5th Symposium on Advances in Approximate Bayesian Inference, 2023 1–20

# Linearized Laplace Inference in Neural Additive Models

**Kouroche Bouchiat** KBOUCHIAT@STUDENT.ETHZ.CH
*ETH Zürich*
**Alexander Immer** ALEXANDER.IMMER@INF.ETHZ.CH
*ETH Zürich & Max Planck Institute for Intelligent Systems*
**Hugo Yèche** HYECHE@INF.ETHZ.CH
*ETH Zürich*
**Vincent Fortuin** VINCENT.FORTUIN@HELMHOLTZ-MUNICH.DE
*Helmholtz AI*

## Abstract

Deep neural networks are highly effective but suffer from a lack of interpretability due to their black-box nature. Neural additive models (NAMs) solve this by separating into additive sub-networks, revealing the interactions between features and predictions. In this paper, we approach the NAM from a Bayesian perspective in order to quantify the uncertainty in the recovered interactions. Linearized Laplace approximation enables inference of these interactions directly in function space and yields a tractable estimate of the marginal likelihood, which can be used to perform implicit feature selection through an empirical Bayes procedure. Empirically, we show that Laplace-approximated NAMs (LA-NAM) are both more robust to noise and easier to interpret than their non-Bayesian counterpart for tabular regression and classification tasks.

## 1. Introduction

Over the past decade, deep neural networks (DNNs) have found successful applications in numerous fields, ranging from computer vision and speech recognition to natural language processing and recommendation systems. This success is often attributed to the growing availability of data in all areas of science and industry. However, their opaque nature has impeded their use in domains where comprehending the reasoning behind their decision-making process is crucial (Pumplun et al., 2021; Veale et al., 2018).

Model-agnostic methods, such as partial dependence (Friedman, 2001), SHAP (Lundberg and Lee, 2017), and LIME (Ribeiro et al., 2016) provide a standardized approach to explaining predictions in machine learning, but the explanations they generate for DNNs are not faithful representations of their full complexity (Rudin, 2019). Instead, one can promote interpretability in DNNs by acting directly on their architecture and training procedure. In generalized additive models (Hastie and Tibshirani, 1999), the response variable $y$ is associated to the predictor variables $x_1$, $x_2$, ..., $x_p$ using a structure of the form

$$g(\mathbb{E}[y \mid x_1, ..., x_p]) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p). \tag{1}$$

The neural additive models (NAMs) proposed by Agarwal et al. (2021) build on this premise. In these models, each dimension of the input is handled by a separate sub-network, exposing

Newer version available at https://arxiv.org/abs/2305.16905.

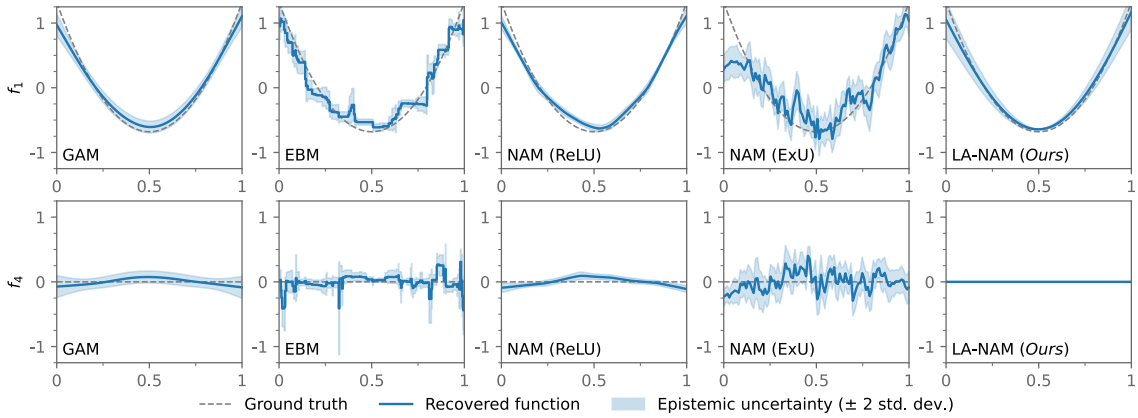


Figure 1: Regression on a synthetic dataset with known additive structure. The proposed LA-NAM fits the data well, provides useful uncertainty estimates, and correctly ignores the uninformative feature ($f_4$, bottom).

the relationship between the features and the model predictions. However, these models are not inherently safeguarded against overconfidence, and it is also desirable that they be able to express the uncertainty in the relationships they uncover. In this work, we show that linearized Laplace inference (Immer et al., 2021b) of Bayesian subnetworks leads to better uncertainty estimates and enables marginal-likelihood-based feature selection, thus improving performance and interpretability. A detailed treatment of the related work is deferred to Appendix C.

## 2. Linearized Laplace Inference in Neural Additive Models

We propose the Laplace-approximated neural additive model (LA-NAM), a generalized additive model with Bayesian neural networks that relies on the linearized Laplace approximation for inference (MacKay, 1991; Khan et al., 2019; Foong et al., 2019). We approximate the log marginal likelihood and posterior over the individual additive feature networks. Optimizing the log marginal likelihood results in a selection of the feature networks by virtue of adapting their respective regularization strength. The posterior predictive decomposes across features and can therefore be easily visualized as is common for additive models. We choose to use the Kronecker-factored Gauss-Newton approximation (Martens and Grosse, 2015) for estimating the log marginal likelihood and make use of the associated linearized predictive (Immer et al., 2021a,b).

### 2.1. Bayesian Neural Additive Model

Consider a tabular dataset $\mathcal{D} = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$ consisting of $N$ pairs of inputs $\mathbf{x} = [x_1, \ldots, x_D]^\top$ and labels $y$. In the Bayesian NAM, the networks of the NAM (Agarwal et al., 2021) are replaced with Bayesian neural networks $f_1, f_2 \ldots f_D$ parameterized by $\boldsymbol{\theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_D]^\top$, with $\boldsymbol{\theta}_d \in \mathbb{R}^{P_d}$ and $f_d : \mathbb{R} \times \mathbb{R}^{P_d} \to \mathbb{R}$. In most cases, all $f_d$ have the same architecture but may also slightly vary, when attending to categorical features for example. Using an inverse

link function $g : \mathbb{R} \to \mathbb{R}$, such as the logit function in classification, we have

$$g(\mathbb{E}[y \mid x_1, \ldots, x_D]) = f(\boldsymbol{x};\, \boldsymbol{\theta}) = f_1(x_1;\, \boldsymbol{\theta}_1) + f_2(x_2;\, \boldsymbol{\theta}_2) + \cdots + f_D(x_D;\, \boldsymbol{\theta}_D). \tag{2}$$

We refer to an individual network $f_d$ as a *feature network*. We have a likelihood function of either $p(y \mid \mathbf{x}, \boldsymbol{\theta}) = \mathcal{N}(y;\, f(\mathbf{x};\, \boldsymbol{\theta}), \sigma^2)$ for regression with identity link and observation noise hyperparameter $\sigma$ or $p(y \mid \mathbf{x}, \boldsymbol{\theta}) = \text{Bernoulli}(s(f(\mathbf{x};\, \boldsymbol{\theta})))$ for binary classification with sigmoid link $s$. Other likelihoods and compatible link functions are also possible.

