NOMU: Neural Optimization-based Model Uncertainty

Jakob Heiss [*1,2] **Jakob Weissteiner** [*2,3] **Hanna Wutte** [*1,2] **Sven Seuken** [2,3] **Josef Teichmann** [1,2]

Abstract

We study methods for estimating model uncertainty for neural networks (NNs) in regression. To isolate the effect of model uncertainty, we focus on a noiseless setting with scarce training data. We introduce five important desiderata regarding model uncertainty that any method should satisfy. However, we find that established benchmarks often fail to reliably capture some of these desiderata, even those that are required by Bayesian theory. To address this, we introduce a new approach for capturing model uncertainty for NNs, which we call *Neural Optimization-based Model Uncertainty (NOMU)*. The main idea of NOMU is to design a network architecture consisting of two connected sub-NNs, one for model prediction and one for model uncertainty, and to train it using a carefully-designed loss function. Importantly, our design enforces that NOMU satisfies our five desiderata. Due to its modular architecture, NOMU can provide model uncertainty for any given (previously trained) NN if given access to its training data. We evaluate NOMU in various regressions tasks and noiseless Bayesian optimization (BO) with costly evaluations. In regression, NOMU performs at least as well as state-of-the-art methods. In BO, NOMU even outperforms all considered benchmarks.

1. Introduction

Neural networks (NNs) are becoming increasingly important in machine learning applications (LeCun et al., 2015). In many domains, it is essential to be able to quantify the *model uncertainty (epistemic uncertainty)* of NNs (Neal, 2012; Ghahramani, 2015). Good estimates of model uncertainty are indispensable in Bayesian optimization (BO) and active learning, where exploration is steered by (functions of) these uncertainty estimates. In recent years, BO has been successfully applied in practice to a wide range of problems, including robotics (Martinez-Cantin et al., 2009), sensor networks (Srinivas et al., 2012), and drug development (Gómez-Bombarelli et al., 2018). Better model uncertainty estimates for BO directly translate to improvements in these applications.

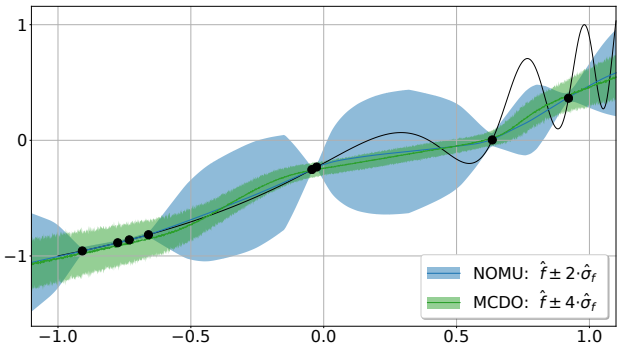


Figure 1. Visualization of estimated model uncertainty $\hat{\sigma}_f$. The unknown true function is depicted as a black solid line with training points as black dots. NOMU's model prediction $\hat{f}$ is shown as a solid blue line and its uncertainty bounds are shown as a blue shaded area. As a benchmark, MC Dropout is shown in green.

However, estimating model uncertainty well for NNs is still an open research problem. For settings with scarce training data and negligible data noise, where model uncertainty is the main source of uncertainty, we uncover deficiencies of widely used state-of-the-art methods for estimating model uncertainty for NNs. Prior work often only measures the performance in data noise dominant settings, and thus does not adequately isolate the pure model uncertainty, thereby overlooking the algorithms' deficiencies. However, in tasks such as BO with costly evaluations, where accurate estimates of model uncertainty are of utmost importance, these deficiencies can drastically decrease performance.

In this paper, we study the problem of estimating model uncertainty for NNs to obtain *uncertainty bounds (UBs)* that estimate Bayesian credible bounds in a setting with negligible data noise and scarce training data. For this, we propose a novel algorithm (NOMU) that is specialized to such a setting. Figure 1 shows UBs for NOMU and the benchmark method MC Dropout.

[*]Equal contribution [1]ETH Zurich [2]ETH AI Center [3]University of Zurich. Correspondence to: Jakob Weissteiner <weissteiner@ifi.uzh.ch>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

1.1. Prior Work on Model Uncertainty for NNs

Over the last decade, researchers have developed various methods to quantify model uncertainty for NNs. One strand of research considers Bayesian Neural Networks (BNNs), where distributions are placed over the NN's parameters (Graves, 2011; Blundell et al., 2015; Hernández-Lobato & Adams, 2015). However, variational methods approximating BNNs are usually computationally prohibitive and require careful hyperparameter tuning. Thus, BNNs are rarely used in practice (Wenzel et al., 2020a).

In practice, *ensemble methods* are more established:

- Gal & Ghahramani (2016) proposed *Monte Carlo dropout (MCDO)* to estimate model uncertainty via stochastic forward passes. Interestingly, they could show that training a NN with dropout can also be interpreted as variational inference approximating a BNN.
- Lakshminarayanan et al. (2017) experimentally evaluated ensembles of NNs and showed that they perform as good or better than BNNs. They proposed using *deep ensembles (DE)*, which use NNs with two outputs for model prediction and data noise, and they estimate model uncertainty via the empirical standard deviation of the ensemble. DE is the most established state-of-the art ensemble method and has been shown to consistently outperform other ensemble methods (Ovadia et al., 2019; Fort et al., 2019; Gustafsson et al., 2020; Ashukha et al., 2020).
- Recently, Wenzel et al. (2020b) proposed *hyper deep ensembles (HDE)*, an extension of DE where additional diversity is created via different hyperparameters, and they showed that HDE outperforms DE.

Despite the popularity of MCDO, DE and HDE, our empirical results suggest that none of them reliably capture all essential features of model uncertainty: MCDO yields tubular bounds that do not narrow at observed data points (which can already be observed in Figure 1); DE and HDE can produce UBs that are sometimes unreasonably narrow in regions far from observed data or unreasonably wide at training points (as we will show in Section 4.1).

1.2. Overview of our Contribution

We present a new approach for estimating model uncertainty for NNs, which we call *neural optimization-based model uncertainty (NOMU)*. In contrast to a fully Bayesian approach (e.g. BNNs), where approximating the posterior for a realistic prior is in general very challenging, we estimate posterior credible bounds by directly enforcing essential properties of model uncertainty. Specifically, we make the following contributions:

1. We first introduce five desiderata that we argue model UBs should satisfy (Section 3.1).
2. We then introduce NOMU, which consists of a network architecture (Section 3.2) and a carefully-designed loss function (Section 3.3), such that the estimated UBs fulfill these five desiderata. NOMU is easy to implement, scales well to large NNs, and can be represented as a *single* NN without the need for further ensemble distillation (in contrast to MCDO, DE and HDE). Because of its modular architecture, it can easily be used to obtain UBs for already trained NNs.
3. We experimentally evaluate NOMU in various regression settings: in scarce and noiseless settings to isolate model uncertainty (Sections 4.1.1 and 4.1.2) and on real-word data-sets (Sections 4.1.3 and 4.1.4). We show that NOMU performs well across all these settings while state-of-the-art methods (MCDO, DE, and HDE) exhibit several deficiencies.[1]
4. Finally, we evaluate the performance of NOMU in high-dimensional Bayesian optimization (BO) and show that NOMU performs as well or better than all considered benchmarks (Section 4.2).

Our source code is available on GitHub: https://github.com/marketdesignresearch/NOMU.

1.3. Further Related Work

Nix & Weigend (1994) were among the first to introduce NNs with two outputs: one for model prediction and one for *data noise (aleatoric uncertainty)*, using the Gaussian negative log-likelihood as loss function. However, such a data noise output cannot be used as an estimator for model uncertainty (epistemic uncertainty); see Appendix G for details. To additionally capture model uncertainty, Kendall & Gal (2017) combined the idea of Nix & Weigend (1994) with MCDO.

