2nd Symposium on Advances in Approximate Bayesian Inference, 2019 1–25

# Benchmarking the Neural Linear Model for Regression

**Sebastian W. Ober** swo25@cam.ac.uk
*University of Cambridge, Cambridge, UK*

**Carl Edward Rasmussen** cer54@cam.ac.uk
*University of Cambridge, Cambridge, UK*
*PROWLER.io, Cambridge, UK*

## Abstract

The neural linear model is a simple adaptive Bayesian linear regression method that has recently been used in a number of problems ranging from Bayesian optimization to reinforcement learning. Despite its apparent successes in these settings, to the best of our knowledge there has been no systematic exploration of its capabilities on simple regression tasks. In this work we characterize these on the UCI datasets, a popular benchmark for Bayesian regression models, as well as on the recently introduced UCI "gap" datasets, which are better tests of out-of-distribution uncertainty. We demonstrate that the neural linear model is a simple method that shows generally good performance on these tasks, but at the cost of requiring good hyperparameter tuning.

## 1. Introduction

Despite the recent successes that neural networks have shown in an impressive range of tasks, they tend to be overconfident in their predictions (Guo et al., 2017). Bayesian neural networks (BNNs; Neal (1995)) attempt to address this by providing a principled framework for uncertainty estimation in predictions. However, inference in BNNs is intractable to compute, requiring approximate inference techniques. Of these, Monte Carlo methods and variational methods, including Monte Carlo dropout (MCD) (Gal and Ghahramani, 2016), are popular; however, the former are difficult to tune, and the latter are often limited in their expressiveness (Foong et al., 2019b; Yao et al., 2019; Foong et al., 2019a).

The neural linear model represents a compromise between tractability and expressiveness for BNNs in regression settings: instead of attempting to perform approximate inference over the entire set of weights, it performs exact inference on only the last layer, where prediction can be done in closed form. It has recently been used in active learning (Pinsler et al., 2019), Bayesian optimization (Snoek et al., 2015), reinforcement learning (Riquelme et al., 2018), and AutoML (Zhou and Precioso, 2019), among others; however, to the best of our knowledge, there has been no systematic attempt to benchmark the model in the simple regression setting. In this work we do so, first demonstrating the model on a toy example, followed by experiments on the popular UCI datasets (as in Hernández-Lobato and Adams (2015)) and the recent UCI gap datasets from Foong et al. (2019b), who identified (along with Yao et al. (2019)) well-calibrated 'in-between' uncertainty as a desirable feature of BNNs.

© S.W. Ober & C.E. Rasmussen.

## 2. Methods

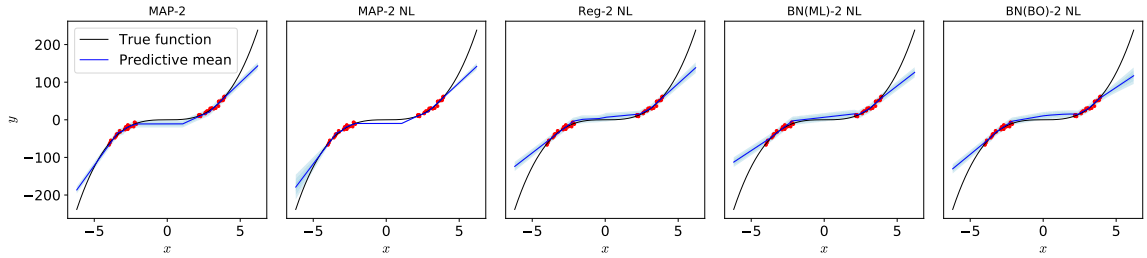


Figure 1: Predictive distributions for the toy problem. Each shaded region represents one predictive standard deviation.

In this section, we briefly describe the different models we train in this work, which are variations of the neural linear (NL) model, in which a neural network extracts features from the input to be used as basis functions for Bayesian linear regression. The central issue in the neural linear model is how to train the network: in this work, we provide three different models, with a total of four different training methods. For a more complete mathematical description of the models, refer to Appendix A; we summarize the models in Appendix C.

**MAP NL** Following the work of Snoek et al. (2015), we can first train the neural network using maximum a posteriori (MAP) estimation. After this training phase, the outputs of the last hidden layer of the network are used as the features for Bayesian linear regression. To reduce overfitting, the noise variance and prior variance (for the Bayesian linear regression) are subsequently marginalized out by slice sampling (Neal et al., 2003) according to the tractable marginal likelihood, using uniform priors. We refer to this model as the *maximum a posteriori neural linear model* (which we abbreviate as MAP-$L$ NL, where $L$ is the number of hidden layers in the network). We tune the hyperparameters for the MAP estimation via Bayesian optimization (Snoek et al., 2012).

**Regularized NL** The MAP NL model's basis functions are learned independently of the final model's predictions. This is an issue for uncertainty quantification, as MAP training has no incentive to learn features useful for providing uncertainty in out-of-distribution areas. To address this issue, we propose to learn the features by optimizing the (tractable) marginal likelihood with respect to the network weights (previous to the output layer), treating them as hyperparameters of the model in an approach analogous to hyperparameter optimization in Gaussian process (GP) regression (Rasmussen and Williams, 2006). However, unlike in GP regression, the per-iteration computational cost of this method is linear in the size of the data. We additionally regularize the weights to reduce overfitting, resulting in a model we call *regularized neural linear* (which we abbreviate as *Reg-L NL*). As in the MAP NL model, we marginalize out the noise and prior variances via slice sampling. We tune the regularization and other hyperparameters via Bayesian optimization.

**Bayesian noise NL** Instead of using slice sampling for the noise variance, we can place a normal-inverse-gamma (N-$\Gamma^{-1}$) prior on the weights and noise variance. This formulation is still tractable, and integrates the marginalization of the noise variance into the model itself, rather than having it implemented after the features are learned. Additionally, the N-$\Gamma^{-1}$ prior can act as a regularizer, meaning that we can avoid using Bayesian optimization to tune the prior parameters by jointly optimizing the marginal likelihood over all hyperparameters. However, this risks overfitting. Therefore, we consider training this model, which we call the *Bayesian noise (BN) neural linear* model, both by maximizing the marginal likelihood for all parameters (including prior parameters), and by tuning the prior parameters with Bayesian optimization. We abbreviate the first as *BN(ML)-L NL* and the second as *BN(BO)-L NL*. Finally, in both cases we slice sample the remaining (non-weight) hyperparameters.

## 3. Experiments

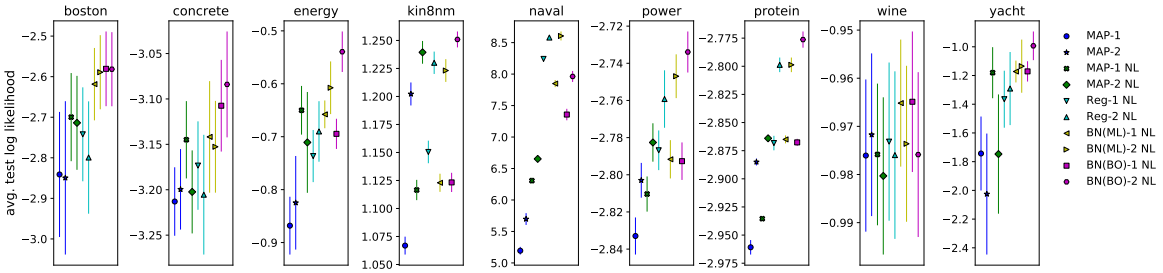


Figure 2: Average test log likelihoods (nats) for the UCI datasets

We compare these models on a toy problem, the UCI datasets, and the UCI "gap" datasets (Foong et al., 2019b). In all experiments, we consider 1- and 2-layer ReLU fully-connected networks with 50 hidden units in each layer (except for the toy problem, where we only consider 2-layer networks). We also provide results for simple MAP inference as a baseline. For experimental details, refer to Appendix B. We provide additional experimental results, including detailed statistical comparisons of the models, in Appendix D.

**Toy problem** We construct a synthetic 1-D dataset comprising 100 train and 100 test pairs $(x, y)$, where $x$ is sampled i.i.d. in the range $[-4, -2] \cup [2, 4]$ and $y$ is generated as $y = x^3 + \epsilon$, $\epsilon \sim \mathcal{N}(0, 9)$. This follows the example from Hernández-Lobato and Adams (2015), with the exception of the "gap" added in the range for $x$, which was motivated by Foong et al. (2019b) and Yao et al. (2019). We plot predictive distributions for each model in Figure 1. Somewhat surprisingly, the MAP-2 NL model seems to struggle more than MAP with uncertainty in the gap, while having better uncertainty quantification at the edges. Of the marginal likelihood-based methods, the BN(BO)-2 NL model qualitatively seems to perform the best.

**UCI datasets** We next provide results on the UCI datasets in Hernández-Lobato and Adams (2015) (omitting the 'year' dataset due to its size), a popular benchmark for Bayesian regression models in recent years. We report average test log likelihoods and RMSEs for

all the models in Appendix D.1, for both 1- and 2-layer architectures. We visualize average test log likelihoods for the models in Figure 2; we tabulate the log likelihoods and RMSEs in Tables 2 and 3 in Appendix D.1, respectively.

From the figure and tables, we see that the BN(ML)-2 NL and BN(BO)-2 NL models have the best performance on these metrics, with reasonable log likelihoods and RMSEs compared to those in the literature for other BNN-based methods (Hernández-Lobato and Adams, 2015; Gal and Ghahramani, 2016; Bui et al., 2016; Tomczak et al.). In fact, these neural linear methods tend to achieve state-of-the-art or near state-of-the-art neural network performance on the 'energy' and 'naval' datasets. While the performance of the Reg-$L$ NL model is decent, it performs worse than the BN-$L$ NL models, showing the advantage of a Bayesian treatment of the noise variance.

**UCI gap datasets** Finally, we provide results on the UCI "gap" datasets proposed by Foong et al. (2019b), which consists of training and testing splits that artificially contain gaps in the training set, ensuring that the model will only succeed if it can represent uncertainty in-between gaps in the data. We again visualize test log likelihoods in Figure 3 while tabulating log likelihoods and RMSEs in Tables 6 and 7 in Appendix D.1.

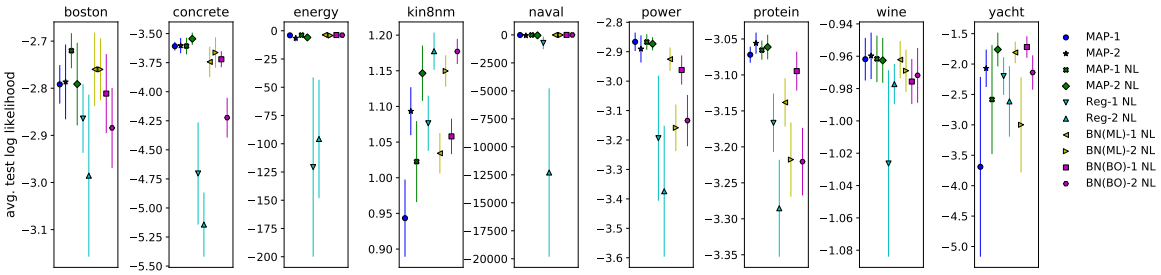


Figure 3: Average test log likelihoods (nats) for the UCI gap datasets

Our results on the MAP-based models in Figure 3 echo those of Foong et al. (2019b), showing catastrophic failure to express in-between uncertainty for some datasets (particularly 'energy' and 'naval'). Somewhat surprisingly, the Reg-$L$ NL models perform the worst of all the models. However, the BN NL models do not seem to fail catastrophically, with the BN(BO)-2 NL model having by far the best performance.

## 4. The Effect of Hyperparameter Tuning

In all of these models we used some form of hyperparameter tuning (Bayesian optimization for all models except the BN(ML)-$L$ NL models, where we used a grid search) to obtain the results shown. However, for the practitioner, performing an oftentimes costly hyperparameter search is not desirable, particularly where one of the main motivations for using the model is its simplicity, as in this case. We therefore investigate the effect of the hyperparameter tuning on the models' performance.

Figure 4 shows the difference in average test log likelihoods and test RMSEs between the tuned models and models whose hyperparameters were set to "reasonable" values that a practitioner might choose by intuition (see Appendix D.2 for details) for the UCI datasets.