We use a zero-mean Gaussian prior with different precisions for each of the $D$ feature network parameters to adaptively regularize the networks and select features in a fashion not dissimilar to automatic relevance determination (ARD; MacKay, 1994; Neal, 1995). In this case we have individual priors per feature network i.e., $p(\boldsymbol{\theta} \mid \boldsymbol{\delta}) = \prod_{d=1}^{D} \mathcal{N}(\boldsymbol{\theta}_d;\, \mathbf{0},\, \delta_d^{-1}\mathbf{I})$. Large values of $\delta_d$ push the corresponding function $f_d$ toward zero and low values encourage a highly non-linear fit. In practice, one can use a prior with precisions assigned to each layer of each feature network as this setup has been shown to be beneficial in linearized Laplace (Immer et al., 2021a; Antorán et al., 2022). The joint distribution of the Bayesian NAM is given by $p(\mathcal{D},\, \boldsymbol{\theta} \mid \boldsymbol{\delta}) = p(\boldsymbol{\theta} \mid \boldsymbol{\delta}) \prod_n p(y_n \mid \mathbf{x}_n,\, \boldsymbol{\theta})$.

### 2.2. Laplace-Approximated Neural Additive Model

We use the linearized Laplace approximation (Laplace, 1774; MacKay, 1991; Khan et al., 2019) to the posterior and log marginal likelihood because it provides differentiable log marginal likelihood estimates (Immer et al., 2021a) that can be optimized to select observation noise and feature network prior precisions. Moreover, the corresponding linearized posterior predictive (Foong et al., 2019; Immer et al., 2021b) is known to provide calibrated estimates of uncertainty (Daxberger et al., 2021). The first step is to linearize the Bayesian NAM, $f(\mathbf{x};\, \boldsymbol{\theta})$, around a parameter estimate $\boldsymbol{\theta}^*$,

$$f^{\text{lin}}(\mathbf{x};\, \boldsymbol{\theta}) = f(\mathbf{x};\, \boldsymbol{\theta}^*) + \mathbf{J}(\mathbf{x};\, \boldsymbol{\theta}^*)(\boldsymbol{\theta} - \boldsymbol{\theta}^*) = \textstyle\sum_d f_d(x_d;\, \boldsymbol{\theta}_d^*) + \mathbf{J}_d(x_d;\, \boldsymbol{\theta}_d^*)(\boldsymbol{\theta}_d - \boldsymbol{\theta}_d^*), \tag{3}$$

where $\mathbf{J}(\mathbf{x};\, \boldsymbol{\theta})$ denotes the Jacobian of the Bayesian NAM. Taking the second derivative of the negative log likelihood $\lambda_n = -\frac{\partial^2}{\partial f^2} \log p(y_n \mid f_n)$ yielding $n$-dimensional vector $\boldsymbol{\lambda}$, and taking the $\mathbf{J} \in \mathbb{R}^{N \times P}$ stacked Jacobians, we have as log marginal likelihood approximation

$$\begin{aligned} \log p(\mathcal{D} \mid \boldsymbol{\delta}) &\approx \log p(\mathcal{D} \mid \boldsymbol{\theta}^*) + \log p(\boldsymbol{\theta}^* \mid \boldsymbol{\delta}) - \tfrac{1}{2} \log |\mathbf{J}^\intercal \operatorname{diag}[\boldsymbol{\lambda}] \mathbf{J} + \operatorname{diag}[\boldsymbol{\delta}]| + \tfrac{P}{2} \log 2\pi \\ &\leq \log p(\mathcal{D} \mid \boldsymbol{\theta}^*) + \textstyle\sum_d \log p(\boldsymbol{\theta}_d^* \mid \delta_d) - \tfrac{1}{2} \log |\underbrace{\mathbf{J}_d^\intercal \operatorname{diag}[\boldsymbol{\lambda}] \mathbf{J}_d + \delta_d \mathbf{I}}_{\stackrel{\text{def}}{=} \mathbf{P}_d}| + \tfrac{P_d}{2} \log 2\pi, \end{aligned} \tag{4}$$

where $\operatorname{diag}[\,\cdot\,]$ turns a vector into the corresponding diagonal matrix. The Laplace approximation separates over the $D$ networks yielding additive structure and a lower bound in the approximate log marginal likelihood (Immer et al., 2023). The corresponding approximate posterior is given by $q(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \mathcal{N}(\boldsymbol{\theta};\, \boldsymbol{\theta}^*,\, \boldsymbol{\Sigma})$ with block-diagonal covariance matrix

$$\boldsymbol{\Sigma} \stackrel{\text{def}}{=} \begin{bmatrix} \boldsymbol{\Sigma}_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \boldsymbol{\Sigma}_D \end{bmatrix} = \begin{bmatrix} \mathbf{P}_1^{-1} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{P}_D^{-1} \end{bmatrix}. \tag{5}$$

#### 2.2.1. Automatic Feature Net Selection

Automatic comparison and selection of feature networks is attained by virtue of adaptive regularization and optimization of the regularization strength using the approximate log marginal likelihood. This procedure is closely related to automatic relevance determination (MacKay, 1994; Neal, 1995) in which parameters of the first layer are grouped according to the input feature and regularized. In this case, this applies to the entire individual feature networks and provides a method for implicitly comparing and selecting features. In Fig. 1, we show that uninformative features of a prediction task can be entirely ignored by our method. Ignoring noisy or uninformative features can improve interpretability of the resulting model as it moves focus to a subset of relevant features.

We maximize the log marginal likelihood w.r.t. the prior precision, $\max_{\boldsymbol{\delta}} \log p(\mathcal{D} \mid \boldsymbol{\delta})$, by taking gradient-based updates during training (Immer et al., 2021a)[1]:

$$\frac{\partial}{\partial \delta_d} \log p(\mathcal{D} \mid \boldsymbol{\delta}) = \frac{P_d}{\delta_d} - \|\boldsymbol{\theta}_d^*\|_2^2 - \operatorname{Tr}(\boldsymbol{\Sigma}_d). \tag{6}$$

An intuition of the corresponding closed-form update derived by MacKay (1991) is given in Tipping (2001). The optimal value of $\delta_d$ is a measure of the concentration of $\boldsymbol{\Sigma}_d$ relative to the prior and depends on how well the data determines the parameters $\boldsymbol{\theta}_d$.

#### 2.2.2. Predictive Decomposition per Feature

The linearized Laplace approximation also yields functional uncertainty estimates which can further aid interpretability (Bhatt et al., 2021). In particular, we can visualize the epistemic uncertainty of the individual feature networks. Fig. 2 shows that the uncertainty tends to be higher when less observations are present, an important aspect which is missed when only predicting the mean contribution to $f$. For a newly observed sample $\mathbf{x}^*$, we have that the predictive variance of $f^*$ is given by

$$\mathbb{V}[f^{\text{lin},*} \mid \mathbf{x}^*] = \mathbf{J}(\mathbf{x}^*;\ \boldsymbol{\theta}^*)^\intercal \boldsymbol{\Sigma}\, \mathbf{J}(\mathbf{x}^*;\ \boldsymbol{\theta}^*) = \sum_{d=1}^{D} \mathbf{J}_d(x_d^*) \boldsymbol{\Sigma}_d \mathbf{J}_d(x_d^*) \overset{\text{def}}{=} \sum_{d=1}^{D} \mathbb{V}[f_d^{\text{lin},*} \mid x_d^*], \tag{7}$$

a consequence of the block-diagonal structure of the posterior covariance $\boldsymbol{\Sigma}$. See Appendix B for a discussion on the importance of the independence of feature networks.

## 3. Experiments

We empirically evaluate the proposed LA-NAM on a collection of synthetic and real-world benchmarks. We compare against the NAM of Agarwal et al. (2021) and other popular methods of the generalized additive model class, namely, a smoothing spline model with smoothing parameters selected via cross-validation (GAM; Hastie and Tibshirani, 1999; Servén et al., 2018) and an implementation of a gradient-boosting model using the hyperparameters recommended by the authors (EBM; Lou et al., 2012; Nori et al., 2019). The epistemic uncertainty of the model is determined by bootstrapping for the EBM and by taking the standard deviation of recovered functions across ensemble members for the NAM. Further experimental details are provided in Appendix A.4.