Similarly, NNs with two outputs for lower and upper UBs, trained on specifically-designed loss functions, were previously considered by Khosravi et al. (2010) and Pearce et al. (2018). However, the method by Khosravi et al. (2010) again only accounts for data noise and does not consider model uncertainty. The method by Pearce et al. (2018) also does not take model uncertainty into account in the design of their loss function and only incorporates it via ensembles (as in DE).

Besides the state-of-the art ensemble methods HDE and DE, there exist several other papers on ensemble methods that,

[1] We also conducted experiments with (Blundell et al., 2015). However, we found that this method did not perform as well as the other considered benchmarks. Moreover, it was shown in (Gal & Ghahramani, 2016; Lakshminarayanan et al., 2017) that *deep ensembles* and *MC dropout* outperform the methods by (Hernández-Lobato & Adams, 2015) and (Graves, 2011), respectively. Therefore, we do not include (Graves, 2011; Blundell et al., 2015; Hernández-Lobato & Adams, 2015) in our experiments.

for example, promote their diversity on the function space (Wang et al., 2019) or reduce their computational cost (Wen et al., 2020; Havasi et al., 2021).

For classification, Malinin & Gales (2018) introduced prior networks, which explicitly model in-sample and out-of-distribution uncertainty, where the latter is realized by minimizing the reverse KL-distance to a selected flat point-wise defined prior. In a recent working paper (and concurrent to our work), Malinin et al. (2020a) report on progress extending their idea to regression. While the idea of introducing a separate loss for learning model uncertainty is related to NOMU, there are several important differences (loss, architecture, behavior of the model prediction, theoretical motivation; see Appendix E for details). Furthermore, their experiments suggest that DE still performs as good or better than their proposed method.

In contrast to BNNs, which perform approximate inference over the entire set of weights, Neural Linear Models (NLMs) perform *exact* inference on only the last layer. NLMs have been extensively benchmarked in (Ober & Rasmussen, 2019) against MCDO and the method from (Blundell et al., 2015). Their results suggest that MCDO and (Blundell et al., 2015) perform competitive, even to carefully-tuned NLMs.

Neural processes, introduced by Garnelo et al. (2018a;b), have been used to express model uncertainty for image completion tasks, where one has access to 1000s of different images interpreted as functions f_i instead of input points x_i. See Appendix F for a detailed comparison of their setting to the setting we consider in this paper.

2. Preliminaries

In this section, we briefly review the classical Bayesian uncertainty framework for regression.

Let $X \subset \mathbb{R}^d, Y \subset \mathbb{R}$ and let $f\colon X \to Y$ denote the unknown ground truth function. Let $D^{\text{train}} := \{(x_i^{\text{train}}, y_i^{\text{train}}) \in X \times Y, i \in \{1, \dots, n^{\text{train}}\}\}$, with $n^{\text{train}} \in \mathbb{N}$ be i.i.d samples from the data generating process $y = f(x) + \varepsilon$, where $\varepsilon | x \sim \mathcal{N}(0, \sigma_{\text{n}}^2(x))$. We use σ_{n} to refer to the *data noise (aleatoric uncertainty)*. We refer to $(x_i^{\text{train}}, y_i^{\text{train}})$ as a *training point* and to x_i^{train} as an *input training point*.

In the remainder of this paper, we follow the classic Bayesian uncertainty framework by modelling the unknown ground truth function f as a random variable. Hence, with a slight abuse of notation, we use the symbol f to denote *both* the unknown ground truth function as well as the corresponding random variable. In Appendix I, we provide a mathematically rigorous formulation of the considered Bayesian uncertainty framework.

Given a prior distribution for f and known data noise σ_{n}, the posterior of f and y are well defined. The *model uncertainty (epistemic uncertainty)* $\sigma_f(x)$ is the posterior standard deviation of $f(x)$, i.e.,

$$\sigma_f(x) := \sqrt{\mathbb{V}[f(x)|D^{\text{train}}, x]}, \quad x \in X. \tag{1}$$

Assuming independence between f and ε, the variance of the predictive distribution of y can be decomposed as $\mathbb{V}[y|D^{\text{train}}, x] = \sigma_f^2(x) + \sigma_{\text{n}}^2(x)$. We present our algorithm for estimating model uncertainty $\hat{\sigma}_f$ for the case of zero or negligible data noise, i.e., $\sigma_{\text{n}} \approx 0$ (see Appendix C for an extension to $\sigma_{\text{n}} \gg 0$). For a given model prediction $\hat{f}$, the induced uncertainty bounds (UBs) are then given by $\left(\underline{UB}_c(x), \overline{UB}_c(x)\right) := \left(\hat{f}(x) \mp c\, \hat{\sigma}_f(x)\right)$, for $x \in X$ and a calibration parameter $c \geq 0$.[2]

3. The NOMU Algorithm

We now present NOMU. We design NOMU to yield a model prediction $\hat{f}$ and a model uncertainty prediction $\hat{\sigma}_f$, such that the resulting UBs $(\underline{UB}_c, \overline{UB}_c)$ fulfill five desiderata.

3.1. Desiderata

D1 (Non-Negativity) *The upper/lower UB between two training points lies above/below the model prediction $\hat{f}$, i.e., $\underline{UB}_c(x) \leq \hat{f}(x) \leq \overline{UB}_c(x)$ for all $x \in X$ and for $c \geq 0$. Thus, $\hat{\sigma}_f \geq 0$.*

By definition, for any given prior, the exact posterior model uncertainty σ_f is positive, and therefore Desideratum D1 should also hold for any estimate $\hat{\sigma}_f$.

D2 (In-Sample) *In the noiseless case ($\sigma_n \equiv 0$), there is zero model uncertainty at each input training point x^{train}, i.e., $\hat{\sigma}_f(x^{train}) = 0$. Thus, $\overline{UB}_c(x^{train}) = \underline{UB}_c(x^{train}) = \hat{f}(x^{train})$ for $c \geq 0$.*

In Appendix D.2, we prove that, for any prior that does not contradict the training data, the exact σ_f satisfies D2. Thus, D2 should also hold for any estimate $\hat{\sigma}_f$, and we argue that even in the case of non-zero small data noise, model uncertainty should be small at input training data points.

D3 (Out-of-Sample) *The larger the distance of a point $x \in X$ to the input training points in D^{train}, the wider the UBs at x, i.e., model uncertainty $\hat{\sigma}_f$ increases out-of-sample.*[3]

For D3 it is often not obvious which metric to choose to measure distances. Some aspects of this metric can be speci-

[2] Note that our UBs estimate *credible bounds (CBs)* $\underline{CB}$ and $\overline{CB}$, which, for $\alpha \in [0, 1]$, fulfill that $\mathbb{P}[f(x) \in [\underline{CB}, \overline{CB}] | D^{\text{train}}, x] = \alpha$. For $\sigma_{\text{n}} \equiv 0$, CBs are equal to *predictive bounds* $\underline{PB}, \overline{PB}$ with $\mathbb{P}[y \in [\underline{PB}, \overline{PB}] | D^{\text{train}}, x] = \alpha$. See Appendix A for an explanation.

[3] Importantly, D3 also promotes model uncertainty in gaps *between* input training points.

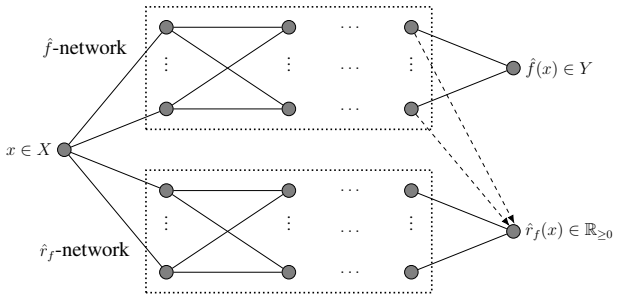


Figure 2. NOMU's network architecture

fied via the architecture (e.g., shift invariance for CNNs). In many applications, it is best to *learn* further aspects of this metric from training data, motivating our next desideratum.