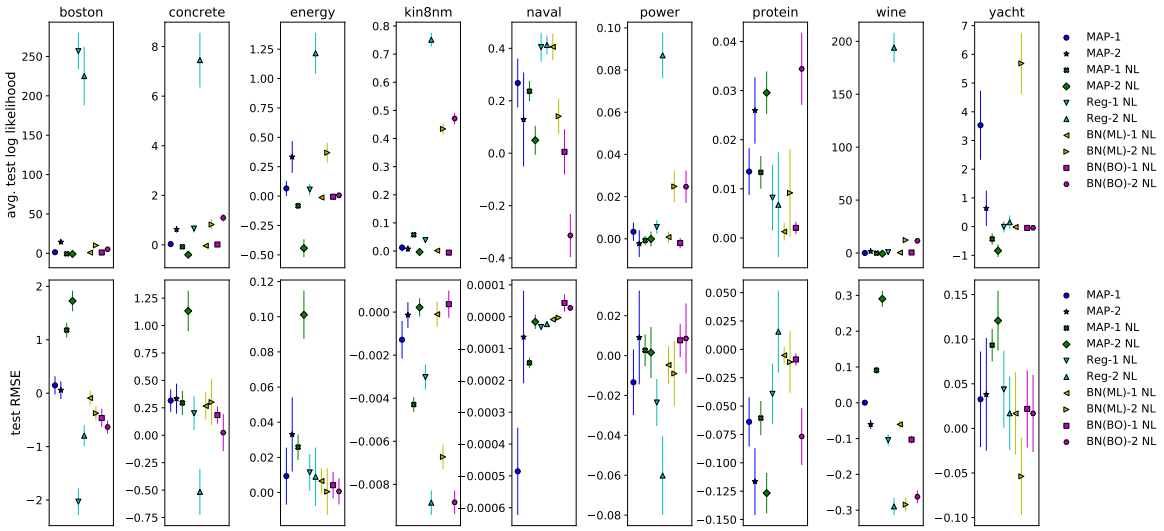


Figure 4: Differences in average test log likelihoods (nats) and RMSEs given by hyperparameter tuning for the UCI datasets

We observe that for each of the two-layer models there exists at least one dataset where the performance in terms of test log likelihood is *significantly* worsened by omitting hyperparameter tuning. The performance difference for RMSEs is not as drastic, although it still exists. In Appendix D.2 we show that these results extend to the UCI gap datasets and that the difference in performance is statistically significant for nearly all models across both the UCI and UCI gap datasets, for both log likelihood and RMSE performance. Finally, in Appendix D.2.1 we show that mean field variational inference (MFVI) (Graves, 2011; Hinton and Van Camp, 1993; Blundell et al., 2015) and MCD can still obtain reasonable, although not state-of-the-art, performance on the UCI datasets without hyperparameter tuning: in many cases the performance is even competitive with the tuned NL models. However, these suffer from the pathologies identified in Foong et al. (2019b); Yao et al. (2019); Foong et al. (2019a) on the gap datasets.

## 5. Conclusions & Future Work

We have shown benchmark results for different variants of the neural linear model in the regression setting. Our results show that the successes these models have seen in other areas such as reinforcement and active learning are not unmerited, with the models achieving generally good performance despite their simplicity. Furthermore, they are not as susceptible to the the inability to express gap uncertainty as MFVI or MCD. However, we have shown that to obtain reasonable performance extensive hyperparameter tuning is often required, unlike MFVI or MCD. Finally, our work suggests that exact inference on a subset of parameters can perform better than approximate inference on the entire set, at least for BNNs. We believe this broader issue is worthy of further investigation.

## Acknowledgments

We thank Andrew Y.K. Foong, David R. Burt, Ross M. Clarke, and Thang D. Bui for helpful discussions. We additionally thank Andrew Y.K. Foong for providing the "gap" splits. SWO acknowledges the support of the Gates Cambridge Trust for his doctoral studies.

## References

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. *arXiv preprint arXiv:1505.05424*, 2015.

Thang Bui, Daniel Hernández-Lobato, José Miguel Hernández-Lobato, Yingzhen Li, and Richard Turner. Deep Gaussian processes for regression using approximate expectation propagation. In *International Conference on Machine Learning*, pages 1472–1481, 2016.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

Andrew YK Foong, David R Burt, Yingzhen Li, and Richard E Turner. Pathologies of factorised Gaussian and MC dropout posteriors in Bayesian neural networks. *arXiv preprint arXiv:1909.00719*, 2019a.

Andrew YK Foong, Yingzhen Li, José Miguel Hernández-Lobato, and Richard E Turner. 'In-between' uncertainty in Bayesian neural networks. *arXiv preprint arXiv:1906.11537*, 2019b.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059, 2016.

Alex Graves. Practical variational inference for neural networks. In *Advances in neural information processing systems*, pages 2348–2356, 2011.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1321–1330. JMLR. org, 2017.

José Miguel Hernández-Lobato and Ryan Adams. Probabilistic backpropagation for scalable learning of Bayesian neural networks. In *International Conference on Machine Learning*, pages 1861–1869, 2015.

Geoffrey Hinton and Drew Van Camp. Keeping neural networks simple by minimizing the description length of the weights. In *in Proc. of the 6th Ann. ACM Conf. on Computational Learning Theory*. Citeseer, 1993.

Diederik P Kingma and Jimmy Ba. ADAM: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. In *Advances in Neural Information Processing Systems*, pages 2575–2583, 2015.

Radford M Neal. *Bayesian learning for neural networks*. 1995.

Radford M Neal et al. Slice sampling. *The Annals of Statistics*, 31(3):705–767, 2003.

Robert Pinsler, Jonathan Gordon, Eric Nalisnick, and José Miguel Hernández-Lobato. Bayesian batch active learning as sparse subset approximation. *arXiv preprint arXiv:1908.02144*, 2019.

Carl Edward Rasmussen and Christopher KI Williams. Gaussian Processes for Machine Learning. *ISBN-13 978-0-262-18253-9*, 2006.

Carlos Riquelme, George Tucker, and Jasper Snoek. Deep Bayesian bandits showdown: An empirical comparison of Bayesian deep networks for Thompson sampling. *arXiv preprint arXiv:1802.09127*, 2018.

Jasper Snoek, Hugo Larochelle, and Ryan P Adams. Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, pages 2951–2959, 2012.

Jasper Snoek, Oren Rippel, Kevin Swersky, Ryan Kiros, Nadathur Satish, Narayanan Sundaram, Mostofa Patwary, Mr Prabhat, and Ryan Adams. Scalable Bayesian optimization using deep neural networks. In *International Conference on Machine Learning*, pages 2171–2180, 2015.

Marcin B Tomczak, Siddharth Swaroop, and Richard E Turner. Neural network ensembles and variational inference revisited.

Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer, 1992.

Jiayu Yao, Weiwei Pan, Soumya Ghosh, and Finale Doshi-Velez. Quality of uncertainty quantification for Bayesian neural network inference. *arXiv preprint arXiv:1906.09686*, 2019.

Weilin Zhou and Frederic Precioso. Adaptive Bayesian linear regression for automated machine learning. *arXiv preprint arXiv:1904.00577*, 2019.

## Appendix A. The Neural Linear Model

The neural linear model uses a neural network to parameterize basis functions for Bayesian linear regression by treating the output weights and bias of the network probabilistically, while treating the rest of the network's parameters $\theta$ as hyperparameters. This can be used as an approximation to full Bayesian inference of the neural network's parameters, with the main advantage being that this simplified case is tractable (assuming Gaussian prior and likelihood). Given the fact that there are significant redundancies in the weight-space posterior for BNNs, this tradeoff may not be a completely unreasonable approximation.

We now describe the model mathematically. Let $\mathcal{D} \coloneqq (X, Y) = \{(x_n, y_n)\}_{n=1}^N$, where $(x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$, be the training data, and let

$$\phi_\theta(x_n) = [\phi_{\theta,1}(x_n), \ldots, \phi_{\theta,N_L}(x_n)]^T$$

represent the outputs (post-activations) of the last hidden layer of the neural network, which will be parameterized by all the weights and biases up to the last layer, $\theta$. We then define a weight vector $w \in \mathbb{R}^M = \mathbb{R}^{N_L+1}$ (this includes a bias term, augmenting $\phi_\theta(x)$ with a 1). If we define a design matrix $\Phi_\theta = [\phi_\theta(x_1), \ldots, \phi_\theta(x_N)]^T$, we can then define our model as

$$Y = \Phi_\theta w + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2),$$

where we treat $Y$ as a column vector of the $y_n$. Given an appropriate $\theta$, Bayesian inference of the weights $w$ is straightforward: given a prior $p(w) = \mathcal{N}(w;\, 0,\, \alpha \mathrm{I}_M)$ on the weights, the posterior is given by

$$\begin{aligned} p(w|\mathcal{D}, \sigma^2) &= \mathcal{N}(w|w_N, \mathrm{V}_N) \propto \mathcal{N}(w;\, 0,\, \alpha \mathrm{I}_M)\, \mathcal{N}(Y;\, \Phi_\theta w,\, \sigma^2), \\ w_N &= \frac{1}{\sigma^2} \mathrm{V}_N \Phi_\theta^T Y, \\ \mathrm{V}_N^{-1} &= \frac{1}{\alpha} \mathrm{I}_M + \frac{1}{\sigma^2} \Phi_\theta^T \Phi_\theta. \end{aligned}$$

The posterior predictive for a test input $x_*$ is then given by

$$p(y_*|x_*, \mathcal{D}, \sigma^2) = \mathcal{N}\left(y_*; w_N^T \phi_\theta(x_*),\, \sigma^2 + \phi_\theta(x_*)^T \mathrm{V}_N \phi_\theta(x_*)\right).$$

It now remains to be determined how to learn $\theta$.

### A.1. MAP Neural Linear

As described in Snoek et al. (2015), we can learn $\theta$ by simply setting it to the values of the corresponding weights and biases in a maximum a posteriori (MAP)-trained network, maximizing the objective

$$\mathcal{L}_{MAP}(\theta_{Full}) = \log \mathcal{N}(Y;\, \Phi_\theta w,\, \sigma^2) - \gamma \|\theta_{Full}\|_2^2,$$

with respect to $\theta_{Full}$ and $\sigma^2$, where $\theta_{Full}$ represents the parameters of the full network (which includes the output weights and bias), and $\gamma$ is a regularization parameter. As in Snoek et al. (2015), once we have obtained $\theta$ from $\theta_{Full}$, we use can use Bayesian linear regression as outlined above. However, the question of setting $\alpha$ still remains. To address

this, we marginalize $\alpha$ and $\sigma^2$ out by slice sampling them according to the log marginal likelihood of the data:

$$\begin{aligned}\mathcal{L}_{\theta,\alpha,\sigma^2}(\mathcal{D}) &= \log p(Y|X,\sigma^2) = \log \int p(Y|X,w,\sigma^2)\, p(w)\, \mathrm{d}w \\ &= -\frac{M}{2}\log\alpha - \frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}||Y - \Phi_\theta w_N||_2^2 - \frac{1}{2\alpha} w_N^T w_N \\ &\quad - \frac{1}{2}\log|\mathrm{V}_N| - \frac{N}{2}\log 2\pi.\end{aligned}$$

In order to learn a suitable value of $\gamma$, along with learning rates and number of epochs, we use Bayesian optimization. For a complete description of the experimental details, see Appendix B.

One key disadvantage of this approach is that it separates the feature learning from prediction: in particular, there is no reason for the network to learn features relevant for out-of-distribution prediction, particularly when it comes to uncertainty estimates.

### A.2. Regularized Neural Linear

From a Bayesian perspective, the neural linear model can be interpreted as a Gaussian process model with a covariance kernel determined by a finite number of basis functions $\phi_{\theta,i}$ with hyperparameters $\theta$. Therefore, as in Gaussian process regression, we propose to maximize the log marginal likelihood of the data, $\mathcal{L}_{\theta,\alpha,\sigma^2}(\mathcal{D})$, with respect to $\theta$ and $\sigma$ as the hyperparameters of the model for an empirical Bayes approach.