1. We have also experimented with MacKay's updates (MacKay, 1991) and obtained similar results.

| Dataset | Linear | GAM | EBM | NAM (ReLU) | NAM (ExU) | LA-NAM (Ours) |
|---|---|---|---|---|---|---|
| autompg | 2.59 (±0.06) | **2.43 (±0.09)** | 2.64 (±0.10) | **2.47 (±0.09)** | 2.69 (±0.16) | **2.46 (±0.08)** |
| concrete | 3.78 (±0.04) | **3.13 (±0.05)** | **3.20 (±0.12)** | **3.21 (±0.06)** | 3.46 (±0.12) | 3.25 (±0.03) |
| energy | 2.46 (±0.02) | **1.46 (±0.02)** | **1.46 (±0.02)** | 1.48 (±0.02) | 1.48 (±0.02) | **1.44 (±0.02)** |
| kin8nm | -0.18 (±0.01) | **-0.20 (±0.01)** | **-0.20 (±0.01)** | **-0.20 (±0.01)** | -0.18 (±0.01) | **-0.20 (±0.00)** |
| naval | -3.72 (±0.01) | **-8.09 (±0.02)** | -3.15 (±0.00) | -4.67 (±0.02) | -3.87 (±0.01) | -7.24 (±0.01) |
| wine | **1.00 (±0.03)** | **0.98 (±0.03)** | **0.99 (±0.03)** | **0.98 (±0.03)** | **1.02 (±0.04)** | **0.98 (±0.03)** |
| yacht | 3.64 (±0.07) | **1.87 (±0.10)** | **1.93 (±0.13)** | **1.95 (±0.12)** | 2.24 (±0.08) | **1.81 (±0.10)** |
| australian | **0.35 (±0.02)** | **0.35 (±0.04)** | **0.33 (±0.03)** | **0.35 (±0.02)** | **0.38 (±0.04)** | **0.34 (±0.03)** |
| breast | **0.10 (±0.01)** | **0.09 (±0.02)** | **0.12 (±0.02)** | 0.12 (±0.01) | 0.16 (±0.03) | **0.10 (±0.02)** |
| heart | **0.39 (±0.04)** | 0.43 (±0.08) | 0.40 (±0.02) | 0.41 (±0.03) | 0.41 (±0.04) | **0.33 (±0.02)** |
| ionosphere | 0.33 (±0.03) | 0.27 (±0.02) | **0.23 (±0.02)** | **0.23 (±0.02)** | 0.31 (±0.04) | **0.25 (±0.04)** |
| parkinsons | 0.33 (±0.02) | 0.36 (±0.02) | **0.28 (±0.05)** | **0.28 (±0.02)** | **0.29 (±0.04)** | **0.26 (±0.03)** |

Table 1: Negative test log likelihood (lower is better) on UCI regression (top) and classification (bottom) benchmarks. Our proposed LA-NAM is competitive with the best baselines and often outperforms the non-Bayesian NAM.

### 3.1. Illustrative Example

First, we illustrate the recovery of purely additive structure from noisy data by constructing a synthetic regression dataset for which the true additive terms are known. Generalized additive models should be able to recover the additive functions of such a dataset precisely since it is designed in such a way that there are no interactions between the input features.

In Fig. 1, we show the recovered functions $f_1$ and $f_4$ along with the quadratic ground truth for $f_1$ and constant for $f_4$. The ReLU variant of the NAM tends to fit the data better, but has bad epistemic uncertainty due to poor diversity among its ensemble members. The ExU variant has better diversity but yields a very poor mean fit. In contrast, the proposed LA-NAM fits the data accurately all-the-while maintaining a good estimate of epistemic uncertainty. It is also less susceptible to misattributing noise to the functions compared to the other baselines. This is particularly striking for the uninformative $f_4$, since only the LA-NAM correctly predicts that it should have no effect in this example. The full details and visualization of this experiment are deferred to Appendix A.1.

### 3.2. UCI Regression and Binary Classification

We benchmark the LA-NAM and baselines on a selection of UCI regression and binary classification tasks. Each UCI dataset is split into 5 cross-validation folds. We use the library defaults for the EBM (Nori et al., 2019) and retain 15% of the training data as validation for the NAM. Extra validation data is not needed for the LA-NAM and GAM since the former is tuned using the estimated log marginal likelihood and the latter through generalized cross-validation scoring (GCV; Golub et al., 1979).

In Table 1, we report the average negative log likelihood and standard error across folds. The NAM and EBM baselines do not provide an estimate of the observation noise in regression, so we have assigned them a maximum likelihood fit using their training data. We also provide results for linear and logistic regression to determine the performance attainable without resorting to nonlinear relationships. The poor log likelihood of the ExU NAM compared to ReLU suggests that it is badly calibrated due to overly aggressive fitting.

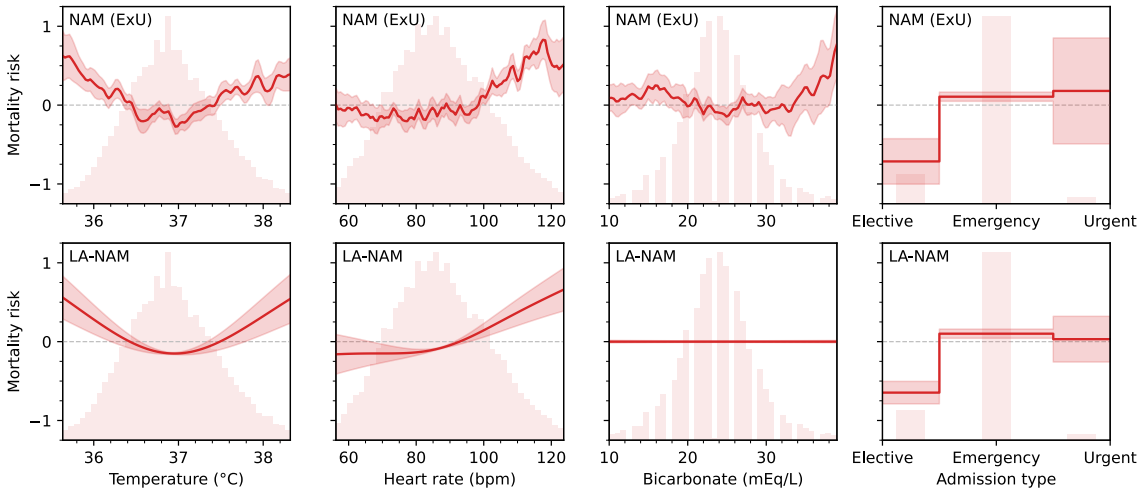


Figure 2: Risk of mortality and associated epistemic uncertainty (± 2 std. deviations) on the MIMIC-III mortality prediction task. The LA-NAM generates smoother and thus more interpretable feature curves, provides useful uncertainties, and ignores the uninformative feature.

In regression, the LA-NAM performs comparably well when compared to the ReLU NAM, and consistently outperforms the ExU NAM. In classification it outperforms all baselines, including places where the ReLU NAM performs worse than logistic regression.