D4 (Metric Learning) *Changes in those features of x that have high predictive power on the training set have a large effect on the distance metric used in D3.*[4]

D4 is not required for any application. However, specifically in *deep* learning applications, where it is a priori not clear which features are important, D4 is particularly desirable.

D5 (Vanishing) *As the number of training points n^{train} tends to infinity, model uncertainty vanishes for each x in the support of the input data distribution, i.e.,* $\lim_{n^{train}\to\infty} \hat{\sigma}_f(x) = 0$ *for a fixed $c \geq 0$. Thus, for a fixed c,* $\lim_{n^{train}\to\infty} |\overline{UB}_c(x) - \underline{UB}_c(x)| = 0$.

In Appendix D, we discuss all desiderata in more detail (see Appendix D.4 for a visualization of D4).

3.2. The Network Architecture

For NOMU, we construct a network $\mathcal{NN}_\theta$ with two outputs: *the model prediction* $\hat{f}$ (e.g. mean prediction) and a *raw* model uncertainty prediction $\hat{r}_f$. Formally: $\mathcal{NN}_\theta \colon X \to Y \times \mathbb{R}_{\geq 0}$, with $x \mapsto \mathcal{NN}_\theta(x) := (\hat{f}(x), \hat{r}_f(x))$. NOMU's architecture consists of two almost separate sub-networks: the $\hat{f}$-network and the $\hat{r}_f$-network (see Figure 2). For each sub-network, any network architecture can be used (e.g., feed-forward NNs, CNNs). This makes NOMU highly modular and we can plug in any previously trained NN for $\hat{f}$, or we can train $\hat{f}$ simultaneously with the $\hat{r}_f$-network. The $\hat{r}_f$-network learns the raw model uncertainty and is connected with the $\hat{f}$-network through the last hidden layer (dashed lines in Figure 2). This connection enables $\hat{r}_f$ to re-use features that are important for the model prediction $\hat{f}$, implementing Desideratum D4 (Metric Learning).[5]

Remark 3.1 *NOMU's network architecture can be modified to realize D4 (Metric Learning) in many different ways. For example, if low-level features were important for predicting the model uncertainty, one could additionally add connections from earlier hidden layers of the $\hat{f}$-network to layers of the $\hat{r}_f$-network. Furthermore, one can strengthen D4 (Metric Learning) by increasing the regularization of the $\hat{r}_f$-network (see Appendix D.4).*

After training $\mathcal{NN}_\theta$, we apply the readout map $\varphi(z) = \ell_{\text{max}}(1 - \exp(-\frac{\max(0,z)+\ell_{\text{min}}}{\ell_{\text{max}}}))$, $\ell_{\text{min}} \geq 0, \ell_{\text{max}} > 0$ to the *raw* model uncertainty output $\hat{r}_f$ to obtain NOMU's *model uncertainty prediction*

$$\hat{\sigma}_f(x) := \varphi(\hat{r}_f(x)), \forall x \in X. \tag{2}$$

The readout map φ monotonically interpolates between a minimal ℓ_{min} and a maximal ℓ_{max} model uncertainty. Here, ℓ_{min} is used for numerical stability, and ℓ_{max} defines the maximal model uncertainty far away from input training points (similarly to the prior variance for GPs). With NOMU's model prediction $\hat{f}$, its model uncertainty prediction $\hat{\sigma}_f$ defined in (2), and given a calibration parameter[6] $c \in \mathbb{R}_{\geq 0}$, we can now define for each $x \in X$ NOMU's UBs as

$$\left(\underline{UB}_c(x), \overline{UB}_c(x)\right) := \left(\hat{f}(x) \mp c\, \hat{\sigma}_f(x)\right). \tag{3}$$

It is straightforward to construct a *single* NN that directly outputs the upper/lower UB, by extending the architecture shown in Figure 2: we monotonically transform and scale the output $\hat{r}_f(x)$ and then add/subtract this to/from the other output $\hat{f}(x)$. It is also straightforward to compute NOMU's UBs for any *given, previously trained NN*, by attaching the $\hat{r}_f$-network to the trained NN, and only training the $\hat{r}_f$-network on the same training points as the original NN.

Remark 3.2 *The readout map φ can be modified depending on the subsequent use of the estimated UBs. For example, for BO over discrete domains (e.g. $X = \{0,1\}^d$) (Baptista & Poloczek, 2018), we propose the linearized readout map $\varphi(z) = \ell_{min} + \max(0, z - \ell_{min}) - \max(0, z - \ell_{max})$. With this φ and ReLU activations, one can encode NOMU's UBs as a mixed integer program (MIP) (Weissteiner & Seuken, 2020). This enables optimizing the upper UB as acquisition*

[4] Consider the task of learning facial expressions from images. For this, eyes and mouth are important features, while background color is not. A CNN automatically learns which features are important for model prediction. The same features are also important for model uncertainty: Consider an image with pixel values similar to those of an image of the training data, but where mouth and eyes are very different. We should be substantially more uncertain about the model prediction for such an image than for one which is almost identical to a training image except that it has a different background color, even if this change of background color results in a huge Euclidean distance of the pixel vectors. D4 requires that a more useful metric is learned instead.

[5] To prevent that $\hat{r}_f$ impacts $\hat{f}$, the dashed lines should only make forward passes when trained.

[6] Like all other methods, NOMU outputs *relative* UBs that should be calibrated, e.g., via a parameter $c \geq 0$. See also Kuleshov et al. (2018) for a non-linear calibration method.

function without the need for further approximation via ensemble distillations(Malinin et al., 2020b).

3.3. The Loss Function

We now introduce the loss function L^π we use for training NOMU's architecture. Let X be such that $0 < \lambda_d(X) < \infty$, where λ_d denotes the d-dimensional Lebesgue measure. We train $\mathcal{NN}_\theta$ with loss L^π and L2-regularization parameter $\lambda > 0$, i.e., minimizing $L^\pi(\mathcal{NN}_\theta) + \lambda \|\theta\|_2^2$ via a gradient descent-based algorithm.

Definition 3.3 (NOMU Loss) *Let $\pi := (\pi_{sqr}, \pi_{exp}, c_{exp}) \in \mathbb{R}^3_{\geq 0}$ denote a tuple of hyperparameters. Given a training set D^{train}, the loss function L^π is defined as*

$$L^\pi(\mathcal{NN}_\theta) := \underbrace{\sum_{i=1}^{n^{train}} (\hat{f}(x_i^{train}) - y_i^{train})^2}_{(a)} + \pi_{sqr} \cdot \underbrace{\sum_{i=1}^{n^{train}} (\hat{r}_f(x_i^{train}))^2}_{(b)} + \pi_{exp} \cdot \underbrace{\frac{1}{\lambda_d(X)} \int_X e^{-c_{exp} \cdot \hat{r}_f(x)} \, dx}_{(c)} . \quad (4)$$

In the following, we explain how the three terms of L^π promote the desiderata we introduced in Section 3.1. Note that the behaviour of $\hat{r}_f$ directly translates to that of $\hat{\sigma}_f$.