Note that the computational complexity of this expression is $\mathcal{O}(N + M^3)$, as opposed to the $\mathcal{O}(N^3)$ cost typically seen in GP regression. This is because we are able to apply the Woodbury identity to obtain the determinant in terms of $\mathrm{V}_N$, which is $M \times M$, due to the fact that there is a finite number of basis functions. Since we typically have that $N \gg M$, this results in significant computational savings.

One issue with this Type-2 maximum likelihood approach is that it will tend to overfit to the training data due to the large number of hyperparameters $\theta$. As a result, the noise variance $\sigma^2$ will tend to be pushed towards zero. One way of addressing this is by introducing a regularization scheme. There are many potential regularization schemes that could be introduced: we could regularize $\theta$, $\alpha$, or $\sigma$ individually, or using any combination of the three. We found empirically that of these, simply regularizing $\theta$ alone via $L^2$ regularization seemed the most promising approach. This results in a Type-2 MAP approach wherein we maximize

$$\hat{\mathcal{L}}_{\theta,\alpha,\sigma^2}(\mathcal{D}) = \mathcal{L}_{\theta,\alpha,\sigma^2}(\mathcal{D}) - \gamma_W\|\theta_W\|_2^2 - \gamma_b\|\theta_b\|_2^2,$$

where we have divided $\theta$ into weights $\theta_W$ and biases $\theta_b$ and introduced regularization hyperparameters $\gamma_W$ and $\gamma_b$.

### A.3. Bayesian Noise Neural Linear

An alternative to regularization would be to treat the noise variance in a Bayesian manner by integrating it out. Fortunately, for Bayesian linear regression this is still tractable with

the use of a normal-inverse-gamma prior on the outputs weights and parameters

$$\begin{aligned} p(w, \sigma^2) &= \text{N-}\Gamma^{-1}(w, \sigma^2;\, 0,\, \alpha \mathrm{I}_M,\, a_0,\, b_0) \\ &= \mathcal{N}(w;\, 0,\, \sigma^2 \alpha \mathrm{I}_M)\Gamma^{-1}(\sigma^2;\, a_0,\, b_0). \end{aligned}$$

The posterior has the form

$$\begin{aligned} p(w, \sigma^2 | \mathcal{D}) &= \text{N-}\Gamma^{-1}(w, \sigma^2; w_N, \mathrm{V}_N, a_N, b_N), \\ w_N &= \mathrm{V}_N \Phi_\theta^T Y, \\ \mathrm{V}_N^{-1} &= \frac{1}{\alpha}\mathrm{I}_M + \Phi_\theta^T \Phi_\theta, \\ a_N &= a_0 + N/2, \\ b_N &= b_0 + \frac{1}{2}(Y^T Y - w_N^T V_N^{-1} w_N), \end{aligned}$$

with posterior predictive

$$p(y_* | x_*, \mathcal{D}) = \mathcal{T}\left(y_*;\, w_N^T \phi_\theta(x_*),\, \frac{b_N}{a_N}(\mathrm{I}_M + \phi_\theta(x_*)^T \mathrm{V}_N \phi_\theta(x_*)),\, 2a_N\right),$$

where $\mathcal{T}(\,\cdot\,;\, \mu,\, \Sigma,\, \nu)$ is a Student's t-distribution with mean $\mu$, scale $\Sigma$, and degrees of freedom $\nu$. As before, we train the network using empirical Bayes, where the marginal likelihood is given by

$$p(Y|X) = \mathcal{T}\left(Y;\, 0,\, \frac{b_0}{a_0}(\mathrm{I}_M + \alpha \Phi_\theta \Phi_\theta^T),\, 2a_0\right).$$

Note that by using the Woodbury identity it is possible to compute this in $\mathcal{O}(N + M^3)$ computational cost as before.

## Appendix B. Experimental Details

All neural networks tested were ReLU networks with one or two 50-unit hidden layers. When using a validation set, we set its size to be one fifth of the size of the training set, except for the toy example, where we used half the training set. We now describe the experimental setup for each model we used.

**MAP** For the MAP baseline, we select a batch size of 32. We subsequently use Bayesian optimization (see section B.1 for a description of the Bayesian optimization algorithm we use) to optimize four hyperparameters using validation log likelihood: the regularization parameter $\gamma$, a learning rate for the weights, a learning rate for the noise variance, and the number of epochs. The regularization parameter is allowed to vary within the range corresponding to a log prior variance between -5 and 5. The learning rates are also optimized in log space in the range $[\log \text{1e-4}, \log 1e-2]$. Finally, the number of epochs is set to vary between zero and the number required to obtain at least 10000 gradient steps (the number of epochs will thus vary with the size of the dataset given a constant batch size). We initialize the regularization parameter to 0.5, the learning rates at 1e-3, the noise variance at $e^{-3}$, and the number of epochs at the maximum value. The network itself is optimized using ADAM (Kingma and Ba, 2014).

**MAP NL** For the MAP neural linear model, we take the above optimal MAP network and obtain 200 slice samples of $\alpha_W$ (the output weight prior variance), $\alpha_b$ (the output bias prior variance), and $\sigma^2$ for Bayesian linear regression. We initialize $\alpha_W = 1/50$ and $\alpha_b = 1$, to match the scaling used in Neal (1995).

**Regularized NL** For the regularized NL model, there are five hyperparameters which we tune via Bayesian optimization: $\gamma_W$, $\gamma_b$, a learning rate for $\theta$, a learning rate for $\sigma^2$, and the number of epochs. We allow $\gamma_W$ and $\gamma_b$ to vary within a range of log prior variances between -10 and 10, and the number of epochs to be in the range of $[0, 5000]$ (since each epoch corresponds to one gradient step). The ranges for the other parameters remain the same. We initialize $\gamma_W$ and $\gamma_b$ to 1, and the remaining parameters the same way as in the MAP model. We again initialize $\alpha_W = 1/50$ and $\alpha_b = 1$. As before, we use 200 slice samples to marginalize out $\sigma^2$, $\alpha_W$, and $\alpha_b$ after the Bayesian optimization was completed.

**Bayesian noise NL (ML)** Here we optimize the parameters $\theta$, $a_0$, $b_0$, $\alpha_W$, and $\alpha_b$ directly and jointly via the log marginal likelihood. We employ early stopping by tracking the validation log likelihood up to 5000 epochs, and also maximize the validation log likelihood over a grid of 10 learning rates ranging logarithmically from $\log$ 1e-4 to $\log$ 1e-2. We also initialize $a_0 = b_0 = 1$ and $\alpha_W = \alpha_b = 1$. Finally, we use slice sampling to obtain 200 samples to marginalize out these hyperparameters.

**Bayesian noise NL (BO)** Instead of optimizing over the hyperparameters jointly as in the BN(ML) model, we keep all except $\theta$ fixed over each iteration of Bayesian optimization. We retain the same initializations, and allow the following ranges for the hyperparameters: $a_0 \in [0, 20]$, $b_0 \in [0, 10]$, $\log \alpha_W \in [-10, 10]$, $\log \alpha_b \in [-10, 10]$, with the ranges for the learning rate and number of epochs being the same as before. We retain the same initializations as before as well. The slice sampling also remains the same.

### B.1. Bayesian Optimization

Here we describe the Bayesian optimization algorithm that we used throughout. In each case we attempt to maximize the validation log likelihood. We largely follow the formulation set out in Snoek et al. (2012). We use a Gaussian process with a Matérn-5/2 kernel with the model hyperparameters as inputs and the validation log likelihoods as outputs (normalizing the inputs and outputs). We first learn the kernel hyperparameters (including a noise variance) by maximizing the marginal likelihood of the GP, using 5000 iterations of ADAM (Kingma and Ba, 2014) with a learning rate of 1e-2. We then obtain 20 slice samples of the GP hyperparameters, before using the expected improvement acquisition function to find the next set of network hyperparameters to test. In total, we use 50 iterations of Bayesian optimization for each model, initialized with 10 iterations of random search.

## Appendix C. Summary of the Models

In Table 1, we provide a summary of the models we use, describing which parameters are optimized and how (we exclude learning rates and the number of epochs from this table), and which are then slice sampled.

| Model | Opt. objective | Params optimized | Bayes Opt. | Params slice sampled |
|---|---|---|---|---|
| **MAP-$L$** | MAP | $\theta$, $\sigma^2$ | $\gamma$ | None |
| **MAP-$L$ NL** | MAP | $\theta$, $\sigma^2$ | $\gamma$ | $\alpha_W$, $\alpha_b$, $\sigma^2$ |
| **Reg-$L$ NL** | Reg. ML | $\theta$, $\alpha_W$, $\alpha_b$, $\sigma^2$ | $\gamma_W$, $\gamma_b$ | $\alpha_W$, $\alpha_b$, $\sigma^2$ |
| **BN(ML)-$L$ NL** | ML | $\theta$, $\alpha_w$, $\alpha_b$, $a_0$, $b_0$ | None | $\alpha_W$, $\alpha_b$, $a_0$, $b_0$ |
| **BN(BO)-$L$ NL** | ML | $\theta$ | $\alpha_W$, $\alpha_b$, $a_0$, $b_0$ | $\alpha_W$, $\alpha_b$, $a_0$, $b_0$ |

Table 1: Summary of the models presented. The first column lists the model; the second shows the optimization objective, while the third shows which parameters were optimized using this objective. Meanwhile, the fourth lists the parameters that were tuned using Bayesian optimization, while the final lists the parameters that slice sampling was performed on.

## Appendix D. Additional Experimental Results

In this appendix, we provide the full results from the main text, before briefly describing empirically the effect of slice sampling on the models.

### D.1. Tabulated Results and Average Rank Analysis

On the next pages, we present tables of average test log likelihoods and test RMSEs for the UCI and UCI gap datasets for all models. For the UCI datasets, we present the average test log likelihoods and test RMSEs in Tables 2 and 3 , as well as train log likelihoods and RMSEs in Tables 4 and 5 . Following Bui et al. (2016), we also compute average ranks of the models across all splits of the standard UCI datasets. As in Bui et al. (2016), we additionally follow the procedure for the Friedman test as described in Demšar (2006), generating the plots shown in Figures 5 and 6. These plots show the average rank of each method across all splits, where the difference between models is not statistically significant ($p < 0.05$) if the models are connected by a dark line, which is determined by the *critical difference (CD)*.

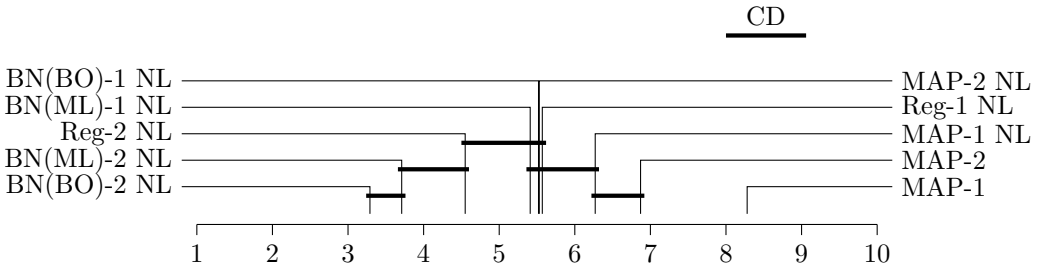


Figure 5: Average ranks on the UCI datasets according to average test log likelihoods, generated as described in Demšar (2006).