### 3.3. MIMIC-III Mortality Prediction

Finally, we explore our method's behavior in a real world clinical setting. Using the pre-processing from Lengerich et al. (2022) of the MIMIC-III database (Johnson et al., 2016), we predict the mortality risk and associated uncertainty 24 hours after admission into the intensive care unit. Here, we compare the LA-NAM to the NAM with ExU activation since our objective is to obtain useful uncertainty estimates from the recovered additive structure. On this task, the LA-NAM outperforms the ExU NAM obtaining a negative log likelihood of 0.264 compared to 0.274, and an area under the ROC curve of 79.6% to 78.9% on a 70-30% train-test split. In Fig. 2, we show a subset of the additive structure recovered from this dataset (Complete model is shown in Appendix A.2). In the background of each subplot, we display the histogram of the distribution of feature values. We find that the epistemic uncertainty of the LA-NAM is consistent with the presence of samples or lack thereof.

Additionally, this experiment empirically demonstrates that the LA-NAM can decide on the usefulness of features by selecting feature nets: Because of their linear dependency, both high bicarbonate levels and low anion gap are indicators of metabolic acidosis (Kraut and Madias, 2010), but in this case, the LA-NAM has determined that the risk associated with bicarbonate can be determined solely through the measurement of the anion gap. This is why bicarbonate is fully ignored by the LA-NAM in Fig. 2. See Appendix A.3 for an ablation experiment confirming this.

## 4. Conclusion

In this work, we have shown that using linearized Laplace inference in neural additive models leads to a natural decomposition of the epistemic uncertainty of the additive subnetworks, and enables implicit selection of features when optimizing the log marginal likelihood. We have provided evidence that the proposed Laplace-approximated neural additive models (la-nam) are more robust to noise and easier to interpret than their non-Bayesian counterpart, and are thus viable alternatives for use in safety-critical settings and as tools for data-driven scientific discovery. Ultimately, we hope that this work inspires future research at the intersection of interpretable machine learning and Bayesian inference.

### Acknowledgments

We thank Ben Lengerich for providing us with the pre-processed version of the MIMIC-III dataset used in the paper. A.I. gratefully acknowledges funding by the Max Planck ETH Center for Learning Systems (CLS). V.F. was supported by a Branco Weiss Fellowship.

This project was supported by grant #2022-278 of the Strategic Focus Area "Personalized Health and Related Technologies (PHRT)" of the ETH Domain.

## References

Rishabh Agarwal, Levi Melnick, Nicholas Frosst, Xuezhou Zhang, Ben Lengerich, Rich Caruana, and Geoffrey E Hinton. Neural Additive Models: Interpretable Machine Learning with Neural Nets. In *Advances in Neural Information Processing Systems*, volume 34, pages 4699–4711. Curran Associates, Inc., 2021.

Javier Antorán, David Janz, James U Allingham, Erik Daxberger, Riccardo Rb Barbano, Eric Nalisnick, and José Miguel Hernández-Lobato. Adapting the linearised laplace model evidence for modern deep learning. In *International Conference on Machine Learning*, pages 796–821. PMLR, 2022.

Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, et al. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 401–413, 2021.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In *ICML*, 2015.

Leo Breiman and Jerome H. Friedman. Estimating Optimal Transformations for Multiple Regression and Correlation. *Journal of the American Statistical Association*, 80(391): 580–598, September 1985. ISSN 0162-1459. doi: 10.1080/01621459.1985.10478157.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1721–1730, New York, NY,

USA, August 2015. Association for Computing Machinery. ISBN 978-1-4503-3664-2. doi: 10.1145/2783258.2788613.

Kamil Ciosek, Vincent Fortuin, Ryota Tomioka, Katja Hofmann, and Richard Turner. Conservative uncertainty estimation by fitting prior networks. In *International Conference on Learning Representations*, 2020.

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, December 1989. ISSN 1435-568X. doi: 10.1007/BF02551274.

Francesco D'Angelo and Vincent Fortuin. Repulsive deep ensembles are bayesian. *Advances in Neural Information Processing Systems*, 34:3451–3465, 2021.

Francesco D'Angelo, Vincent Fortuin, and Florian Wenzel. On stein variational neural network ensembles. *arXiv preprint arXiv:2106.10760*, 2021.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace Redux - Effortless Bayesian Deep Learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 20089–20103. Curran Associates, Inc., 2021.

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*, 2019.

Vincent Fortuin. Priors in bayesian deep learning: A review. *International Statistical Review*, 90(3):563–591, 2022.

Vincent Fortuin, Adrià Garriga-Alonso, Mark van der Wilk, and Laurence Aitchison. Bnnpriors: A library for bayesian neural network inference with different prior distributions. *Software Impacts*, 9:100079, 2021.

Vincent Fortuin, Adrià Garriga-Alonso, Sebastian W Ober, Florian Wenzel, Gunnar Ratsch, Richard E Turner, Mark van der Wilk, and Laurence Aitchison. Bayesian neural network priors revisited. In *International Conference on Learning Representations*, 2022.

Jerome H. Friedman. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. ISSN 0090-5364.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016.

Adrià Garriga-Alonso and Vincent Fortuin. Exact Langevin dynamics with stochastic gradients. *arXiv preprint arXiv:2102.01691*, 2021.

Gene H. Golub, Michael Heath, and Grace Wahba. Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter. *Technometrics*, 21(2):215–223, May 1979. ISSN 0040-1706, 1537-2723. doi: 10.1080/00401706.1979.10489751.

Alex Graves. Practical variational inference for neural networks. In *NIPS*, 2011.

Trevor Hastie and Robert Tibshirani. *Generalized Additive Models*. Chapman & Hall/CRC, Boca Raton, Fla, 1999. ISBN 978-0-412-34390-2.

Bobby He, Balaji Lakshminarayanan, and Yee Whye Teh. Bayesian deep ensembles via the neural tangent kernel. *Advances in neural information processing systems*, 33:1010–1022, 2020.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016.

Alexander Immer, Matthias Bauer, Vincent Fortuin, Gunnar Rätsch, and Khan Mohammad Emtiyaz. Scalable Marginal Likelihood Estimation for Model Selection in Deep Learning. In *Proceedings of the 38th International Conference on Machine Learning*, pages 4563–4573. PMLR, July 2021a.

Alexander Immer, Maciej Korzepa, and Matthias Bauer. Improving predictions of Bayesian neural nets via local linearization. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pages 703–711. PMLR, March 2021b.

Alexander Immer, Tycho FA van der Ouderaa, Vincent Fortuin, Gunnar Rätsch, and Mark van der Wilk. Invariance learning in deep neural networks with differentiable Laplace approximations. In *NeurIPS*, 2022.

Alexander Immer, Tycho F. A. Van Der Ouderaa, Mark Van Der Wilk, Gunnar Ratsch, and Bernhard Schölkopf. Stochastic marginal likelihood gradients using neural tangent kernels. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *34th Conference on Uncertainty in Artificial Intelligence 2018, UAI 2018*, pages 876–885, 2018.

Pavel Izmailov, Sharad Vikram, Matthew D Hoffman, and Andrew Gordon Wilson. What are Bayesian neural network posteriors really like? In *ICML*, 2021.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassem i, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035, May 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.35.

Laurent Valentin Jospin, Hamid Laga, Farid Boussaid, Wray Buntine, and Mohammed Bennamoun. Hands-on bayesian neural networks - a tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 17(2):29–48, 2022.

Mohammad Khan, Didrik Nielsen, Voot Tangkaratt, Wu Lin, Yarin Gal, and Akash Srivastava. Fast and scalable Bayesian deep learning by weight-perturbation in Adam. In *ICML*, 2018.

Mohammad Emtiyaz E Khan, Alexander Immer, Ehsan Abedi, and Maciej Korzepa. Approximate Inference Turns Deep Networks into Gaussian Processes. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

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

Diederik P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. In *Advances in Neural Information Processing Systems 28*, 2015.