- Term (a) solves the regression task, i.e., learning a smooth function $\hat{f}$ given D^{train}. If $\hat{f}$ is given as a pre-trained NN, then this term can be omitted.
- Term (b) implements D2 (In-Sample) and D5 (Vanishing) (i.e., $\hat{r}_f(x_i^{\text{train}}) \approx 0$). The hyperparameter π_{sqr} controls the amount of uncertainty at the training points.[7] The larger π_{sqr}, the narrower the UBs at training points.
- Term (c) has two purposes. First, it implements D1 (Non-Negativity) (i.e., $\hat{r}_f \geq 0$). Second, it pushes $\hat{r}_f$ towards infinity across the whole input space X. However, due to the counteracting force of (b) as well as regularization, $\hat{r}_f$ increases continuously as you move away from the training data. The interplay of (b), (c), and regularization thus promotes D3 (Out-of-Sample). The hyperparameters π_{exp} and c_{exp} control the size and shape of the UBs. Concretely, the larger π_{exp}, the wider the UBs; the larger c_{exp}, the narrower the UBs at points x with large $\hat{\sigma}_f(x)$ and the wider the UBs at points x with small $\hat{\sigma}_f(x)$.

In Appendix H.1, we provide detailed visualizations on how the loss hyperparameters π_{sqr}, π_{exp}, and c_{exp} shape NOMU's model uncertainty estimate. In the implementation of L^π, we approximate (c) via MC-integration using additional, *artificial input points* $D^{\text{art}} := \{x_i\}_{i=1}^l \overset{i.i.d}{\sim} \text{Unif}(X)$ by $\frac{1}{l} \cdot \sum_{x \in D^{\text{art}}} e^{-c_{\text{exp}} \cdot \hat{\sigma}_f(x)}$.

[7] In the noiseless case, in theory $\frac{\pi_{\text{sqr}}}{\lambda} = \infty$; we set $\frac{\pi_{\text{sqr}}}{\lambda} = 10^7$. For small non-zero data noise, setting $\frac{\pi_{\text{sqr}}}{\lambda} \ll \infty$ captures *data noise induced model uncertainty* $\sigma_f(x^{\text{train}}) > 0$ (Appendix D.2).

Remark 3.4 *In (c), instead of the Lebesgue-measure, one can also use a different measure ν, i.e., $\frac{1}{\nu(X)} \int_X e^{-c_{exp} \cdot \hat{r}_f(x)} \, d\nu(x)$. This can be relevant in high dimensions, where meaningful data points often lie close to a lower-dimensional manifold (Cayton, 2005); ν can then be chosen to concentrate on that region. In practice, this can be implemented by sampling from a large unlabeled data set D^{art} representing this region or learning the measure ν using GANs (Goodfellow et al., 2014).*

Theory In Appendix D, we prove that NOMU fulfills D1, D2 and D5 (Propositions D.1.a, D.2.c and D.5.a), and discuss how NOMU fulfills D3 and D4. In Appendix A.1, we show that, under certain assumptions, NOMU's UBs can be interpreted as *pointwise worst-case* UBs $\overline{UB}_{\text{pw}}(x) := \sup_{f \in \mathcal{H}_{D^{\text{train}}}} f(x)$ within a hypothesis class $\mathcal{H}_{D^{\text{train}}}$ of data-explaining functions. In Appendix A.2, we explain how $\underline{UB}_{\text{pw}}(x)$ and $\overline{UB}_{\text{pw}}(x)$ estimate posterior CBs of BNNs (with a Gaussian prior on the weights), without performing challenging variational inference. However, while exact posterior CBs of BNNs lose D4 as their width goes to infinity, NOMU's UBs are capable of retaining D4 in this limit.

4. Experimental Evaluation

In this section, we experimentally evaluate NOMU's model uncertainty estimate in multiple synthetic and real-world regression settings (Section 4.1) as well as in high-dimensional Bayesian optimization (Section 4.2).

Benchmarks We compare NOMU against four benchmarks, each of which gives a model prediction $\hat{f}$ and a model uncertainty prediction $\hat{\sigma}_f$ (see Appendix B.1 for formulas). We calculate model-specific UBs at $x \in X$ as $(\hat{f}(x) \mp c \, \hat{\sigma}_f(x))$ with calibration parameter $c \in \mathbb{R}_{\geq 0}$ and use them to evaluate all methods. We consider three algorithms that are specialized to model uncertainty for NNs: (i) MC dropout (MCDO) (ii) deep ensembles (DE) and (iii) hyper deep ensembles (HDE) and a non-NN-based benchmark: (iv) Gaussian process (GP) with RBF kernel.

4.1. Regression

To develop intuition, we first study the *model* UBs of all methods on synthetic test functions with 1D–2D scarce input training points without data noise (Section 4.1.1). We then propose a novel generative test-bed and evaluate NOMU within this setting (Section 4.1.2). Next, we analyze NOMU on a real-world time series (Section 4.1.3). Finally, we evaluate NOMU on the real-world UCI data sets (Section 4.1.4).

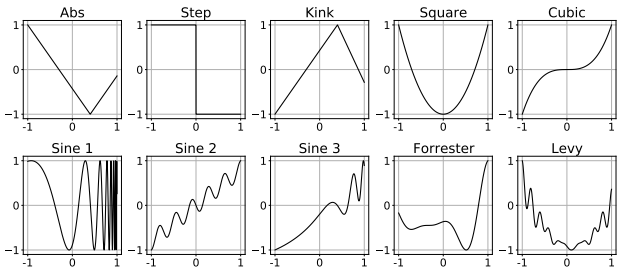


Figure 3. 1D synthetic test functions

4.1.1. Toy Regression

Setting We consider ten different 1D functions whose graphs are shown in Figure 3. Those include the popular Levy and Forrester function with multiple local optima.[8] All functions are transformed to $X := [-1, 1] =: f(X)$. For each function, we conduct 500 runs. In each run, we randomly sample eight noiseless input training points from X, such that the only source of uncertainty is model uncertainty. For each run, we also generate 100 *test points* in the same fashion to assess the quality of the UBs.

Metrics We report the *average negative log (Gaussian) likelihood (NLL)*, minimized over the calibration parameter c, which we denote as NLL_{min}. Following prior work (Khosravi et al., 2010; Kuleshov et al., 2018; Pearce et al., 2018), we further measure the quality of UBs by contrasting their *mean width* (MW) with their *coverage probability* (CP). Ideally, MW should be as small as possible, while CP should be close to 1. Since CP is counter-acting MW, we consider ROC-like curves, plotting MW against CP for a range of calibration parameters c, and report the *area under the curve (AUC)* (see Appendix B.2.1 for details).

Algorithm Setup For each of the two NOMU subnetworks, we use a feed-forward NN with three fully-connected hidden layers à 2^{10} nodes, ReLUs, and hyperparameters $\pi_{\text{exp}} = 0.01, \pi_{\text{sqr}} = 0.1, c_{\text{exp}} = 30$. In practice, the values for $\pi_{\text{exp}}, \pi_{\text{sqr}}$, and c_{exp} can be tuned on a validation set. However, for all synthetic experiments (Section 4.1.1 and Section 4.1.2), we use the same values, which lead to good results across all functions. Moreover, we set $\lambda = 10^{-8}$ accounting for zero data-noise, $\ell_{\text{min}} = 0.001$ and $\ell_{\text{max}} = 2$. In Appendix H.2, we provide an extensive sensitivity analysis for the hyperparameters $\pi_{\text{exp}}, \pi_{\text{sqr}}, c_{\text{exp}}, \ell_{\text{min}}$ and ℓ_{max}. This analysis demonstrates NOMU's robustness within a certain range of hyperparameter values.

Furthermore, to enable a fair comparison of all methods, we use generic representatives and do not optimize any of them for any test function. We also choose the architectures of all NN-based methods, such that the overall number of parameters is comparable. We provide all methods with the same prior information (specifically, this entails knowledge of zero data noise), and set the corresponding parameters accordingly. Finally, we set all remaining hyperparameters of the benchmarks to the values proposed in the literature. Details on all configurations are provided in Appendix B.2.2.

