \mid \mathcal{D}, \boldsymbol{\theta} \sim \mathcal{N}(\cdot \mid \boldsymbol{\mu}_n, \boldsymbol{\Lambda}_n^{-1}), \qquad \boldsymbol{\Lambda}_n = \sigma^{-2} \boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \boldsymbol{\Lambda}_0, \tag{11}$$

with an explicit Gaussian predictive distribution for output $y$ and query $\mathbf{x}$,

$$y \mid \mathbf{x}, \mathcal{D}, \boldsymbol{\theta} \sim \mathcal{N}(\cdot \mid \boldsymbol{\phi}_{\mathbf{x}}^\top \boldsymbol{\mu}_n, \sigma^2 + \boldsymbol{\phi}_{\mathbf{x}}^\top \boldsymbol{\Lambda}_n^{-1} \boldsymbol{\phi}_{\mathbf{x}}). \tag{12}$$

The log-marginal likelihood can be written as

$$\log p(\mathcal{D} \mid \boldsymbol{\theta}) = \frac{1}{2\sigma^2}(\boldsymbol{\mu}_n^\top \boldsymbol{\Lambda}_n \boldsymbol{\mu}_n - \boldsymbol{\mu}_0^\top \boldsymbol{\Lambda}_0 \boldsymbol{\mu}_0 - \mathbf{y}^\top \mathbf{y}) + \frac{1}{2} \log \frac{|\boldsymbol{\Lambda}_0|}{|\boldsymbol{\Lambda}_n|} - \frac{n}{2} \log 2\pi\sigma^2. \tag{13}$$

## Appendix B. Optimization Parameters

Gradient-based optimization was performed using Adam (Kingma and Ba (2014)). The GPs were trained until convergence for up to 3000 epochs at a learning rate of 1e-2. The NLMs were trained at a learning rate of 1e-3, and the number of training epochs was determined by tracking the validation log-likelihood of 10% of the training data for up to 10000 epochs for ML training and up to 15000 epochs for all fVI variants. The fKL was evaluated using 100 samples per epoch from a uniform distribution consistent with Sun et al. (2019).