Jeffrey A. Kraut and Nicolaos E. Madias. Metabolic acidosis: Pathophysiology, diagnosis and management. *Nature Reviews Nephrology*, 6(5):274–285, May 2010. ISSN 1759-507X. doi: 10.1038/nrneph.2010.33.

Jeffrey A. Kraut and Nicolaos E. Madias. Treatment of acute metabolic acidosis: A pathophysiologic approach. *Nature Reviews. Nephrology*, 8(10):589–601, October 2012. ISSN 1759-507X. doi: 10.1038/nrneph.2012.186.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NIPS*, 2017.

Pierre Simon Laplace. Mémoire sur la probabilité des causes par les évènements. In *Mémoires de l'Académie Royale Des Sciences de Paris (Savants Étrangers)*, volume 6, pages 621–656. 1774.

Benjamin J. Lengerich, Rich Caruana, Mark E. Nunnally, and Manolis Kellis. Death by Round Numbers: Glass-Box Machine Learning Uncovers Biases in Medical Practice, November 2022.

Yin Lou, Rich Caruana, and Johannes Gehrke. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 150–158, New York, NY, USA, August 2012. Association for Computing Machinery. ISBN 978-1-4503-1462-6. doi: 10.1145/2339530.2339556.

Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 623–631, New York, NY, USA, August 2013. Association for Computing Machinery. ISBN 978-1-4503-2174-7. doi: 10.1145/2487575.2487579.

Christos Louizos and Max Welling. Structured and efficient variational deep learning with matrix Gaussian posteriors. In *ICML*, 2016.

Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The Expressive Power of Neural Networks: A View from the Width. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Mattias Luber, Anton Thielmann, and Benjamin Säfken. Structural Neural Additive Models: Enhanced Interpretable Machine Learning, February 2023.

Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

David J. C. MacKay. Bayesian model comparison and backprop nets. In J. Moody, S. Hanson, and R.P. Lippmann, editors, *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann, 1991.

David J. C. MacKay. A practical Bayesian framework for backpropagation networks. *Neural Computation*, 4(3), 1992.

David J. C. MacKay. Bayesian nonlinear modeling for the prediction competition. *ASHRAE transactions*, 100(2):1053–1062, 1994.

Wesley J Maddox, Pavel Izmailov, Timur Garipov, Dmitry P Vetrov, and Andrew Gordon Wilson. A simple baseline for Bayesian uncertainty in deep learning. In *NeurIPS*, 2019.

Vitaly Maiorov and Allan Pinkus. Lower bounds for approximation by MLP neural networks. *Neurocomputing*, 25(1):81–91, April 1999. ISSN 0925-2312. doi: 10.1016/S0925-2312(98)00111-8.

James Martens and Roger Grosse. Optimizing Neural Networks with Kronecker-factored Approximate Curvature. In *International conference on machine learning*, pages 2408–2417. PMLR, 2015.

Seth Nabarro, Stoil Ganev, Adrià Garriga-Alonso, Vincent Fortuin, Mark van der Wilk, and Laurence Aitchison. Data augmentation in bayesian neural networks and the cold posterior effect. In *Uncertainty in Artificial Intelligence*, pages 1434–1444. PMLR, 2022.

Radford M Neal. Bayesian learning via stochastic dynamics. In *Advances in neural information processing systems*, pages 475–482, 1993.

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

Radford M Neal et al. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11), 2011.

Harsha Nori, Samuel Jenkins, Paul Koch, and Rich Caruana. InterpretML: A Unified Framework for Machine Learning Interpretability, September 2019.

Kazuki Osawa, Siddharth Swaroop, Mohammad Emtiyaz E Khan, Anirudh Jain, Runa Eschenhagen, Richard E Turner, and Rio Yokota. Practical deep learning with Bayesian principles. In *NeurIPS*, 2019.

Luisa Pumplun, Mariska Fecho, Nihal Wahl, Felix Peters, and Peter Buxmann. Adoption of Machine Learning Systems for Medical Diagnostics in Clinics: Qualitative Interview Study. *Journal of Medical Internet Research*, 23(10):e29301, October 2021. doi: 10.2196/29301.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144, New York, NY, USA, August 2016. Association for Computing Machinery. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939778.

Jonas Rothfuss, Vincent Fortuin, Martin Josifoski, and Andreas Krause. Pacoh: Bayes-optimal meta-learning with pac-guarantees. In *International Conference on Machine Learning*, pages 9116–9126. PMLR, 2021.

Jonas Rothfuss, Martin Josifoski, Vincent Fortuin, and Andreas Krause. Pac-bayesian meta-learning: From theory to practice. *arXiv preprint arXiv:2211.07206*, 2022.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Pola Schwöbel, Martin Jørgensen, Sebastian W Ober, and Mark Van Der Wilk. Last layer marginal likelihood for invariance learning. In *International Conference on Artificial Intelligence and Statistics*, pages 3542–3555. PMLR, 2022.

Daniel Servén, Charlie Brummitt, Hassan Abedi, and hlink. pyGAM: Generalized Additive Models in Python. Zenodo, October 2018.

Mrinank Sharma, Tom Rainforth, Yee Whye Teh, and Vincent Fortuin. Incorporating unlabelled data into bayesian neural networks. *arXiv preprint arXiv:2304.01762*, 2023.

Anton Thielmann, René-Marcel Kruse, Thomas Kneib, and Benjamin Säfken. Neural Additive Models for Location Scale and Shape: A Framework for Interpretable Neural Regression Beyond the Mean, January 2023.

Michael E Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1(Jun):211–244, 2001.

Tycho FA van der Ouderaa and Mark van der Wilk. Learning invariant weights in neural networks. In *Uncertainty in Artificial Intelligence*, pages 1992–2001. PMLR, 2022.

Michael Veale, Max Van Kleek, and Reuben Binns. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 1–14, New York, NY, USA, April 2018. Association for Computing Machinery. doi: 10.1145/3173574.3174014.

Grace Wahba. Bayesian "Confidence Intervals" for the Cross-Validated Smoothing Spline. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(1):133–150, 1983. ISSN 2517-6161. doi: 10.1111/j.2517-6161.1983.tb01239.x.

Ziyu Wang, Tongzheng Ren, Jun Zhu, and Bo Zhang. Function space particle optimization for bayesian neural networks. In *International Conference on Learning Representations*, 2019.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *ICML*, 2011.

Andrew Gordon Wilson and Pavel Izmailov. Bayesian deep learning and a probabilistic perspective of generalization. In *NeurIPS*, 2020.

Shiyun Xu, Zhiqi Bu, Pratik Chaudhari, and Ian J. Barnett. Sparse Neural Additive Model: Interpretable Deep Learning with Feature Selection via Group Sparsity, February 2022.

Zebin Yang, Aijun Zhang, and Agus Sudjianto. GAMI-Net: An Explainable Neural Network based on Generalized Additive Models with Structured Interactions. *Pattern Recognition*, 120:108192, December 2021. ISSN 0031-3203. doi: 10.1016/j.patcog.2021.108192.