[8] sfu.ca/ ssurjano/optimization.html

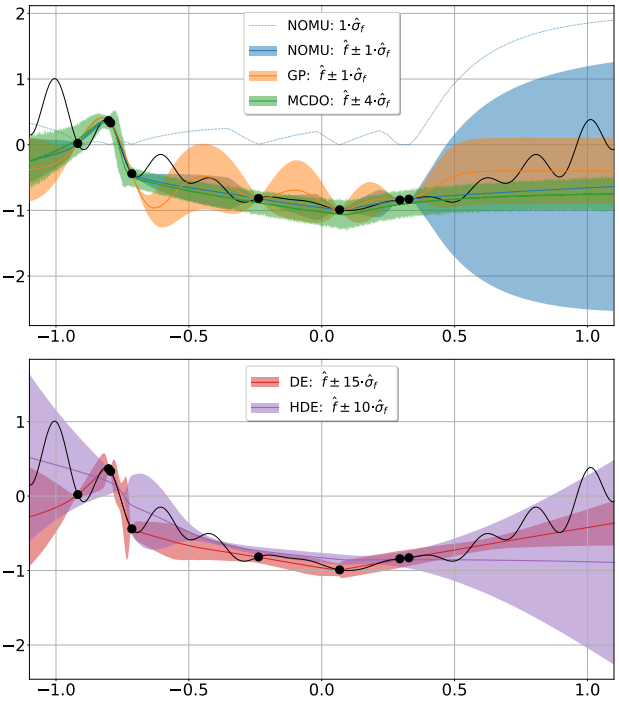


Figure 4. UBs resulting from NOMU, GP, MCDO, and DE, HDE for the Levy function (solid black line). For NOMU, we also show $\hat{\sigma}_f$ as a dotted blue line. Training points are shown as black dots.

Results Figure 4 exemplifies our findings, showing typical UBs for the Levy function as obtained in one run. In Appendix B.2.4 we provide further visualisations. We find that MCDO consistently yields tube-like UBs; in particular, its UBs do not narrow at training points, i.e., failing in D2 (In-Sample) even though MCDO's Bayesian interpretation requires D2 to hold (see Appendix D.2). Moreover, it only fulfills D3 (Out-of-Sample) to a limited degree. We frequently observe that DE leads to UBs of somewhat arbitrary shapes. This can be seen most prominently in Figure 4 around $x \approx -0.75$ and at the edges of its input range, where DE's UBs are very different in width with no clear justification. Thus, also DE is limited in D3 (Out-of-Sample). In addition, we sometimes see that also DE's UBs do not narrow sufficiently at training points, i.e., not fulfilling D2 (In-Sample). HDE's UBs are even more random, i.e., predicting large model uncertainty at training points and some-

Table 1. Ranks (1=best to 5=worst) for AUC and NLL_{min}.

	NOMU		GP		MCDO		DE		HDE	
FUNCTION	AUC	NLL_{MIN}	AUC	NLL_{MIN}	AUC	NLL_{MIN}	AUC	NLL_{MIN}	AUC	NLL_{MIN}
ABS	1	1	3	1	3	4	1	1	5	5
STEP	2	2	4	3	2	3	1	1	5	5
KINK	1	1	3	1	4	4	1	1	5	5
SQUARE	2	2	2	1	4	4	2	3	5	5
CUBIC	2	2	1	1	3	4	3	3	5	5
SINE 1	2	1	2	1	1	1	2	1	5	5
SINE 2	2	2	3	1	1	2	3	3	5	5
SINE 3	1	1	4	1	3	4	1	1	5	5
FORRESTER	1	2	1	1	3	4	3	3	5	5
LEVY	1	1	4	3	1	4	3	1	5	5

times zero model uncertainty in gaps between them (e.g. $x \approx -0.75$). In contrast, NOMU displays the behaviour it is designed to show. Its UBs nicely tighten at training points and expand in-between (D1–D3, for D4 (Metric Learning) see Appendix D.4). Like NOMU, the GP fulfills D2 (In-Sample) and D3 (Out-of-Sample) well, but cannot account for D4 (Metric Learning) (a fixed kernel does not depend on the model prediction). Table 1 provides the ranks achieved by each algorithm (see Appendix B.2.3 for corresponding metrics). We calculate the ranks based on the medians and a 95% bootstrap CI of AUC and NLL_{min}. An algorithm loses one rank to each other algorithm that significantly dominates it. Winners are marked in grey. We observe that NOMU is the only algorithm that never comes in third place or worse. Thus, while some algorithms do particularly well in capturing uncertainties of functions with certain characteristics (e.g. RBF-GPs for polynomials), NOMU is the only algorithm that consistently performs well. HDE's bad performance can be explained by its randomness and the fact that it sometimes predicts zero model uncertainty out-of-sample. For 2D, we provide results and visualizations in Appendix B.2.3 and B.2.4 highlighting similar characteristics of all the algorithms as in 1D.

4.1.2. GENERATIVE TEST-BED

Setting Instead of only relying on a *limited* number of data-sets, we also evaluate all algorithms on a generative test-bed, which provides an unlimited number of test-functions. The importance of using a test-bed (to avoid over-fitting on specific data-sets) has also been highlighted in a recent work of Osband et al. (2021). For our test-bed, we generate 200 different data-sets by randomly sampling 200 different test-functions from a BNN with i.i.d centered Gaussian weights and three fully-connected hidden layers with nodes $[2^{10}, 2^{11}, 2^{10}]$ and ReLU activations. From each of these test-functions we uniformly at random sample $n^{\text{train}} = 8 \cdot d$ input training points and $100 \cdot d$ test data points, where d refers to the input dimension. We train all algorithms on these training sets and determine the NLL on the corresponding test sets averaged over the 200 test-functions. This metric converges to the Kullback-Leibler divergence to the exact BNN-posterior (see Theorem B.7). We calibrate by choosing per dimension an optimal value of c in terms of average NLL, which does not depend on the test-function.

Table 2. Average NLL (without const. $\ln(2\pi)/2$) and a 95% CI over 200 BNN samples. Winners are marked in grey.

FUNCTION	NOMU	GP	MCDO	DE	HDE
BNN1D	-1.65±0.10	-1.08±0.22	-0.34±0.23	-0.38±0.36	8.47±1.00
BNN2D	-1.16±0.05	-0.52±0.11	-0.33±0.13	-0.77±0.07	9.11±0.39
BNN5D	-0.37±0.02	-0.33±0.02	-0.05±0.04	-0.13±0.03	8.41±1.00

Results In Table 2, we provide the results for input dimensions 1, 2 and 5. We see that NOMU outperforms all other algorithms including MCDO, which is a variational inference method to approximate this posterior (Gal & Ghahramani, 2016). See Appendix B.2.5 for a detailed explanation of the experiment setting and further results in modified settings including a discussion of higher dimensional settings.

4.1.3. SOLAR IRRADIANCE TIME SERIES

Setting Although the current version of NOMU is specifically designed for scarce settings without data noise, we are also interested to see how well it performs in settings where these assumptions are not satisfied. To this end, we now study a setting with many training points and small non-zero data noise. This allows us to analyze how well NOMU captures D5 (Vanishing). We consider the popular task of interpolating the solar irradiance data (Steinhilber et al., 2009) also studied in Gal & Ghahramani (2016). We scale the data to $X = [-1, 1]$ and split it into 194 training and 197 test points. As in Gal & Ghahramani (2016), we choose five intervals to contain only test points. Since the true function is likely of high frequency, we set $\lambda = 10^{-19}$ for NOMU and the benchmarks' regularization accordingly. To account for small non-zero data noise, we set $\pi_{\text{exp}} = 0.05$, $\ell_{\text{min}} = 0.01$ and use otherwise the same hyperparameters as in Section 4.1.1.