| Dataset | MAP-1 | MAP-2 | MAP-1 NL | MAP-2 NL | Reg-1 NL | Reg-2 NL | BN(ML)-1 NL | BN(ML)-2 NL | BN(BO)-1 NL | BN(BO)-2 NL |
|---|---|---|---|---|---|---|---|---|---|---|
| boston | -2.84 ± 0.15 | -2.85 ± 0.19 | -2.70 ± 0.11 | -2.71 ± 0.11 | -2.74 ± 0.11 | -2.80 ± 0.14 | -2.62 ± 0.09 | -2.59 ± 0.09 | **−2.58 ± 0.09** | **−2.58 ± 0.09** |
| concrete | -3.21 ± 0.04 | -3.20 ± 0.04 | -3.14 ± 0.04 | -3.20 ± 0.05 | -3.17 ± 0.05 | -3.21 ± 0.07 | -3.14 ± 0.06 | -3.15 ± 0.05 | -3.11 ± 0.05 | **−3.08 ± 0.06** |
| energy | -0.87 ± 0.05 | -0.82 ± 0.09 | -0.65 ± 0.05 | -0.71 ± 0.09 | -0.74 ± 0.05 | -0.69 ± 0.06 | -0.66 ± 0.03 | -0.61 ± 0.05 | -0.69 ± 0.03 | **−0.54 ± 0.04** |
| kin8nm | 1.07 ± 0.01 | 1.20 ± 0.01 | 1.12 ± 0.01 | 1.24 ± 0.01 | 1.15 ± 0.01 | 1.23 ± 0.01 | 1.12 ± 0.01 | 1.22 ± 0.01 | 1.12 ± 0.01 | **1.25 ± 0.01** |
| naval | 5.19 ± 0.05 | 5.70 ± 0.09 | 6.31 ± 0.04 | 6.65 ± 0.04 | 8.24 ± 0.04 | 8.58 ± 0.03 | 7.85 ± 0.05 | **8.61 ± 0.06** | 7.36 ± 0.08 | 7.96 ± 0.08 |
| power | -2.83 ± 0.01 | -2.80 ± 0.01 | -2.81 ± 0.01 | -2.78 ± 0.01 | -2.79 ± 0.01 | -2.76 ± 0.02 | -2.79 ± 0.01 | -2.75 ± 0.01 | -2.79 ± 0.01 | **−2.73 ± 0.01** |
| protein | -2.96 ± 0.01 | -2.89 ± 0.00 | -2.94 ± 0.00 | -2.86 ± 0.00 | -2.87 ± 0.01 | -2.80 ± 0.01 | -2.86 ± 0.00 | -2.80 ± 0.01 | -2.87 ± 0.00 | **−2.78 ± 0.01** |
| wine | -0.98 ± 0.02 | -0.97 ± 0.02 | -0.98 ± 0.01 | -0.98 ± 0.02 | -0.97 ± 0.02 | -0.98 ± 0.02 | -0.97 ± 0.01 | -0.97 ± 0.02 | **−0.96 ± 0.01** | -0.98 ± 0.02 |
| yacht | -1.74 ± 0.25 | -2.03 ± 0.42 | -1.18 ± 0.17 | -1.75 ± 0.41 | -1.37 ± 0.20 | -1.29 ± 0.25 | -1.17 ± 0.07 | -1.13 ± 0.18 | -1.17 ± 0.07 | **−0.99 ± 0.10** |
| **Avg. Rank** | 8.28 ± 0.18 | 6.87 ± 0.19 | 6.27 ± 0.20 | 5.53 ± 0.19 | 5.57 ± 0.19 | 4.55 ± 0.24 | 5.41 ± 0.19 | 3.71 ± 0.19 | 5.52 ± 0.20 | **3.29 ± 0.19** |

Table 2: Average Test Log Likelihoods (nats) on the UCI Datasets (Best Values in Bold, Errors are ±1 Standard Error)

| Dataset | MAP-1 | MAP-2 | MAP-1 NL | MAP-2 NL | Reg-1 NL | Reg-2 NL | BN(ML)-1 NL | BN(ML)-2 NL | BN(BO)-1 NL | BN(BO)-2 NL |
|---|---|---|---|---|---|---|---|---|---|---|
| boston | 3.28 ± 0.23 | 3.25 ± 0.24 | 3.16 ± 0.19 | 3.17 ± 0.22 | 3.18 ± 0.22 | 3.20 ± 0.26 | 3.13 ± 0.21 | 3.05 ± 0.23 | 2.97 ± 0.19 | **2.91 ± 0.19** |
| concrete | 5.41 ± 0.12 | 5.13 ± 0.12 | 5.14 ± 0.13 | 5.05 ± 0.11 | 5.03 ± 0.16 | 4.82 ± 0.14 | 5.08 ± 0.13 | 5.17 ± 0.12 | 4.96 ± 0.15 | **4.78 ± 0.19** |
| energy | 0.52 ± 0.02 | 0.47 ± 0.02 | 0.44 ± 0.01 | 0.42 ± 0.02 | 0.46 ± 0.01 | 0.43 ± 0.02 | 0.46 ± 0.01 | 0.42 ± 0.01 | 0.48 ± 0.01 | **0.40 ± 0.01** |
| kin8nm | 0.08 ± 0.00 | **0.07 ± 0.00** | 0.08 ± 0.00 | **0.07 ± 0.00** | 0.08 ± 0.00 | **0.07 ± 0.00** | 0.08 ± 0.00 | **0.07 ± 0.00** | 0.08 ± 0.00 | **0.07 ± 0.00** |
| naval | **0.00 ± 0.00** | **0.00 ± 0.00** | **0.00 ± 0.00** | **0.00 ± 0.00** | **0.00 ± 0.00** | **0.00 ± 0.00** | **0.00 ± 0.00** | **0.00 ± 0.00** | **0.00 ± 0.00** | **0.00 ± 0.00** |
| power | 4.11 ± 0.04 | 3.99 ± 0.03 | 4.01 ± 0.04 | 3.90 ± 0.04 | 3.91 ± 0.04 | 3.74 ± 0.04 | 3.94 ± 0.04 | 3.73 ± 0.04 | 3.94 ± 0.04 | **3.70 ± 0.04** |
| protein | 4.67 ± 0.03 | 4.33 ± 0.01 | 4.56 ± 0.01 | 4.24 ± 0.01 | 4.25 ± 0.02 | 3.94 ± 0.02 | 4.24 ± 0.01 | 3.94 ± 0.02 | 4.25 ± 0.01 | **3.88 ± 0.02** |
| wine | 0.64 ± 0.01 | **0.63 ± 0.01** | 0.64 ± 0.01 | **0.63 ± 0.01** | 0.64 ± 0.01 | **0.63 ± 0.01** | **0.63 ± 0.01** | **0.63 ± 0.01** | **0.63 ± 0.01** | **0.63 ± 0.01** |
| yacht | 0.73 ± 0.06 | 0.66 ± 0.06 | 0.61 ± 0.05 | 0.63 ± 0.05 | 0.64 ± 0.04 | 0.58 ± 0.06 | 0.79 ± 0.06 | **0.55 ± 0.05** | 0.77 ± 0.06 | 0.66 ± 0.06 |
| **Avg. Rank** | 8.36 ± 0.18 | 6.40 ± 0.20 | 6.73 ± 0.20 | 4.95 ± 0.18 | 5.47 ± 0.19 | 3.62 ± 0.21 | 6.13 ± 0.17 | 3.77 ± 0.18 | 6.11 ± 0.20 | **3.45 ± 0.20** |

Table 3: Test RMSEs on the UCI Datasets (Best Values in Bold, Errors are ±1 Standard Error)

| Dataset | MAP-1 | MAP-2 | MAP-1 NL | MAP-2 NL | Reg-1 NL | Reg-2 NL | BN(ML)-1 NL | BN(ML)-2 NL | BN(BO)-1 NL | BN(BO)-2 NL |
|---|---|---|---|---|---|---|---|---|---|---|
| boston | -2.31 ± 0.05 | -2.25 ± 0.07 | -2.22 ± 0.05 | -2.14 ± 0.08 | -2.11 ± 0.09 | -2.00 ± 0.09 | -2.16 ± 0.06 | -2.20 ± 0.05 | -1.93 ± 0.04 | -1.77 ± 0.08 |
| concrete | -2.81 ± 0.03 | -2.65 ± 0.04 | -2.69 ± 0.03 | -2.55 ± 0.05 | -2.61 ± 0.04 | -2.21 ± 0.07 | -2.57 ± 0.04 | -2.62 ± 0.04 | -2.47 ± 0.03 | -2.15 ± 0.05 |
| energy | -0.49 ± 0.04 | -0.30 ± 0.05 | -0.20 ± 0.02 | -0.01 ± 0.05 | -0.29 ± 0.04 | 0.10 ± 0.06 | -0.33 ± 0.02 | 0.14 ± 0.04 | -0.43 ± 0.02 | 0.29 ± 0.06 |
| kin8nm | 1.15 ± 0.00 | 1.36 ± 0.01 | 1.22 ± 0.00 | 1.43 ± 0.01 | 1.29 ± 0.01 | 1.47 ± 0.01 | 1.25 ± 0.00 | 1.43 ± 0.01 | 1.25 ± 0.00 | 1.53 ± 0.01 |
| naval | 5.20 ± 0.05 | 5.70 ± 0.09 | 6.33 ± 0.04 | 6.68 ± 0.04 | 8.33 ± 0.04 | 8.68 ± 0.03 | 7.91 ± 0.05 | 8.67 ± 0.06 | 7.43 ± 0.08 | 8.03 ± 0.09 |
| power | -2.81 ± 0.00 | -2.77 ± 0.00 | -2.79 ± 0.00 | -2.74 ± 0.00 | -2.73 ± 0.00 | -2.60 ± 0.01 | -2.75 ± 0.00 | -2.60 ± 0.01 | -2.75 ± 0.00 | -2.55 ± 0.01 |
| protein | -2.95 ± 0.00 | -2.86 ± 0.00 | -2.92 ± 0.00 | -2.84 ± 0.00 | -2.82 ± 0.00 | -2.68 ± 0.01 | -2.83 ± 0.00 | -2.68 ± 0.01 | -2.82 ± 0.00 | -2.66 ± 0.01 |
| wine | -0.89 ± 0.01 | -0.86 ± 0.01 | -0.87 ± 0.01 | -0.83 ± 0.02 | -0.85 ± 0.01 | -0.86 ± 0.02 | -0.83 ± 0.01 | -0.83 ± 0.02 | -0.80 ± 0.01 | -0.82 ± 0.02 |
| yacht | -0.67 ± 0.13 | -0.32 ± 0.12 | 0.12 ± 0.07 | 0.49 ± 0.13 | -0.00 ± 0.09 | 0.93 ± 0.16 | -0.45 ± 0.06 | 0.76 ± 0.10 | -0.74 ± 0.05 | 0.43 ± 0.10 |

Table 4: Average Train Log Likelihoods (nats) on the UCI Datasets (Errors are ±1 Standard Error)

| Dataset | MAP-1 | MAP-2 | MAP-1 NL | MAP-2 NL | Reg-1 NL | Reg-2 NL | BN(ML)-1 NL | BN(ML)-2 NL | BN(BO)-1 NL | BN(BO)-2 NL |
|---|---|---|---|---|---|---|---|---|---|---|
| boston | 2.51 ± 0.14 | 2.35 ± 0.18 | 2.29 ± 0.13 | 2.19 ± 0.18 | 2.15 ± 0.20 | 1.98 ± 0.22 | 2.20 ± 0.15 | 2.26 ± 0.13 | 1.69 ± 0.07 | 1.52 ± 0.13 |
| concrete | 4.03 ± 0.12 | 3.42 ± 0.16 | 3.61 ± 0.11 | 3.18 ± 0.17 | 3.35 ± 0.15 | 2.29 ± 0.14 | 3.20 ± 0.11 | 3.37 ± 0.12 | 2.86 ± 0.09 | 2.07 ± 0.10 |
| energy | 0.39 ± 0.02 | 0.32 ± 0.02 | 0.30 ± 0.01 | 0.25 ± 0.01 | 0.33 ± 0.01 | 0.23 ± 0.02 | 0.32 ± 0.01 | 0.20 ± 0.01 | 0.37 ± 0.01 | 0.18 ± 0.01 |
| kin8nm | 0.08 ± 0.00 | 0.06 ± 0.00 | 0.07 ± 0.00 | 0.06 ± 0.00 | 0.07 ± 0.00 | 0.06 ± 0.00 | 0.07 ± 0.00 | 0.06 ± 0.00 | 0.07 ± 0.00 | 0.05 ± 0.00 |
| naval | 0.00 ± 0.00 | 0.00 ± 0.00 | 0.00 ± 0.00 | 0.00 ± 0.00 | 0.00 ± 0.00 | 0.00 ± 0.00 | 0.00 ± 0.00 | 0.00 ± 0.00 | 0.00 ± 0.00 | 0.00 ± 0.00 |
| power | 4.03 ± 0.01 | 3.85 ± 0.01 | 3.92 ± 0.01 | 3.76 ± 0.01 | 3.72 ± 0.01 | 3.26 ± 0.04 | 3.78 ± 0.01 | 3.25 ± 0.03 | 3.79 ± 0.01 | 3.08 ± 0.03 |
| protein | 4.60 ± 0.02 | 4.23 ± 0.01 | 4.48 ± 0.01 | 4.15 ± 0.01 | 4.07 ± 0.01 | 3.52 ± 0.04 | 4.10 ± 0.00 | 3.52 ± 0.04 | 4.07 ± 0.01 | 3.45 ± 0.04 |
| wine | 0.59 ± 0.00 | 0.57 ± 0.01 | 0.58 ± 0.01 | 0.56 ± 0.01 | 0.57 ± 0.01 | 0.57 ± 0.01 | 0.56 ± 0.01 | 0.56 ± 0.01 | 0.54 ± 0.01 | 0.55 ± 0.01 |
| yacht | 0.42 ± 0.05 | 0.27 ± 0.05 | 0.22 ± 0.01 | 0.16 ± 0.03 | 0.27 ± 0.02 | 0.13 ± 0.02 | 0.36 ± 0.02 | 0.09 ± 0.02 | 0.49 ± 0.03 | 0.15 ± 0.02 |