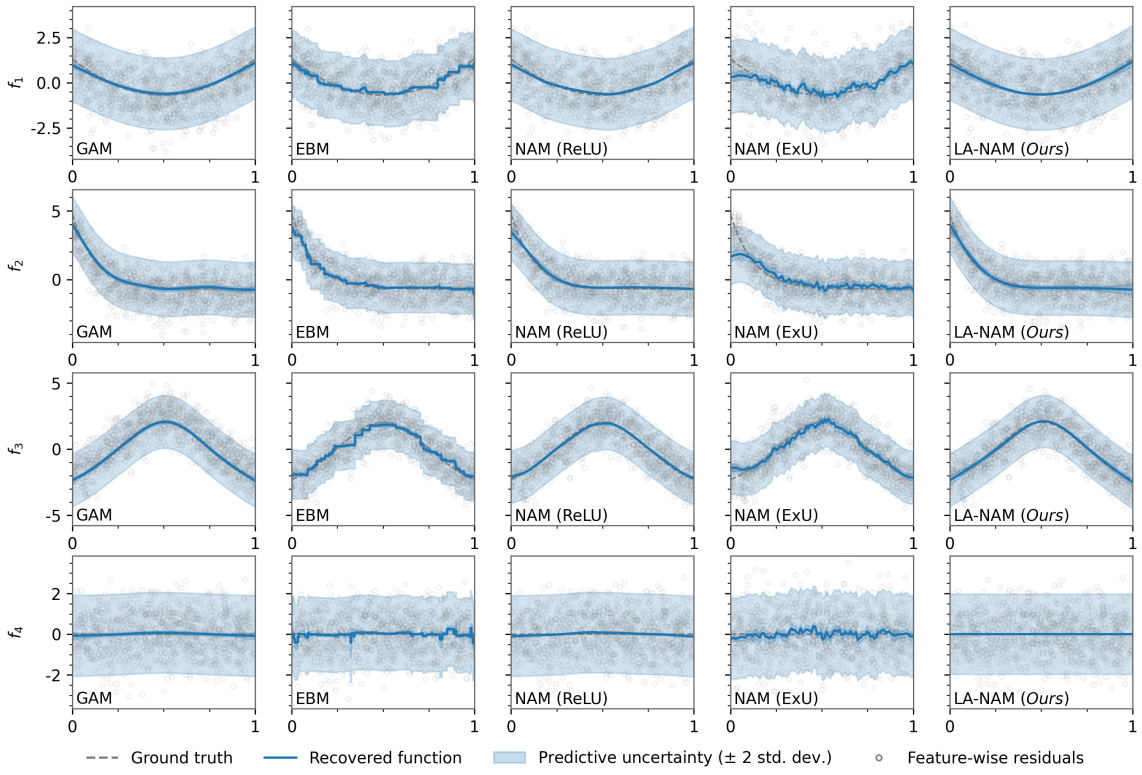


Figure 3: Recovery of the additive structure of the synthetic dataset of A.1. The feature-wise residuals are the generated data points with the mean contribution of the other feature networks subtracted.

## Appendix A. Experimental Details

### A.1. Dataset of the Illustrative Example

In Section 3.1 we present an illustrative example to motivate the capacity of the LA-NAM and baselines to recover purely additive structure from noisy data. We provide further details on the generation of the synthetic dataset used here. Consider the function $\hat{f} : \mathbb{R}^4 \to \mathbb{R}$, where $\hat{f}(x_1, x_2, x_3, x_4) = \hat{f}_1(x_1) + \hat{f}_2(x_2) + \hat{f}_3(x_3) + \hat{f}_4(x_4)$, and

$$\hat{f}_1(x_1) = 8(x_1 - \tfrac{1}{2})^2, \qquad \hat{f}_2(x_2) = \tfrac{1}{10}\exp[-8x_2 + 4] \tag{8}$$

$$\hat{f}_3(x_3) = 5\exp[-2(2x_3 - 1)^2], \qquad \hat{f}_4(x_4) = 0. \tag{9}$$

We generate $N = 1\,000$ noisy observations $\{\mathbf{x}_n, y_n\}_{n=1}^N$ by sampling inputs $\mathbf{x}_n$ uniformly from $\mathcal{U}([0, 1]^4)$ and generating targets $y_n = \hat{f}(\mathbf{x}_n) + \epsilon_n$, where $\epsilon_n \sim \mathcal{N}(0, 1)$ is random Gaussian noise. Fig. 3 shows the recovered functions along with the associated predictive uncertainty for the LA-NAM and baseline models.

## A.2. Predicted Mortality Risk in MIMIC-III

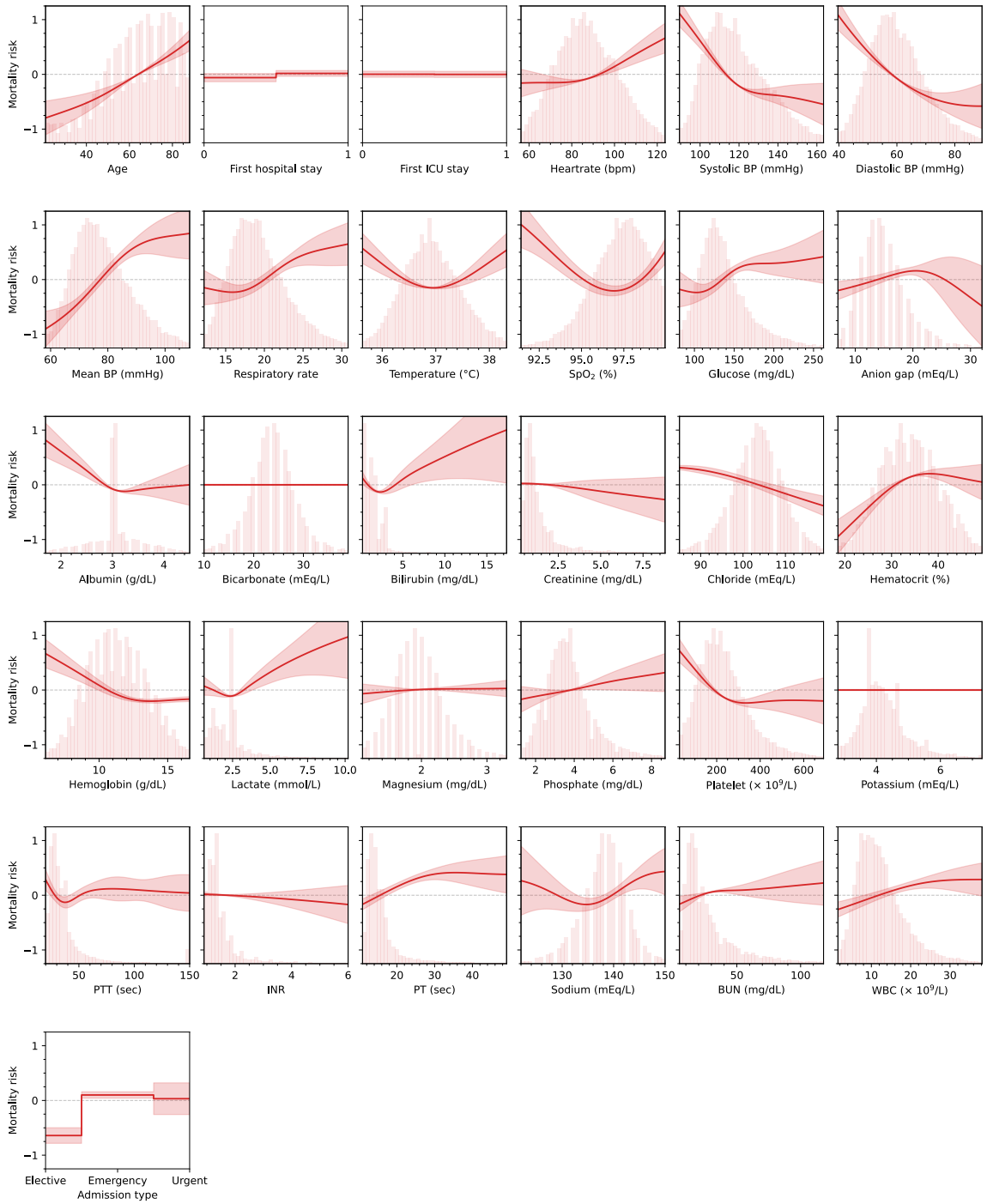


Figure 4: Complete visualization of predicted mortality risk in the LA-NAM of Section 3.3.