Results Figure 5 visualizes NOMU's UBs. We see that NOMU manages to fit the training data well while capturing model uncertainty between input training points. In particu-

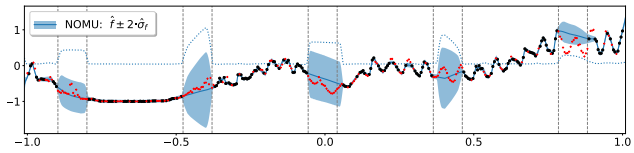


Figure 5. NOMU's model prediction (solid), model uncertainty (dotted) and UBs (shaded area) on the solar irradiance data. Training and test points are shown as black dots and red crosses.

Table 3. Average NLL and a 95% normal-CI over 20 runs for UCI data sets. Winners are marked in grey.

DATASET	NOMU	DE	MCDO	MCDO2	LL	NLM-HPO	NLM
BOSTON	2.68 ±0.11	2.41 ±0.49	2.46 ±0.11	2.40 ±0.07	2.57 ±0.09	2.58 ±0.17	3.63 ±0.39
CONCRETE	3.05 ±0.06	3.06 ±0.35	3.04 ±0.03	2.97 ±0.03	3.05 ±0.07	3.11 ±0.09	3.12 ±0.09
ENERGY	0.77 ±0.06	1.38 ±0.43	1.99 ±0.03	1.72 ±0.01	0.82 ±0.05	0.69 ±0.05	0.69 ±0.05
KIN8NM	-1.08 ±0.01	-1.20 ±0.03	-0.95 ±0.01	-0.97 ±0.00	-1.23 ±0.01	-1.12 ±0.01	-1.13 ±0.01
NAVAL	-5.63 ±0.39	-5.63 ±0.09	-3.80 ±0.01	-3.91 ±0.01	-6.40 ±0.11	-7.36 ±0.15	-7.35 ±0.01
CCPP	2.79 ±0.01	2.79 ±0.07	2.80 ±0.01	2.79 ±0.01	2.83 ±0.01	2.79 ±0.01	2.79 ±0.01
PROTEIN	2.79 ±0.01	2.83 ±0.03	2.89 ±0.00	2.87 ±0.00	2.89 ±0.00	2.78 ±0.01	2.81 ±0.00
WINE	1.08 ±0.04	0.94 ±0.23	0.93 ±0.01	0.92 ±0.01	0.97 ±0.03	0.96 ±0.01	1.48 ±0.09
YACHT	1.38 ±0.28	1.18 ±0.41	1.55 ±0.05	1.38 ±0.01	1.01 ±0.09	1.17 ±0.13	1.13 ±0.09

lar, large gaps *between* input training points are successfully modeled as regions of high model uncertainty (D3 (Out-of-Sample)). Moreover, in regions highly populated by input training points, NOMU's model uncertainty vanishes as required by D5 (Vanishing). Plots for the other algorithms are provided in Appendix B.2.6, where we observe similar patterns as in Section 4.1.1.

4.1.4. UCI DATA SETS

Setting Recall that NOMU is specifically tailored to noiseless settings with scarce input training data. However, even the current version of NOMU, which does not explicitly model data noise, already performs on par with existing benchmarks on real-world regression tasks *with* data noise. To demonstrate this, we test its performance on (a) the UCI data sets proposed in Hernández-Lobato & Adams (2015), a common benchmark for uncertainty quantification in noisy, real-world regression, and (b) the UCI gap data set extension proposed in Foong et al. (2019). We consider exactly the same experiment setup as proposed in these works, with a $70/20/10$-train-validation-test split, equip NOMU with a shallow architecture of 50 hidden nodes, and train it for 400 epochs. Validation data are used to calibrate the constant c on NLL. See Appendix B.2.7 for details on NOMU's setup.

Results Table 3 reports NLLs on test data, averaged across 20 splits, compared to the current state-of-the-art. We report the NLL for MCDO (Gal & Ghahramani, 2016) and DE (Lakshminarayanan et al., 2017) from the original papers. Moreover, we reprint the best results (for comparable network sizes) of neural linear models (NLMs) with and without hyperparameter optimization (HPO) from (Ober & Rasmussen, 2019) (NLM-HPO, NLM), linearized laplace (LL) (Foong et al., 2019) and a recent strong MCDO baseline (MCDO2) from (Mukhoti et al., 2018). It is surprising that, even though the current design of NOMU does not yet explicitly incorporate data noise, it already performs comparably to state-of-the-art results. We obtain similar results for the UCI gap data. See Appendix B.2.7 for more details on this experiment.

4.2. Bayesian Optimization

In this section, we assess the performance of NOMU in high-dimensional noiseless *Bayesian optimization (BO)*.

Setting In BO, the goal is to maximize an unknown expensive-to-evaluate function, given a budget of function queries. We use a set of test functions with different characteristics from the same library as before, but now in 5 to 20 dimensions d, transformed to $X = [-1, 1]^d$, $f(X) = [-1, 1]$.[9] Additionally, we again use Gaussian BNNs to create a generative test-bed consisting of a large variety of test functions (see Section 4.1.2). For each test function, we randomly sample 8 initial points $(x_i, f(x_i))$ and let each algorithm choose 64 further function evaluations (one by one) using its upper UB as acquisition function. This corresponds to a setting where one can only afford 72 expensive function evaluations in total. We provide details regarding the selected hyperparameters for each algorithm in Appendix B.3.1. We measure the performance of each algorithm based on its *final regret* $|\max_{x\in X} f(x) - \max_{i\in\{1,\dots,72\}} f(x_i)|/|\max_{x\in X} f(x)|$.

For each algorithm, the UBs must be calibrated by choosing appropriate values of c. We do so in the following straightforward way: First, after observing the 8 random initial points, we determine those two values of c for which the resulting mean width (MW) of the UBs is 0.05 and 0.5, respectively (MW SCALING).[10] We perform one BO run for both resulting initial values of c. Additionally, if in a BO run, an algorithm were to choose a point $x_{i'}$ very close to an already observed point x_i, we dynamically increase c to make it select a different one instead (DYNAMIC C; see Appendix B.3.2 for details). A value of 0.05 in MW SCALING corresponds to small model uncertainty, such that exploration is mainly due to DYNAMIC C. Smaller values than 0.05 thus lead to similar outcomes. In contrast, a value of 0.5 corresponds to large model uncertainties, such that DYNAMIC C is rarely used. Only for the "plain GP (pGP)" we use neither MW SCALING nor DYNAMIC C, as pGP uses its default calibration (c is determined by the built-in hyperparameter optimization in every step). However, a comparison of GP and pGP suggests that MW SCALING and DYNAMIC C surpass the built-in calibration (see Table 4). As a baseline, we also report random search (RAND).

Results In Table 4, we present the BO results in 5D, 10D and 20D. We show the average final regret per dimension across the five functions. For each algorithm and dimension, we give the results corresponding to the MW scaling parameter (0.05 or 0.5) that minimizes the average final

[9] These functions are designed for minimization. We multiply them by -1 and equivalently maximize instead.

[10] We fix MW instead of c, since the scales of the algorithms vary by orders of magnitude.

Table 4. BO results: average final regrets per dimension and ranks for each individual function (1=best to 7=worst).