Table 5: Train RMSEs on the UCI Datasets (Errors are ±1 Standard Error)

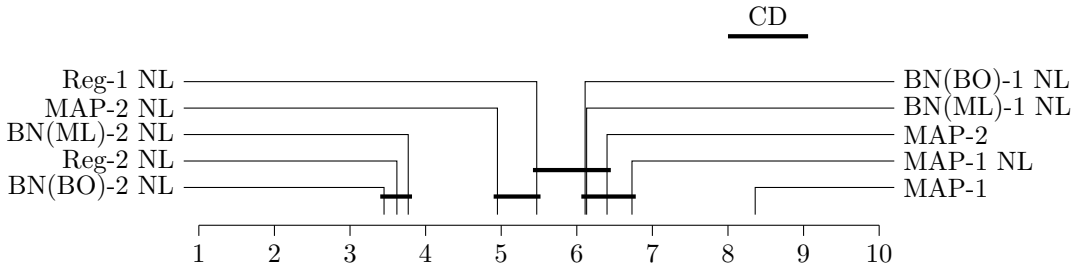


Figure 6: Average ranks on the UCI datasets according to test RMSEs, generated as described in Demšar (2006).

We make a few observations from these results. First, the two-layer marginal likelihood-based methods generally outperform the other methods, with the BN(BO)-2 NL model performing the best of all (although not significantly different from the BN(ML)-2 NL model according to the Friedman test). These are generally followed by the single-layer marginal-likelihood based methods, with the MAP-based methods performing the worst of all. This confirms our intuition that the more Bayesian versions of the models would yield better performance.

From the train log likelihoods and RMSEs, we observe that all the models exhibit overfitting for most of the datasets, with especially noticeable overfitting on 'boston', 'concrete', 'energy', and 'yacht'. The overfitting is generally worse on the two-layer models than the single-layer models, as there are more hyperparamters $\theta$ that can lead to overfitting in these models. Based off this trend, we expect that as the number of layers is increased further that the overfitting would worsen, thereby potentially limiting the use of neural linear models to smaller, shallower neural networks.

In Tables 6 and 7 we show the test log likelihoods for the UCI gap datasets. As we are more concerned about whether the models can capture in-between uncertainty than the performance of the models on these datasets, we do not compute average ranks for these datasets. Additionally, since the test set is not within the same distribution as the training set, we do not show results on the training sets as they cannot be compared to the test performance.

From these tables, we see that the MAP-$L$, MAP-$L$ NL, and Reg-$L$ NL models fail catastrophically on the 'naval' and 'energy' dataset. Additionally, MAP-1 performs especially poorly on 'yacht', although it is not clear whether this can be termed 'catastrophic'. While the BN NL models perform poorly on 'naval' and 'energy', by looking at the log likelihoods on the individual splits themselves we found that they were not actually failing catastrophically. This yet again confirms that the more fully Bayesian models are better, although it is surprising just how poorly the Reg-$L$ NL models perform. Additionally, because the overfitting we observed before worsens as the number of layers is increased, the performance of the BN NL models worsens as more layers are added.

| Dataset | MAP-1 | MAP-2 | MAP-1 NL | MAP-2 NL | Reg-1 NL | Reg-2 NL | BN(ML)-1 NL | BN(ML)-2 NL | BN(BO)-1 NL | BN(BO)-2 NL |
|---|---|---|---|---|---|---|---|---|---|---|
| boston | -2.79 ± 0.04 | -2.79 ± 0.08 | -2.72 ± 0.04 | -2.79 ± 0.09 | -2.86 ± 0.07 | -2.99 ± 0.17 | -2.76 ± 0.08 | -2.76 ± 0.06 | -2.81 ± 0.08 | -2.88 ± 0.08 |
| concrete | -3.61 ± 0.03 | -3.60 ± 0.06 | -3.61 ± 0.07 | -3.54 ± 0.05 | -4.70 ± 0.44 | -5.14 ± 0.27 | -3.74 ± 0.13 | -3.66 ± 0.13 | -3.72 ± 0.06 | -4.22 ± 0.17 |
| energy | -4.06 ± 1.27 | -6.55 ± 2.39 | -3.85 ± 1.20 | -5.89 ± 2.43 | -120.64 ± 78.92 | -95.79 ± 51.96 | -3.67 ± 1.11 | -4.22 ± 1.26 | -3.78 ± 1.16 | -3.87 ± 1.20 |
| kin8nm | 0.94 ± 0.05 | 1.09 ± 0.03 | 1.02 ± 0.06 | 1.15 ± 0.04 | 1.08 ± 0.04 | 1.18 ± 0.03 | 1.03 ± 0.03 | 1.15 ± 0.02 | 1.06 ± 0.02 | 1.18 ± 0.02 |
| naval | -12.06 ± 4.98 | -61.22 ± 28.17 | -7.43 ± 3.87 | -41.73 ± 14.80 | -716.29 ± 516.48 | -12294.53 ± 7478.34 | -4.30 ± 1.32 | -5.84 ± 1.40 | -1.17 ± 0.80 | -2.29 ± 0.99 |
| power | -2.87 ± 0.03 | -2.89 ± 0.04 | -2.87 ± 0.02 | -2.87 ± 0.02 | -3.19 ± 0.21 | -3.37 ± 0.22 | -2.93 ± 0.04 | -3.16 ± 0.08 | -2.96 ± 0.05 | -3.13 ± 0.09 |
| protein | -3.07 ± 0.01 | -3.06 ± 0.01 | -3.07 ± 0.01 | -3.06 ± 0.02 | -3.17 ± 0.04 | -3.29 ± 0.07 | -3.14 ± 0.03 | -3.22 ± 0.05 | -3.09 ± 0.03 | -3.22 ± 0.05 |
| wine | -0.96 ± 0.01 | -0.96 ± 0.01 | -0.96 ± 0.01 | -0.96 ± 0.01 | -1.03 ± 0.06 | -0.98 ± 0.01 | -0.96 ± 0.01 | -0.97 ± 0.01 | -0.98 ± 0.01 | -0.97 ± 0.02 |
| yacht | -3.69 ± 1.47 | -2.07 ± 0.30 | -2.58 ± 0.89 | -1.76 ± 0.27 | -2.20 ± 0.30 | -2.61 ± 0.57 | -1.81 ± 0.17 | -3.00 ± 0.77 | -1.72 ± 0.17 | -2.14 ± 0.28 |

Table 6: Average Test Log Likelihoods (nats) on the UCI Gap Datasets (Errors are ±1 Standard Error)

| Dataset | MAP-1 | MAP-2 | MAP-1 NL | MAP-2 NL | Reg-1 NL | Reg-2 NL | BN(ML)-1 NL | BN(ML)-2 NL | BN(BO)-1 NL | BN(BO)-2 NL |
|---|---|---|---|---|---|---|---|---|---|---|
| boston | 3.82 ± 0.16 | 3.54 ± 0.14 | 3.71 ± 0.15 | 3.59 ± 0.14 | 3.69 ± 0.17 | 3.71 ± 0.21 | 3.67 ± 0.18 | 3.62 ± 0.17 | 3.85 ± 0.23 | 3.87 ± 0.21 |
| concrete | 7.79 ± 0.18 | 7.78 ± 0.23 | 7.68 ± 0.23 | 7.44 ± 0.17 | 8.21 ± 0.48 | 8.27 ± 0.39 | 7.69 ± 0.51 | 7.33 ± 0.36 | 7.74 ± 0.31 | 9.20 ± 0.55 |
| energy | 2.83 ± 0.99 | 3.70 ± 1.33 | 3.09 ± 1.17 | 3.48 ± 1.21 | 4.24 ± 2.11 | 3.83 ± 1.49 | 4.15 ± 1.64 | 4.10 ± 1.64 | 4.76 ± 1.98 | 4.58 ± 1.87 |
| kin8nm | 0.09 ± 0.01 | 0.08 ± 0.00 | 0.09 ± 0.01 | 0.07 ± 0.00 | 0.08 ± 0.00 | 0.07 ± 0.00 | 0.09 ± 0.00 | 0.08 ± 0.00 | 0.08 ± 0.00 | 0.07 ± 0.00 |
| naval | 0.02 ± 0.00 | 0.03 ± 0.00 | 0.02 ± 0.00 | 0.03 ± 0.00 | 0.01 ± 0.00 | 0.01 ± 0.00 | 0.01 ± 0.00 | 0.01 ± 0.00 | 0.01 ± 0.00 | 0.01 ± 0.00 |
| power | 4.24 ± 0.12 | 4.33 ± 0.18 | 4.25 ± 0.09 | 4.27 ± 0.08 | 5.17 ± 0.60 | 5.23 ± 0.43 | 4.49 ± 0.15 | 5.17 ± 0.28 | 4.66 ± 0.21 | 5.27 ± 0.36 |
| protein | 5.16 ± 0.04 | 5.07 ± 0.06 | 5.13 ± 0.05 | 5.08 ± 0.06 | 5.23 ± 0.12 | 5.33 ± 0.16 | 5.27 ± 0.12 | 5.37 ± 0.17 | 5.14 ± 0.10 | 5.46 ± 0.17 |
| wine | 0.63 ± 0.01 | 0.63 ± 0.01 | 0.63 ± 0.01 | 0.63 ± 0.01 | 0.66 ± 0.02 | 0.64 ± 0.01 | 0.63 ± 0.01 | 0.64 ± 0.01 | 0.65 ± 0.01 | 0.64 ± 0.01 |
| yacht | 1.31 ± 0.14 | 1.05 ± 0.09 | 1.28 ± 0.14 | 1.01 ± 0.09 | 1.24 ± 0.11 | 1.22 ± 0.13 | 1.15 ± 0.11 | 1.31 ± 0.16 | 1.37 ± 0.15 | 1.59 ± 0.23 |

Table 7: Test RMSEs on the UCI Gap Datasets (Errors are ±1 Standard Error)

### D.2. Additional Results on the Effect of Hyperparameter Tuning

We describe the setup for our experiments on the effect of hyperparameter tuning as well as provide additional results not in the main text. We first describe the "reasonable" hyperparameter values that we selected:

**MAP** For the MAP baseline, we select a batch size of 32. We set $\gamma = 0.5$, corresponding to a unity prior variance. We set the two learning rates to the ADAM default of 1e-3 (Kingma and Ba, 2014). Finally, we allow for approximately 10000 gradient steps (we ensure that the last epoch is completed, so that there are at least 10000 gradient steps).

**MAP NL** For the MAP neural linear model, we take the above optimal MAP network and obtain 200 slice samples of $\alpha_W$ (the output weight prior variance), $\alpha_b$ (the output bias prior variance), and $\sigma^2$ for Bayesian linear regression. We initialize $\alpha_W = 1/50$ and $\alpha_b = 1$, to match the scaling used in Neal (1995).

**Regularized NL** We set $\gamma_W = \gamma_b = 0.5$, the learning rates to 1e-3 and the number of epochs to 5000.

**Bayesian noise NL (ML)** We initialize $a_0 = b_0 = \alpha_W = \alpha_b = 1$, the learning rate to 1e-3, and the number of epochs to 5000.