### A.3. Ablation of the Anion Gap in MIMIC-III

The anion gap is a measure of the serum concentration of sodium subtracted with the serum concentrations of chloride and bicarbonate,

$$\text{Anion gap} = [\text{Na}^+] - ([\text{Cl}^-] + [\text{HCO}_3^-]) \text{ mEq/L}. \tag{10}$$

Both low bicarbonate levels and thus high anion gap are indicators of acute metabolic acidosis. This is a known risk factor for intensive care mortality with very poor prognosis (Kraut and Madias, 2010, 2012). Fig. 5 shows that the predicted mortality risk increases steadily as the anion gap grows but becomes uncertain above 20 mEq/L due to low sample size.

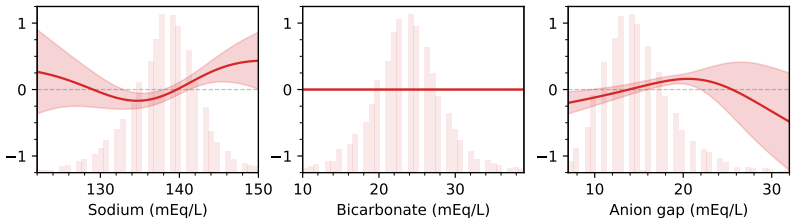


Figure 5: Sodium, bicarbonate and anion gap mortality risk predicted by the LA-NAM.

When presented both anion gap and bicarbonate in the mortality risk dataset of Section 3.3 the LA-NAM uses high anion gap as a proxy for the risk of low bicarbonate. We confirm this visually by performing an ablation experiment in which the LA-NAM is re-trained with the feature network attending to the anion gap removed. Fig. 6 shows that in the ablated model the anion gap risk is moved into the low levels for bicarbonate. The bicarbonate risk increases below 20 mEq/L and becomes uncertain around 15 mEq/L.

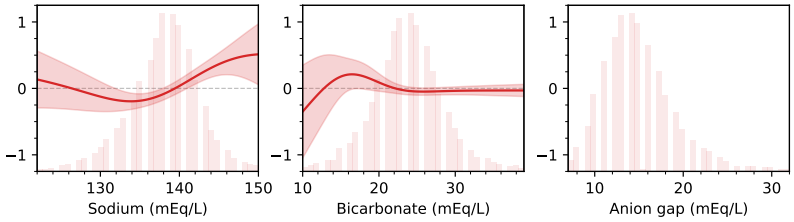


Figure 6: Sodium and bicarbonate mortality risk withs anion gap feature network ablated.

### A.4. Experimental Setup

We provide additional details on the choice of implementation and hyperparameters used in the experiments of this paper:

**GAM.** We use an open source implementation (`pygam`; Servén et al., 2018). The smoothing parameters are grid-searched. We sample one thousand candidates uniformly from the recommended range of $[10^{-3}, 10^3]$ and select using generalized cross-validation scoring (GCV; Golub et al., 1979).

**EBM.** For the gradient-boosted-based generalized additive model, we use the explainable boosting machine (EBM) which is an open source and modern implementation that is available as a part of the `InterpretML` library (Nori et al., 2019). The library defaults are used for the hyperparameters.

**NAM.** We test both the standard ReLU variant with hidden layer sizes of 64, 64 and 32, and the exponential ExU variant with ReLU-1 activation and a single hidden layer containing 1024 units. We perform grid-search for selecting learning rate and regularization hyperparameters. All other hyperparameters are taken from recommended training procedure in the supplementary material of Agarwal et al. (2021).

**LA-NAM.** The LA-NAM is constructed using feature nets containing a single hidden layer of 64 neurons with GELU activation (Hendrycks and Gimpel, 2016). The feature network parameters and hyperparameters (prior precision, observation noise) are optimized using Adam (Kingma and Ba, 2014), alternating between optimizing both at regular intervals as in Immer et al. (2021a). We select the learning rate in the discrete set of {0.1, 0.01, 0.001} which maximizes the ultimate log marginal likelihood. We use a batch size of 512 and peform early stopping on the log marginal likelihood restoring the best scoring parameters at the end of training. We find that the algorithm is fairly robust to the choice of hyperparameter optimization schedule: We use 0.1 for the hyperparameter learning rate and perform batches of 30 gradient steps on the log marginal likelihood every 100 epochs of regular training.

### A.5. Additional Results on UCI Datasets

We report additional metrics for the 5 cross-validation folds of the UCI regression and classification benchmarks of Section 3.2 in Table 2 and Table 3.

| **Dataset** | Linear | GAM | EBM | NAM (ReLU) | NAM (ExU) | LA-NAM (Ours) |
|---|---|---|---|---|---|---|
| `autompg` | 3.18 (±0.18) | **2.70 (±0.20)** | **2.98 (±0.11)** | **2.76 (±0.16)** | **2.94 (±0.19)** | **2.77 (±0.18)** |
| `concrete` | 10.50 (±0.40) | **5.57 (±0.24)** | **5.15 (±0.25)** | 5.90 (±0.27) | 6.89 (±0.45) | 6.27 (±0.13) |
| `energy` | 2.84 (±0.05) | **1.04 (±0.02)** | **1.04 (±0.02)** | **1.06 (±0.02)** | **1.06 (±0.02)** | **1.04 (±0.02)** |
| `kin8nm` | 0.20 (±0.00) | **0.20 (±0.00)** | **0.20 (±0.00)** | **0.20 (±0.00)** | 0.20 (±0.00) | **0.20 (±0.00)** |
| naval | 6e-3 (±4e-5) | **7e-5 (±2e-6)** | 1e-2 (±5e-5) | 2e-3 (±5e-5) | 5e-3 (±5e-5) | 2e-4 (±2e-6) |
| `wine` | **0.65 (±0.02)** | **0.65 (±0.01)** | **0.64 (±0.02)** | **0.64 (±0.02)** | **0.64 (±0.02)** | **0.64 (±0.02)** |
| `yacht` | 9.09 (±0.54) | **1.51 (±0.16)** | **1.56 (±0.16)** | **1.61 (±0.19)** | 2.20 (±0.15) | **1.45 (±0.17)** |

Table 2: Root mean squared error (RMSE, ± std. error) on UCI regression datasets.

| **Dataset** | Linear | GAM | EBM | NAM (ReLU) | NAM (ExU) | LA-NAM (Ours) |
|---|---|---|---|---|---|---|
| `australian` | **92.5 (±1.0)** | **91.9 (±1.5)** | **93.2 (±1.3)** | **92.5 (±0.9)** | **92.0 (±1.0)** | **92.6 (±1.1)** |
| `breast` | **99.6 (±0.2)** | **99.4 (±0.4)** | **99.2 (±0.2)** | **99.5 (±0.3)** | 99.0 (±0.4) | **99.4 (±0.2)** |
| `heart` | **90.0 (±2.4)** | 89.1 (±2.8) | 90.3 (±1.4) | **91.0 (±1.9)** | **90.3 (±2.0)** | **93.5 (±1.4)** |
| `ionosphere` | 90.4 (±2.1) | **95.4 (±0.9)** | **96.3 (±0.8)** | **96.3 (±0.9)** | **95.1 (±1.3)** | **94.5 (±1.3)** |
| `parkinsons` | 90.0 (±2.2) | 88.5 (±2.6) | **94.6 (±2.0)** | **94.3 (±1.8)** | **94.6 (±1.4)** | **94.5 (±1.7)** |

Table 3: Area under the ROC curve (AUROC, ± std. error) on UCI classification datasets.