FUNCTION	NOMU	GP	MCDO	DE	HDE	PGP	RAND
LEVY5D	1	1	6	3	3	4	7
ROSENBROCK5D	1	1	1	1	2	5	7
G-FUNCTION5D	2	3	1	4	2	3	7
PERM5D	3	1	1	5	7	2	4
BNN5D	1	1	4	1	4	1	7
Average Regret 5D	2.87e−2	5.03e−2	4.70e−2	5.18e−2	7.13e−2	4.14e−2	1.93e−1
LEVY10D	1	3	5	6	1	1	6
ROSENBROCK10D	1	1	2	6	3	2	7
G-FUNCTION10D	2	5	1	3	2	5	7
PERM10D	2	1	2	6	2	2	1
BNN10D	1	2	1	1	3	1	7
Average Regret 10D	8.40e−2	1.17e−1	6.96e−2	1.15e−1	9.32e−2	9.46e−2	2.35e−1
LEVY20D	1	1	5	7	1	1	6
ROSENBROCK20D	2	2	2	6	1	4	6
G-FUNCTION20D	1	4	5	1	1	3	7
PERM20D	3	5	3	2	3	3	1
BNN20D	1	2	2	2	6	1	7
Average Regret 20D	1.12e−1	1.33e−1	1.39e−1	1.71e−1	1.37e−1	1.17e−1	2.80e−1

regret across that dimension (see Appendix B.3.3 for both MWs). In practice, one often only knows the dimensionality of a given BO task, which is why we use the average final regret per dimension as the criterion for setting the optimal MW. For each individual function, we also present the ranks based on the final regret and a 95% CI over 100 (5D) and 50 (10-20D) runs. We see that NOMU performs as well or better than all benchmarks in terms of average final regret. By inspecting the ranks achieved for each individual function, we further observe that NOMU is never ranked worse than 3rd. In contrast, the performance of the benchmarks heavily depends on the test function; and each benchmark is ranked 4th and worse multiple times. For Perm10D/20D we see that RAND performs best. However, due to a flat optimum of Perm, all algorithms achieve similar (very small) final regrets. Finally, we see that NOMU is always ranked first for the BNN test functions. Figure 6 shows the regret plot for BNN20D (see Appendix B.3.4 for all regret plots).

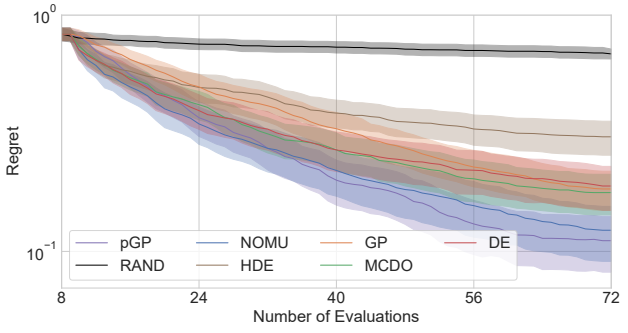


Figure 6. Regret plot for BNN20D. For each BO step, we show the regrets averaged over 50 runs (solid lines) with 95% CIs.

5. Conclusion

We have introduced NOMU, a new algorithm for estimating model uncertainty for NNs, specialized for scarce and noiseless settings. By using a specific architecture and carefully-designed loss function, we have ensured that NOMU satisfies five important desiderata regarding model uncertainty that any method should satisfy. However, when analyzing model uncertainty decoupled from data noise, we have experimentally uncovered that, perhaps surprisingly, established state-of-the-art methods fail to reliably capture some of the desiderata, even those that are required by Bayesian theory. In contrast, NOMU satisfies all desiderata, matches the performance of all benchmarks in regression tasks, and performs as well or better in noiseless BO tasks. We see great potential to further improve NOMU, for example by adapting the loss, or by modifying the connections between the two sub-NNs. We also envision several extensions of NOMU, including its application to classification, employing different architectures (CNNs, GNNs, RNNs or Transformers), and incorporating data noise.

Acknowledgements

We thank Marius Högger and Aurelio Dolfini for insightful discussions and their excellent research assistance in implementing the Bayesian optimization experiments and the UCI data set experiments, respectively. Furthermore, we thank the anonymous reviewers for helpful comments. This paper is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 805542).

References

Ashukha, A., Lyzhov, A., Molchanov, D., and Vetrov, D. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BJxI5gHKDr. 2

Baptista, R. and Poloczek, M. Bayesian optimization of combinatorial structures. In *International Conference on Machine Learning*, pp. 462–471. PMLR, 2018. URL http://proceedings.mlr.press/v80/baptista18a/baptista18a.pdf. 4

Bergstra, J. and Bengio, Y. Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(2), 2012. URL https://www.jmlr.org/papers/v13/bergstra12a.html. 18

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. Weight uncertainty in neural networks. In *32nd International Conference on Machine Learning (ICML)*, 2015. URL http://proceedings.mlr.press/v37/blundell15.pdf. 2, 3, 38

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., and Vandergheynst, P. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4): 18–42, 2017. URL https://arxiv.org/abs/1611.08097. 40, 41

Caruana, R., Niculescu-Mizil, A., Crew, G., and Ksikes, A. Ensemble selection from libraries of models. In *Proceedings of the twenty-first international conference on Machine learning*, pp. 18, 2004. URL https://www.cs.cornell.edu/~caruana/ctp/ct.papers/caruana.icml04.icdm06long.pdf. 18

Cayton, L. Algorithms for manifold learning. *Univ. of California at San Diego Tech. Rep*, 12(1-17):1, 2005. URL https://www.lcayton.com/resexam.pdf. 5, 26

Foong, A. Y., Li, Y., Hernández-Lobato, J. M., and Turner, R. E. "In-between" uncertainty in bayesian neural networks. *arXiv preprint arXiv:1906.11537*, 2019. URL https://arxiv.org/abs/1906.11537. 8, 29, 30

Fort, S., Hu, H., and Lakshminarayanan, B. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019. URL https://arxiv.org/abs/1912.02757. 2

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *33rd International Conference on Machine Learning (ICML)*, pp. 1050–1059, 2016. URL https://proceedings.mlr.press/v48/gal16.html. 2, 7, 8, 20, 27, 29, 38

Garnelo, M., Rosenbaum, D., Maddison, C., Ramalho, T., Saxton, D., Shanahan, M., Teh, Y. W., Rezende, D., and Eslami, S. A. Conditional neural processes. In *International Conference on Machine Learning*, pp. 1704–1713. PMLR, 2018a. URL https://proceedings.mlr.press/v80/garnelo18a.html. 3, 46

Garnelo, M., Schwarz, J., Rosenbaum, D., Viola, F., Rezende, D. J., Eslami, S., and Teh, Y. W. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018b. URL https://arxiv.org/abs/1807.01622. 3, 46

Ghahramani, Z. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452–459, 2015. URL https://www.nature.com/articles/nature14541. 1

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., and Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018. URL https://doi.org/10.1021/acscentsci.7b00572. 1

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. URL https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf. 5

Graves, A. Practical variational inference for neural networks. In *Advances in neural information processing systems*, pp. 2348–2356, 2011. URL http://papers.nips.cc/paper/4329-practical-variational-inference-for-neural-networks.pdf. 2, 38

Gustafsson, F. K., Danelljan, M., and Schon, T. B. Evaluating scalable bayesian deep learning methods for robust computer vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 318–319, 2020. URL https://arxiv.org/abs/1906.01620. 2

Havasi, M., Jenatton, R., Fort, S., Liu, J. Z., Snoek, J., Lakshminarayanan, B., Dai, A. M., and Tran, D. Training independent subnetworks for robust prediction. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=OGg9XnKxFAH. 3

Heiss, J., Teichmann, J., and Wutte, H. How implicit regularization of relu neural networks characterizes the learned function – part i: the 1-d case of two layers with random first layer. *arXiv preprint arXiv:1911.02903*, 2019. URL https://doi.org/10.3929/ethz-b-000402003. 15

Hernández-Lobato, J. M. and Adams, R. Probabilistic backpropagation for scalable learning of bayesian neural networks. In *International Conference on Machine Learning*, pp. 1861–1869, 2015. URL http://proceedings.mlr.press/v37/hernandez-lobatoc15.pdf. 2, 8, 29, 38

Kendall, A. and Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pp. 5574–5584, 2017. URL https://papers.nips.cc/paper/2017/hash/2650d6089a6d640c5e85b2b88265dc2b-Abstract.html. 2, 43

Khosravi, A., Nahavandi, S., Creighton, D., and Atiya, A. F. Lower upper bound estimation method for construction of neural network-based prediction intervals. *IEEE transactions on neural networks*, 22(3):337–346, 2010. URL https://doi.org/10.1109/TNN.2010.2096824. 2, 6