**Bayesian noise NL (BO)** We use the same hyperparameter settings as above, although in this case $a_0, b_0, \alpha_W$, and $\alpha_b$ will remain fixed.

The log likelihoods and RMSEs for each split are then compared to those obtained when hyperparameter tuning is allowed. We show the average test log likelihoods and test RMSEs for the models without hyperparameter tuning in Tables 9 and 10 for the UCI datasets and Tables 11 and 12 for the UCI gap datasets. We also visualize the results for the gap datasets in Figure 7.

These results show that in general the hyperparameter tuning is of essential importance, particularly for the two-layer cases. As a whole, the results are significantly worse than with hyperparameter tuning, and in particular, for each of the two-layer methods, there is at least one dataset where the results are catastrophically bad compared to the models with hyperparameter tuning. Somewhat by contrast, however, while the results for the gap datasets are worse for the methods that did not fail catastrophically, they are not catastrophically worse. For the methods that were not able to represent in-between uncertainty, however, we find that they now fail catastrophically on even more datasets.

We now verify that the differences induced by the hyperparameter tuning are indeed statistically significant. In order to do so, we use the Wilcoxon signed-rank test (Wilcoxon, 1992) as described in Demšar (2006). By comparing each tuned model to its non-tuned counterpart over all splits, we arrive at the table shown in Table 8. This shows that the difference is indeed statistically significant ($p < 0.05$) on both the standard and gap datasets for the vast majority of models, measured both by log likelihoods and RMSEs.

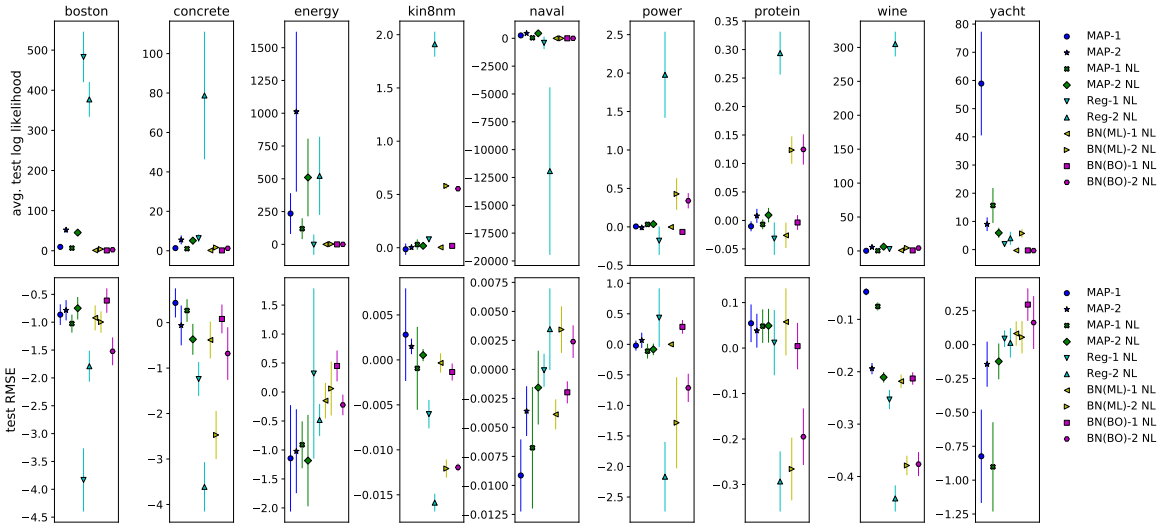


Figure 7: Differences in average test log likelihoods (nats) and RMSEs given by hyperparameter tuning for the UCI gap datasets

| **Model** | UCI: LL | UCI: RMSE | Gap: LL | Gap: RMSE |
|---|---|---|---|---|
| **MAP-1** | I | N | I | I |
| **MAP-2** | I | N | I | I |
| **MAP-1 NL** | I | I | I | I |
| **MAP-2 NL** | I | N | I | I |
| **Reg-1 NL** | I | I | I | I |
| **Reg-2 NL** | I | I | I | I |
| **BN(ML)-1 NL** | I | I | N | I |
| **BN(ML)-2 NL** | I | I | I | I |
| **BN(BO)-1 NL** | I | I | I | I |
| **BN(BO)-2 NL** | I | I | I | I |

Table 8: The Effect of Hyperparameter Tuning on the Models: (I) Improves, (W) Worsens, (N) Not Statistically Significant.

| Dataset | MAP-1 | MAP-2 | MAP-1 NL | MAP-2 NL | Reg-1 NL | Reg-2 NL | BN(ML)-1 NL | BN(ML)-2 NL | BN(BO)-1 NL | BN(BO)-2 NL |
|---|---|---|---|---|---|---|---|---|---|---|
| boston | -4.45 ± 0.42 | -17.22 ± 1.56 | -4.25 ± 0.37 | -16.65 ± 1.78 | -259.51 ± 22.70 | -227.87 ± 36.75 | -3.53 ± 0.17 | -12.69 ± 0.61 | -3.63 ± 0.20 | -7.76 ± 0.32 |
| concrete | -3.19 ± 0.05 | -3.85 ± 0.15 | -3.16 ± 0.05 | -3.88 ± 0.16 | -3.83 ± 0.18 | -10.66 ± 1.10 | -3.11 ± 0.04 | -3.97 ± 0.19 | -3.12 ± 0.05 | -4.18 ± 0.15 |
| energy | -0.89 ± 0.08 | -0.99 ± 0.10 | -0.84 ± 0.06 | -0.88 ± 0.11 | -0.79 ± 0.08 | -1.91 ± 0.20 | -0.65 ± 0.03 | -0.98 ± 0.11 | -0.69 ± 0.03 | -0.55 ± 0.04 |
| kin8nm | 1.05 ± 0.01 | 1.18 ± 0.01 | 1.07 ± 0.01 | 1.22 ± 0.01 | 1.11 ± 0.01 | 0.48 ± 0.02 | 1.12 ± 0.01 | 0.79 ± 0.02 | 1.13 ± 0.01 | 0.78 ± 0.02 |
| naval | 4.96 ± 0.03 | 5.63 ± 0.03 | 6.12 ± 0.04 | 6.52 ± 0.03 | 7.84 ± 0.03 | 8.17 ± 0.02 | 7.44 ± 0.01 | 8.46 ± 0.03 | 7.35 ± 0.01 | 8.28 ± 0.01 |
| power | -2.83 ± 0.01 | -2.80 ± 0.01 | -2.81 ± 0.01 | -2.78 ± 0.01 | -2.79 ± 0.01 | -2.85 ± 0.02 | -2.79 ± 0.01 | -2.77 ± 0.02 | -2.79 ± 0.01 | -2.76 ± 0.02 |
| protein | -2.97 ± 0.01 | -2.91 ± 0.00 | -2.95 ± 0.00 | -2.90 ± 0.00 | -2.88 ± 0.00 | -2.81 ± 0.01 | -2.87 ± 0.00 | -2.81 ± 0.01 | -2.87 ± 0.00 | -2.81 ± 0.00 |
| wine | -1.05 ± 0.02 | -2.36 ± 0.10 | -1.08 ± 0.02 | -2.60 ± 0.10 | -1.76 ± 0.08 | -194.96 ± 13.60 | -1.34 ± 0.04 | -13.20 ± 0.25 | -1.48 ± 0.05 | -12.49 ± 0.21 |
| yacht | -4.31 ± 0.68 | -2.65 ± 0.59 | -3.25 ± 0.42 | -2.21 ± 0.49 | -1.36 ± 0.18 | -1.44 ± 0.24 | -1.16 ± 0.08 | -6.82 ± 1.15 | -1.13 ± 0.05 | -0.95 ± 0.10 |

Table 9: Average Test Log Likelihoods (nats) on the UCI Datasets Without Hyperparameter Tuning

| Dataset | MAP-1 | MAP-2 | MAP-1 NL | MAP-2 NL | Reg-1 NL | Reg-2 NL | BN(ML)-1 NL | BN(ML)-2 NL | BN(BO)-1 NL | BN(BO)-2 NL |
|---|---|---|---|---|---|---|---|---|---|---|
| boston | 3.15 ± 0.15 | 3.12 ± 0.14 | 3.21 ± 0.16 | 3.10 ± 0.15 | 5.21 ± 0.22 | 3.99 ± 0.28 | 3.23 ± 0.15 | 3.43 ± 0.18 | 3.43 ± 0.24 | 3.55 ± 0.15 |
| concrete | 4.97 ± 0.12 | 4.79 ± 0.14 | 4.93 ± 0.12 | 4.69 ± 0.15 | 4.83 ± 0.21 | 5.34 ± 0.22 | 4.82 ± 0.12 | 4.87 ± 0.24 | 4.78 ± 0.12 | 4.76 ± 0.12 |
| energy | 0.49 ± 0.02 | 0.43 ± 0.01 | 0.46 ± 0.01 | 0.40 ± 0.01 | 0.45 ± 0.02 | 0.42 ± 0.02 | 0.46 ± 0.01 | 0.42 ± 0.02 | 0.48 ± 0.02 | 0.40 ± 0.01 |
| kin8nm | 0.08 ± 0.00 | 0.07 ± 0.00 | 0.08 ± 0.00 | 0.07 ± 0.00 | 0.08 ± 0.00 | 0.08 ± 0.00 | 0.08 ± 0.00 | 0.08 ± 0.00 | 0.08 ± 0.00 | 0.08 ± 0.00 |
| naval | 0.00 ± 0.00 | 0.00 ± 0.00 | 0.00 ± 0.00 | 0.00 ± 0.00 | 0.00 ± 0.00 | 0.00 ± 0.00 | 0.00 ± 0.00 | 0.00 ± 0.00 | 0.00 ± 0.00 | 0.00 ± 0.00 |
| power | 4.10 ± 0.03 | 3.99 ± 0.04 | 4.02 ± 0.04 | 3.91 ± 0.03 | 3.94 ± 0.04 | 3.80 ± 0.04 | 3.94 ± 0.04 | 3.74 ± 0.04 | 3.94 ± 0.04 | 3.70 ± 0.04 |
| protein | 4.72 ± 0.02 | 4.45 ± 0.02 | 4.63 ± 0.01 | 4.39 ± 0.01 | 4.29 ± 0.01 | 3.95 ± 0.02 | 4.25 ± 0.02 | 3.95 ± 0.02 | 4.26 ± 0.01 | 3.96 ± 0.01 |
| wine | 0.64 ± 0.01 | 0.68 ± 0.01 | 0.66 ± 0.01 | 0.70 ± 0.01 | 0.74 ± 0.02 | 0.92 ± 0.02 | 0.69 ± 0.01 | 0.92 ± 0.02 | 0.73 ± 0.01 | 0.90 ± 0.02 |
| yacht | 0.66 ± 0.04 | 0.59 ± 0.05 | 0.63 ± 0.04 | 0.58 ± 0.05 | 0.60 ± 0.04 | 0.56 ± 0.05 | 0.77 ± 0.06 | 0.60 ± 0.06 | 0.75 ± 0.05 | 0.64 ± 0.07 |