## Appendix B. Network Independence

In the LA-NAM, we apply the Laplace approximation independently across feature networks and obtain a block-diagonal posterior covariance matrix on which the decomposition of predictive variance in Eq. (7) depends. Here, we further discuss the importance of feature network independence in the posterior distribution.

As a toy experiment, suppose we wanted to obtain an estimate of two variable terms, $b_1$ and $b_2$, such that their sum is equal to some constant $C$. We also desire that neither term $b_1$ or $b_2$ dominate the other, such that they are approximately equally balanced. One possible setup for finding maximum a posteriori (MAP) estimates for $b_1$ and $b_2$ could be to design a cost function $L(b_1, b_2)$ such that the MAP solution is its minimizer,

$$p(b_1, b_2) = \mathcal{N}([b_1, b_2]^\top; 0, \lambda^{-1}\mathbf{I}), \quad p(C \mid b_1, b_2) = \mathcal{N}(C; b_1 + b_2, 1), \tag{11}$$

$$\log p(b_1, b_2 \mid C) \propto \log p(C \mid b_1, b_2) + \log p(b_1, b_2) \tag{12}$$

$$\propto -(b_1 + b_2 - C)^2 - \lambda(b_1^2 + b_2^2) \quad \stackrel{\text{def}}{=} -L(b_1, b_2). \tag{13}$$

For demonstrative purposes we choose to take $C = 20$ and $\lambda = 0.01$. In the left of Fig. 7, we show the values of the cost function $L(b_1, b_2)$, along with the corresponding MAP solution displayed as a white cross.

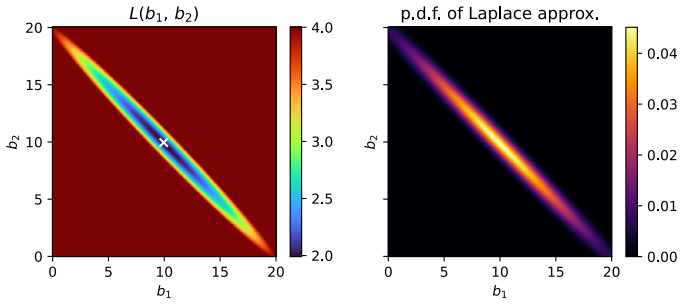


Figure 7: Laplace approximation of the toy example of Appendix B. ($C = 20$, $\lambda = 0.01$)

Now suppose we are interested in finding an approximate posterior distribution for $b_1$ and $b_2$ using a Laplace approximation centered at our MAP estimate. If we consider $b_1$ and $b_2$ jointly, i.e., assume that they are dependent of one another, then we obtain the Gaussian approximate posterior shown on the right of Fig. 7. In this approximation, $b_1$ and $b_2$ are strongly anti-correlated, which can be explained by the fact that translating $b_1$ by some amount $\Delta$ can be accounted for by translating $b_2$ by $-\Delta$, since $b_1 + b_2 = (b_1 + \Delta) + (b_2 - \Delta)$. The variance of $b_1$ and $b_2$ is overtaken by all possible translations $\Delta$.

In the context of Bayesian NAMs, this is an undesirable property. We desire that the credible intervals for a feature network $f_d$ depend only on possible changes of its shape and not translations accounted for in other feature networks. In the example above, this can be avoided by assuming that $b_1$ and $b_2$ are independent by performing a first Laplace approximation for $b_1$ keeping $b_2$ frozen, and then another for $b_2$ keeping $b_1$ frozen. In the LA-NAM the feature networks are guaranteed to be independent given that we have used block-diagonal approximation Kronecker-factored GGN matrix.

## Appendix C. Related Work

**Generalized additive models.** Several constructions have been proposed for the generalized additive models of Hastie and Tibshirani (1999). Originally it was proposed to construct these models using smoothing splines (Wahba, 1983) and to fit in an iterative fashion using "backfitting" (Breiman and Friedman, 1985). One alternative is to construct the smoothing functions using gradient-boosted decision trees (Friedman, 2001). The boosting algorithm can be modified to cycle through functions in its inner loop, which was shown to be preferable to a sequential backfitting of boosted trees (Lou et al., 2012). Boosted trees also enable selection and fitting of feature interactions by separating training into multiple stages (Lou et al., 2013). These second-order interactions are believed to enable gradient-boosted additive models to achieve competitive accuracy when comparing to fully-interacting models for tabular supervised learning (Caruana et al., 2015; Nori et al., 2019).

**Neural additive models.** Neural networks are also compelling candidates for the construction of generalized additive models since it is established through "universal approximation theorems" that they can be made to approximate continuous functions up to arbitrary precision given sufficient complexity (Cybenko, 1989; Maiorov and Pinkus, 1999; Lu et al., 2017). The neural additive model (NAM) proposed by Agarwal et al. (2021) is constructed using ensembles of ReLU and ExU feed-forward networks and fitted through standard backpropagation. The ExU variant, in which weights are learned in logarithmic space, is used for fitting jagged functions. The GAMI-Net proposed around the same time by Yang et al. (2021) is closely related, but single networks are used instead of an ensemble and the model also supports learning of feature interaction terms. A number of extensions have since been suggested: Feature selection through sparse regularization of the feature nets (Xu et al., 2022), generation of confidence intervals using regression spline basis expansion (Luber et al., 2023), and estimation of the skewness, heteroscedasticity, and kurtosis of the underlying data distributions (Thielmann et al., 2023).

**Bayesian neural networks.** Bayesian neural networks promise to marry the expressivity of neural networks with the principled statistical properties of Bayesian inference (MacKay, 1992; Neal, 1993). However, approximate inference in these complex models has remained challenging (Jospin et al., 2022). Approximate inference techniques lie on a spectrum of quality and computational cost, from cheap local approximations like Laplace inference (Laplace, 1774; MacKay, 1992; Khan et al., 2019; Daxberger et al., 2021), stochastic weight averaging (Izmailov et al., 2018; Maddox et al., 2019), and dropout (Gal and Ghahramani, 2016; Kingma et al., 2015), via variational approximations with different levels of complexity (e.g., Graves, 2011; Blundell et al., 2015; Louizos and Welling, 2016; Khan et al., 2018; Osawa et al., 2019), across ensemble-based methods (Lakshminarayanan et al., 2017; Wang et al., 2019; Wilson and Izmailov, 2020; Ciosek et al., 2020; He et al., 2020; D'Angelo et al., 2021; D'Angelo and Fortuin, 2021), up to the very expensive but asymptotically correct Markov Chain Monte Carlo (MCMC) approaches (e.g., Neal, 1993; Neal et al., 2011; Welling and Teh, 2011; Garriga-Alonso and Fortuin, 2021; Izmailov et al., 2021). Apart from the challenges relating to approximate inference, recent work has also studied the question of prior choice for BNNs (e.g., Fortuin et al., 2021, 2022; Nabarro et al., 2022; Sharma et al., 2023; Fortuin, 2022, and references therein) and how to perform model selection in this

framework (e.g., Immer et al., 2021a, 2022; Rothfuss et al., 2021, 2022; van der Ouderaa and van der Wilk, 2022; Schwöbel et al., 2022). In our work, we mainly draw on the linearized Laplace inference (Immer et al., 2021b) and the associated marginal likelihood estimation (Immer et al., 2021a) and apply these methods to the NAM, which to the best of our knowledge has not been tried before.