Table 10: Test RMSEs on the UCI Datasets Without Hyperparameter Tuning

| Dataset | MAP-1 | MAP-2 | MAP-1 NL | MAP-2 NL | Reg-1 NL | Reg-2 NL | BN(ML)-1 NL | BN(ML)-2 NL | BN(BO)-1 NL | BN(BO)-2 NL |
|---|---|---|---|---|---|---|---|---|---|---|
| boston | -12.33 ± 1.48 | -54.66 ± 5.95 | -9.35 ± 1.02 | -48.01 ± 4.93 | -485.85 ± 61.72 | -380.12 ± 42.19 | -3.69 ± 0.14 | -6.82 ± 0.30 | -3.60 ± 0.11 | -5.29 ± 0.16 |
| concrete | -4.94 ± 0.31 | -9.13 ± 2.01 | -4.65 ± 0.27 | -8.68 ± 1.69 | -11.08 ± 1.75 | -83.89 ± 32.24 | -3.93 ± 0.07 | -5.31 ± 0.24 | -3.89 ± 0.11 | -5.46 ± 0.23 |
| energy | -238.97 ± 152.20 | -1018.49 ± 607.15 | -123.12 ± 76.99 | -516.00 ± 293.08 | -118.03 ± 67.94 | -618.22 ± 344.98 | -3.73 ± 1.19 | -6.62 ± 1.93 | -3.40 ± 1.02 | -3.81 ± 1.16 |
| kin8nm | 0.96 ± 0.03 | 1.09 ± 0.03 | 0.99 ± 0.02 | 1.13 ± 0.03 | 1.00 ± 0.03 | -0.73 ± 0.14 | 1.03 ± 0.02 | 0.57 ± 0.03 | 1.04 ± 0.02 | 0.62 ± 0.03 |
| naval | -271.04 ± 56.88 | -511.37 ± 97.14 | -53.08 ± 17.22 | -487.50 ± 97.54 | -309.49 ± 71.79 | -365.79 ± 100.30 | -2.01 ± 0.90 | -3.51 ± 1.15 | -1.48 ± 0.84 | -1.09 ± 0.90 |
| power | -2.87 ± 0.02 | -2.88 ± 0.03 | -2.90 ± 0.03 | -2.91 ± 0.04 | -3.01 ± 0.05 | -5.35 ± 0.76 | -2.92 ± 0.04 | -3.59 ± 0.27 | -2.90 ± 0.03 | -3.47 ± 0.15 |
| protein | -3.06 ± 0.02 | -3.06 ± 0.02 | -3.06 ± 0.02 | -3.07 ± 0.02 | -3.13 ± 0.03 | -3.58 ± 0.09 | -3.11 ± 0.03 | -3.34 ± 0.07 | -3.09 ± 0.03 | -3.35 ± 0.06 |
| wine | -1.20 ± 0.03 | -6.46 ± 0.42 | -1.26 ± 0.04 | -7.24 ± 0.53 | -3.78 ± 0.34 | -305.93 ± 17.49 | -1.78 ± 0.04 | -5.43 ± 0.06 | -1.75 ± 0.04 | -5.19 ± 0.08 |
| yacht | -62.59 ± 17.77 | -11.08 ± 2.36 | -18.25 ± 5.93 | -7.69 ± 1.77 | -4.19 ± 0.61 | -6.65 ± 2.20 | -1.51 ± 0.08 | -8.74 ± 1.42 | -1.44 ± 0.11 | -1.79 ± 0.11 |

Table 11: Average Test Log Likelihoods (nats) on the UCI Gap Datasets Without Hyperparameter Tuning

| Dataset | MAP-1 | MAP-2 | MAP-1 NL | MAP-2 NL | Reg-1 NL | Reg-2 NL | BN(ML)-1 NL | BN(ML)-2 NL | BN(BO)-1 NL | BN(BO)-2 NL |
|---|---|---|---|---|---|---|---|---|---|---|
| boston | 4.69 ± 0.21 | 4.33 ± 0.23 | 4.74 ± 0.21 | 4.33 ± 0.22 | 7.52 ± 0.67 | 5.50 ± 0.31 | 4.59 ± 0.29 | 4.62 ± 0.23 | 4.46 ± 0.28 | 5.39 ± 0.32 |
| concrete | 7.35 ± 0.33 | 7.84 ± 0.42 | 7.42 ± 0.33 | 7.81 ± 0.39 | 9.45 ± 0.51 | 11.88 ± 0.71 | 8.07 ± 0.21 | 9.80 ± 0.43 | 7.66 ± 0.38 | 9.88 ± 0.45 |
| energy | 3.97 ± 1.47 | 4.72 ± 1.96 | 4.00 ± 1.48 | 4.67 ± 1.93 | 3.92 ± 1.63 | 4.31 ± 1.61 | 4.30 ± 1.84 | 4.04 ± 1.46 | 4.31 ± 1.78 | 4.80 ± 2.02 |
| kin8nm | 0.09 ± 0.00 | 0.08 ± 0.00 | 0.09 ± 0.00 | 0.07 ± 0.00 | 0.09 ± 0.00 | 0.09 ± 0.00 | 0.09 ± 0.00 | 0.09 ± 0.00 | 0.08 ± 0.00 | 0.09 ± 0.00 |
| naval | 0.03 ± 0.00 | 0.03 ± 0.00 | 0.03 ± 0.01 | 0.03 ± 0.00 | 0.01 ± 0.00 | 0.01 ± 0.00 | 0.01 ± 0.00 | 0.01 ± 0.00 | 0.01 ± 0.00 | 0.01 ± 0.00 |
| power | 4.26 ± 0.07 | 4.27 ± 0.10 | 4.36 ± 0.12 | 4.35 ± 0.13 | 4.73 ± 0.18 | 7.40 ± 0.96 | 4.49 ± 0.17 | 6.45 ± 0.98 | 4.38 ± 0.13 | 5.98 ± 0.50 |
| protein | 5.10 ± 0.06 | 5.03 ± 0.07 | 5.08 ± 0.06 | 5.03 ± 0.08 | 5.22 ± 0.10 | 5.63 ± 0.17 | 5.21 ± 0.12 | 5.64 ± 0.22 | 5.14 ± 0.10 | 5.66 ± 0.18 |
| wine | 0.68 ± 0.01 | 0.82 ± 0.01 | 0.71 ± 0.01 | 0.84 ± 0.01 | 0.91 ± 0.03 | 1.08 ± 0.02 | 0.85 ± 0.01 | 1.01 ± 0.02 | 0.86 ± 0.01 | 1.02 ± 0.02 |
| yacht | 2.13 ± 0.34 | 1.20 ± 0.16 | 2.18 ± 0.34 | 1.13 ± 0.15 | 1.19 ± 0.11 | 1.21 ± 0.13 | 1.07 ± 0.06 | 1.25 ± 0.16 | 1.08 ± 0.13 | 1.43 ± 0.15 |

Table 12: Test RMSEs on the UCI Gap Datasets Without Hyperparameter Tuning

| Model | boston | concrete | energy | kin8nm | naval | power | protein | wine | yacht |
|---|---|---|---|---|---|---|---|---|---|
| **MFVI-1** | -2.60 ± 0.06 | -3.09 ± 0.03 | -0.74 ± 0.02 | 1.11 ± 0.01 | 5.91 ± 0.04 | -2.82 ± 0.01 | -2.94 ± 0.00 | -0.97 ± 0.01 | -1.25 ± 0.13 |
| **MFVI-2** | -2.82 ± 0.04 | -3.10 ± 0.02 | -0.77 ± 0.02 | 1.24 ± 0.01 | 5.99 ± 0.08 | -2.81 ± 0.01 | -2.87 ± 0.00 | -0.98 ± 0.01 | -1.15 ± 0.05 |
| **MCD-1** | -2.71 ± 0.11 | -3.33 ± 0.02 | -1.89 ± 0.03 | 0.67 ± 0.01 | 3.31 ± 0.01 | -2.98 ± 0.01 | -3.01 ± 0.00 | -0.97 ± 0.02 | -2.48 ± 0.10 |
| **MCD-2** | -2.70 ± 0.12 | -3.17 ± 0.03 | -1.34 ± 0.02 | 0.74 ± 0.01 | 3.91 ± 0.02 | -2.92 ± 0.01 | -2.95 ± 0.00 | -1.16 ± 0.04 | -2.88 ± 0.22 |

Table 13: Average Test Log Likelihoods (nats) on the UCI Datasets for MFVI and Monte Carlo Dropout

| Model | boston | concrete | energy | kin8nm | naval | power | protein | wine | yacht |
|---|---|---|---|---|---|---|---|---|---|
| **MFVI-1** | 3.19 ± 0.18 | 5.24 ± 0.11 | 0.50 ± 0.01 | 0.08 ± 0.00 | 0.00 ± 0.00 | 4.07 ± 0.04 | 4.61 ± 0.02 | 0.64 ± 0.01 | 0.73 ± 0.05 |
| **MFVI-2** | 3.73 ± 0.25 | 5.35 ± 0.10 | 0.51 ± 0.01 | 0.07 ± 0.00 | 0.00 ± 0.00 | 4.02 ± 0.03 | 4.29 ± 0.02 | 0.64 ± 0.01 | 0.76 ± 0.04 |
| **MCD-1** | 3.24 ± 0.22 | 6.65 ± 0.15 | 1.40 ± 0.07 | 0.12 ± 0.00 | 0.01 ± 0.00 | 4.68 ± 0.04 | 4.89 ± 0.02 | 0.63 ± 0.01 | 1.85 ± 0.34 |
| **MCD-2** | 3.06 ± 0.15 | 5.63 ± 0.16 | 0.82 ± 0.04 | 0.11 ± 0.00 | 0.00 ± 0.00 | 4.40 ± 0.04 | 4.55 ± 0.02 | 0.65 ± 0.01 | 3.16 ± 0.62 |

Table 14: Test RMSEs on the UCI Datasets for MFVI and Monte Carlo Dropout

#### D.2.1. Comparison to MFVI and MCD

To ensure that this worse behavior is not because all models require hyperparameter tuning to perform reasonably well, we now compare these results to results for mean field variational inference (MFVI) and Monte Carlo dropout (MCD) without hyperparameter tuning. We implement MFVI according to Blundell et al. (2015) using the local reparameterization trick (Kingma et al., 2015). We set a unity prior variance and use a step size of 1e-3, using ADAM (Kingma and Ba, 2014) with approximately 25000 gradient steps and a batch size of 32. We allow the gradients to be estimated using 10 samples from the approximate posterior at each step. For testing we use 100 samples from the approximate posterior.

For MCD, we follow the implementation in Gal and Ghahramani (2016). We set the dropout rate to $p = 0.05$ with weight decay corresponding to unity prior variance. We again use a learning rate of 1e-3, using ADAM (Kingma and Ba, 2014) with approximately 25000 gradient steps using a batch size of 32. For testing we use 100 samples generated by the neural network.

For the UCI datasets, we tabulate the test log likelihoods and test RMSEs for one- and two-layer architectures in Tables 13 and 14. These generally show reasonable values for each dataset despite the absence of hyperparameter tuning: there is no dataset for which either method can be said to do catastrophically badly. In fact, the results for MFVI are largely competitive with the best results we obtained for the neural linear models using hyperparameter tuning. This difference suggests that the reason hyperparameter tuning is important in the neural linear models is because it is necessary to carefully regularize the weights, whereas being approximately Bayesian over all of the weights is not as sensitive to the choice of hyperparameters. Although the train log likelihoods and RMSEs are not reported here, they confirm this intuition: the neural linear models still suffer from substantial overfitting, whereas the overfitting we observed for MFVI and MCD is far less.

As with the results for the UCI datasets with tuned hyperparameters, we compute average ranks for the models across all splits and use the Friedman test as described in Demšar (2006) to determine whether the differences are statistically significant. We plot the ranking using test log likelihoods in Figure 8 and using test RMSEs in Figure 9. The rankings show that MCD performs poorly on average in terms of test log likelihood, whereas MFVI performs reasonably well for both log likelihood and RMSE. However, these rankings

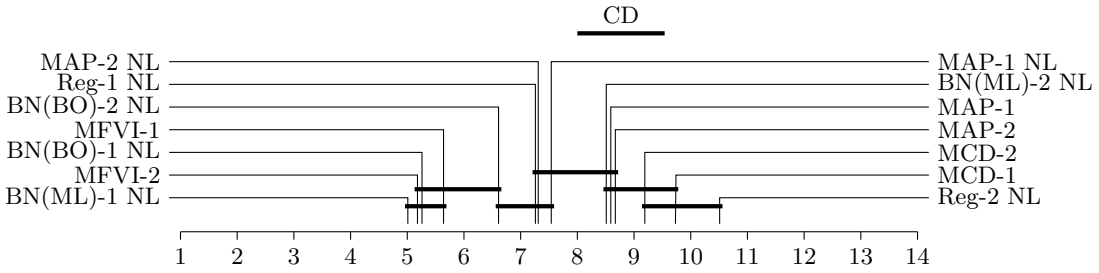


Figure 8: Average ranks of the single-run models on the UCI datasets according to average test log likelihoods, generated as described in Demšar (2006).

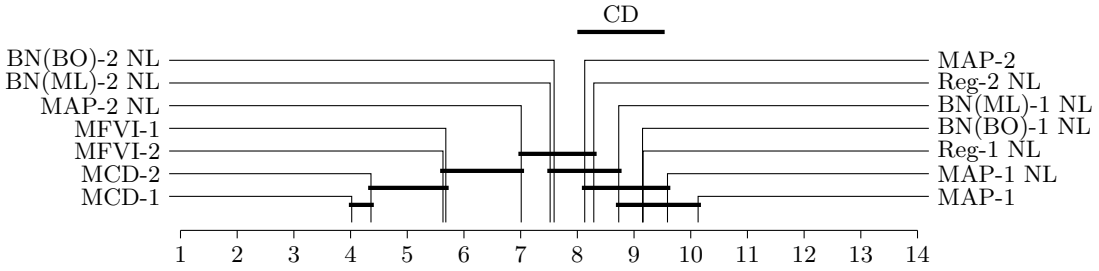


Figure 9: Average ranks of the single-run models on the UCI datasets according to test RMSEs, generated as described in Demšar (2006).

| Model | boston | concrete | energy | kin8nm | naval | power | protein | wine | yacht |
|---|---|---|---|---|---|---|---|---|---|
| **MFVI-1** | -2.71 ± 0.06 | -3.53 ± 0.10 | -83.48 ± 53.70 | 1.05 ± 0.02 | -1318.01 ± 243.27 | -2.85 ± 0.02 | -3.04 ± 0.02 | -0.95 ± 0.01 | -1.84 ± 0.18 |
| **MFVI-2** | -2.88 ± 0.01 | -3.49 ± 0.07 | -11.93 ± 5.53 | 1.19 ± 0.02 | -1090.67 ± 192.18 | -2.94 ± 0.05 | -3.07 ± 0.03 | -0.97 ± 0.01 | -1.71 ± 0.18 |
| **MCD-1** | -2.88 ± 0.07 | -3.62 ± 0.04 | -20.71 ± 11.00 | 0.66 ± 0.02 | 2.46 ± 0.06 | -3.02 ± 0.03 | -3.08 ± 0.01 | -1.03 ± 0.02 | -2.67 ± 0.11 |
| **MCD-2** | -3.25 ± 0.12 | -3.89 ± 0.10 | -42.31 ± 23.94 | 0.73 ± 0.02 | -0.73 ± 0.55 | -3.23 ± 0.18 | -3.10 ± 0.02 | -1.52 ± 0.05 | -2.80 ± 0.16 |

Table 15: Average Test Log Likelihoods (nats) on the UCI Gap Datasets for MFVI and Monte Carlo Dropout

| Model | boston | concrete | energy | kin8nm | naval | power | protein | wine | yacht |
|---|---|---|---|---|---|---|---|---|---|
| **MFVI-1** | 3.78 ± 0.19 | 7.04 ± 0.33 | 4.30 ± 1.82 | 0.08 ± 0.00 | 0.03 ± 0.00 | 4.25 ± 0.12 | 5.02 ± 0.06 | 0.63 ± 0.01 | 1.30 ± 0.12 |
| **MFVI-2** | 3.70 ± 0.16 | 7.33 ± 0.25 | 2.58 ± 0.88 | 0.07 ± 0.00 | 0.03 ± 0.00 | 4.67 ± 0.23 | 5.05 ± 0.11 | 0.63 ± 0.01 | 1.26 ± 0.16 |
| **MCD-1** | 3.66 ± 0.12 | 8.00 ± 0.20 | 5.01 ± 1.72 | 0.12 ± 0.00 | 0.01 ± 0.00 | 4.87 ± 0.14 | 5.20 ± 0.05 | 0.65 ± 0.01 | 3.17 ± 0.55 |
| **MCD-2** | 3.58 ± 0.12 | 8.06 ± 0.24 | 5.18 ± 2.12 | 0.11 ± 0.00 | 0.02 ± 0.00 | 5.54 ± 0.64 | 5.18 ± 0.08 | 0.70 ± 0.01 | 3.75 ± 0.61 |

Table 16: Test RMSEs on the UCI Gap Datasets for MFVI and Monte Carlo Dropout

do not take into account that the neural linear models fail catastrophically on some, but not all, datasets since they only take the ordering of the methods into account and not how well they perform. Therefore, we still argue that for the standard UCI datasets MFVI and MCD are better since they perform relatively well across all datasets without the need for hyperparameter tuning.

Finally, we consider the performance of MFVI and MCD on the UCI gap datasets. We tabulate average test log likelihoods and test RMSEs in Tables 15 and 16. These echo the results in Foong et al. (2019b), showing catastrophic failure of MFVI to express 'in-between' uncertainty for the 'energy' and 'naval' datasets. They also show that MCD fails catastrophically on the 'energy' datasets, as suggested by theoretical results in Foong et al. (2019a); however, we believe these are the first results in the literature that show catastrophic failure to express 'in-between' uncertainty on a real dataset. The gap results therefore show one crucial advantage of the neural linear models over MFVI and MCD: their ability to express 'in-between' uncertainty.

### D.3. Effect of Slice Sampling

In this section, we briefly investigate the effect of slice sampling on the performance of the models. We first make plots of the predictive posterior distribution for each model trained on the toy problem of Section 3. These plots are visible in Figure 10. Note that the MAP-2 NL model simply becomes MAP inference. The most visible difference between Figure 10 and Figure 1 can be seen in the BN(BO)-2 NL model, which seems to have gained certainty at the edges while perhaps becoming slightly more uncertain in the gap; however, the effect in the gap is almost negligible. We observe the opposite effect in the MAP-2 NL model. Additionally, the Reg-2 NL model becomes slightly smoother. In general, however, it would seem that the effect of slice sampling for the toy problem is small.

We then plot the differences in the log likelihoods and RMSEs between the full models (with slice sampling) and the equivalent models without the final slice sampling step, to observe any quantitative differences. These plots are shown in Figure 11 for the UCI datasets and Figure 12 for the UCI gap datasets. These plots do not give a clear picture of whether slice samping improves or worsens the performance of these models: it seems to depend on both the model and the dataset.

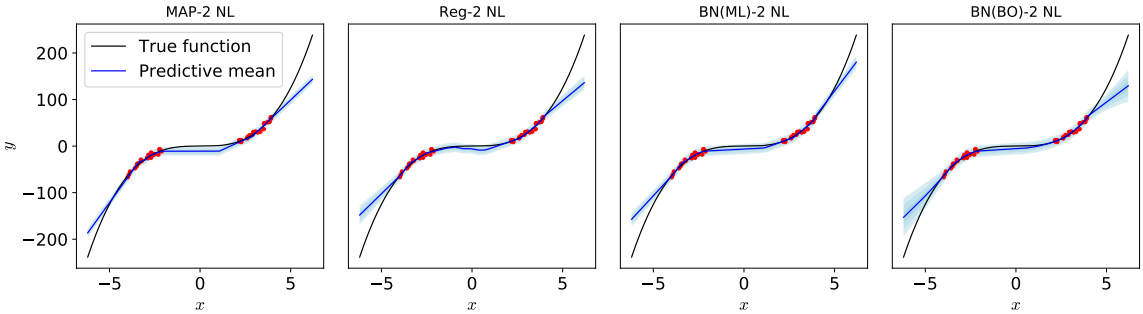


Figure 10: Predictive distributions for the toy problem, without slice sampling. Each shaded region represents one predictive standard deviation. Note that without slice sampling, the MAP-2 NL model simply becomes MAP-2.

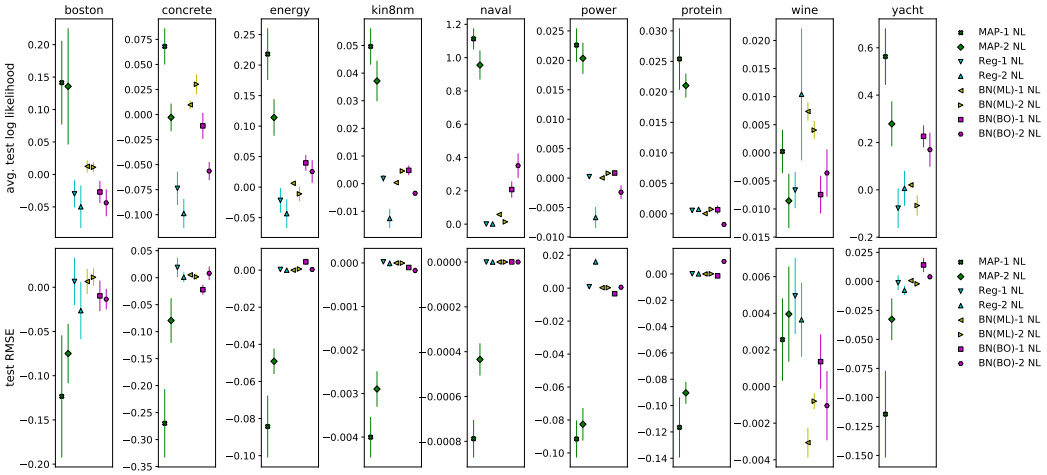


Figure 11: Differences in average test log likelihoods (nats) and RMSEs given by slice sampling for the UCI datasets

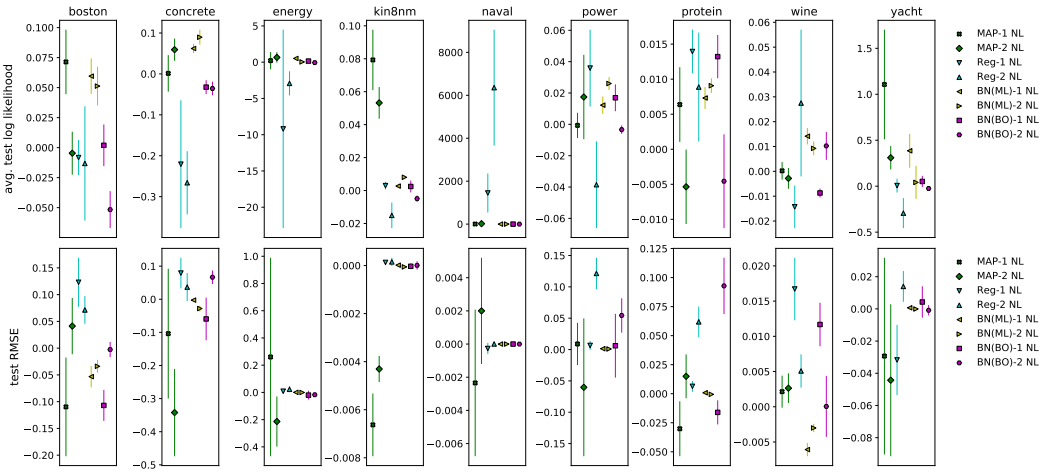


Figure 12: Differences in average test log likelihoods (nats) and RMSEs given by slice sampling for the UCI gap datasets

| **Model** | UCI: LL | UCI: RMSE | Gap: LL | Gap: RMSE |
|---|---|---|---|---|
| **MAP-1 NL** | I | I | I | I |
| **MAP-2 NL** | I | I | I | N |
| **Reg-1 NL** | W | N | I | W |
| **Reg-2 NL** | W | N | N | W |
| **BN(ML)-1 NL** | I | N | I | N |
| **BN(ML)-2 NL** | I | N | I | I |
| **BN(BO)-1 NL** | I | W | N | N |
| **BN(BO)-2 NL** | N | N | W | N |

Table 17: The Effect of Slice Sampling on the Models: (I) Improves, (W) Worsens, (N) Not Statistically Significant.

To gain a clearer insight into whether slice sampling improves the performance of the neural linear models, we once again perform the Wilcoxon signed-rank test to compare the results obtained with slice sampling to those without. The results of this analysis is shown in Table 17. This shows that the majority of models are in fact improved by slice sampling, particularly when the improvement is measured in terms of the log likelihoods. However, in many cases, particularly when performance is measured in terms of RMSE, the effect of slice sampling is not statistically significant. Furthermore, it seems that performance for the Reg-$L$ NL and BN(BO)-2 NL models may be worsened by slice sampling.

In conclusion, in most cases performance will not be worsened by slice sampling. In particular, the MAP-$L$ NL models seem to benefit especially from slice sampling. However, slice sampling is likely detrimental to the Reg-$L$ NL models in all cases and potentially harmful for the BN(BO)-2 NL model when it comes to in-between uncertainty.