Kuleshov, V., Fenner, N., and Ermon, S. Accurate uncertainties for deep learning using calibrated regression. In Dy, J. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2796–2804, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/kuleshov18a.html. 4, 6, 14, 19

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, pp. 6402–6413, 2017. URL http://papers.nips.cc/paper/7219-simple-and-scalable-predictive-uncertainty-estimation-using-deep-ensembles.pdf. 2, 8, 17, 20, 29

LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *Nature*, 521(7553):436–444, 2015. ISSN 1476-4687. doi: 10.1038/nature14539. URL https://www.nature.com/articles/nature14539. 1

Malinin, A. and Gales, M. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, pp. 7047–7058, 2018. URL https://papers.nips.cc/paper/2018/file/3ea2db50e62ceefceaf70a9d9a56a6f4-Paper.pdf. 3, 36, 43

Malinin, A., Chervontsev, S., Provilkov, I., and Gales, M. Regression prior networks, 2020a. URL https://arxiv.org/pdf/2006.11590.pdf. 3, 45

Malinin, A., Mlodozeniec, B., and Gales, M. Ensemble distribution distillation. In *International Conference on Learning Representations*, 2020b. URL https://openreview.net/forum?id=BygSP6Vtvr. 5

Martinez-Cantin, R., De Freitas, N., Brochu, E., Castellanos, J., and Doucet, A. A bayesian exploration-exploitation approach for optimal online sensing and planning with a visually guided mobile robot. *Autonomous Robots*, 27(2):93–103, 2009. URL https://doi.org/10.1007/s10514-009-9130-2. 1

Mukhoti, J., Stenetorp, P., and Gal, Y. On the importance of strong baselines in bayesian deep learning. *arXiv preprint arXiv:1811.09385*, 2018. URL https://arxiv.org/abs/1811.09385. 8, 29

Neal, R. M. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012. URL https://api.semanticscholar.org/CorpusID:60809283. 1

Nix, D. A. and Weigend, A. S. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 ieee international conference on neural networks (ICNN'94)*, volume 1, pp. 55–60. IEEE, 1994. URL https://doi.org/10.1109/ICNN.1994.374138. 2, 46

Ober, S. W. and Rasmussen, C. E. Benchmarking the neural linear model for regression. *arXiv preprint arXiv:1912.08416*, 2019. URL https://arxiv.org/abs/1912.08416. 3, 8, 29

Osband, I., Wen, Z., Asghari, M., Ibrahimi, M., Lu, X., and Van Roy, B. Epistemic neural networks. *arXiv preprint arXiv:2107.08924*, 2021. URL https://arxiv.org/abs/2107.08924. 7, 26, 27

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., and Snoek, J. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/8558cb408c1d76621371888657d2eb1d-Paper.pdf. 2

Pearce, T., Zaki, M., Brintrup, A., and Neely, A. High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. In *35th International Conference on Machine Learning (ICML)*, pp. 4075–4084, 2018. URL https://proceedings.mlr.press/v80/pearce18a.html. 2, 6

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. W. Information-theoretic regret bounds for gaussian process optimization in the bandit setting. *IEEE Transactions on Information Theory*, 58(5):3250–3265, May 2012. ISSN 1557-9654. doi: 10.1109/tit.2011.2182033. URL http://dx.doi.org/10.1109/TIT.2011.2182033. 1

Steinhilber, F., Beer, J., and Fröhlich, C. Total solar irradiance during the holocene. *Geophysical Research Letters*, 36(19), 2009. doi: 10.1029/2009GL040142. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2009GL040142. 7

Wahba, G. Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 40(3):364–372, 1978. URL https://doi.org/10.1111/j.2517-6161.1978.tb01050.x. 36

Wang, Z., Ren, T., Zhu, J., and Zhang, B. Function space particle optimization for bayesian neural networks. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=BkgtDsCcKQ. 3

Weissteiner, J. and Seuken, S. Deep learning-powered iterative combinatorial auctions. In *Proceedings of the 34th AAAI Conference of Artificial Intelligence*, pp. 2284–2293, 2020. URL https://ojs.aaai.org/index.php/AAAI/article/download/5606/5462. 4

Wen, Y., Tran, D., and Ba, J. Batchensemble: An alternative approach to efficient ensemble and lifelong learning, 2020. URL https://arxiv.org/pdf/2002.06715.pdf. 3

Wenzel, F., Roth, K., Veeling, B., Swiatkowski, J., Tran, L., Mandt, S., Snoek, J., Salimans, T., Jenatton, R., and Nowozin, S. How good is the Bayes posterior in deep neural networks really? In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 10248–10259. PMLR, 13–18 Jul 2020a. URL https://proceedings.mlr.press/v119/wenzel20a.html. 2, 18

Wenzel, F., Snoek, J., Tran, D., and Jenatton, R. Hyperparameter ensembles for robustness and uncertainty quantification. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020b. Curran Associates Inc. ISBN 9781713829546. URL https://papers.nips.cc/paper/2020/file/481fbfa59da2581098e841b7afc122f1-Paper.pdf. 2, 18, 20, 21

Williams, C. K. and Rasmussen, C. E. *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006. URL https://gaussianprocess.org/gpml/. 16, 37

Appendix

A. Theoretical Analysis of NOMU

In this section, we

1. first provide a theoretical motivation for the design of NOMU and establish via Theorem A.1 a connection to pointwise worst-case UBs $\underline{UB}_{\text{pw}}, \overline{UB}_{\text{pw}}$ **(Appendix A.1)**.
2. Next, we provide a Bayesian interpretation of those pointwise worst-case UBs $\underline{UB}_{\text{pw}}, \overline{UB}_{\text{pw}}$ and elaborate on *relative uncertainties* **(Appendix A.2)**.
3. Then, we discuss the case if $\mathcal{H}_{D^{\text{train}}}$ is not *upwards directed* as assumed in Theorem A.1 and further show how we deal with this challenge **(Appendix A.3)**.
4. Finally, we prove Theorem A.1 **(Appendix A.4)**.

A.1. Relating NOMU to Pointwise Worst-Case Uncertainty Bounds

In this section, we provide a theoretical motivation for the design of NOMU. To this end, we first define worst-case UBs. We then state a theorem connecting these worst case bounds and NOMU's UBs via NOMU's loss function. In what follows, we assume zero data noise $\sigma_{\text{n}} = 0$.

Consider a *hypothesis class* $\mathcal{H}$ given as a subset of a Banach space of functions $f : X \to Y$. Furthermore, let $\mathcal{H}_{D^{\text{train}}} := \{f \in \mathcal{H} : f(x_i^{\text{train}}) = y_i^{\text{train}}, i = 1, \dots, n^{\text{train}}\}$ denote the set of all functions from the hypothesis class that fit through the training points and let $\hat{f} \in \mathcal{H}_{D^{\text{train}}}$ be a prediction function (e.g., a NN trained on D^{train}). Worst-case bounds within the class $\mathcal{H}_{D^{\text{train}}}$ can be defined **p**oint**w**ise for each $x \in X$ as:

$$\underline{UB}_{\text{pw}}(x) := \inf_{f \in \mathcal{H}_{D^{\text{train}}}} f(x), \tag{5}$$

$$\overline{UB}_{\text{pw}}(x) := \sup_{f \in \mathcal{H}_{D^{\text{train}}}} f(x). \tag{6}$$

By definition, these UBs are the tightest possible bounds that cover every $f \in \mathcal{H}_{D^{\text{train}}}$ (i.e., $\underline{UB}_{\text{pw}}(x) \leq f(x) \leq \overline{UB}_{\text{pw}}(x)\ \forall x \in X$). From a Bayesian perspective, such bounds correspond to credible bounds for $\alpha = 1$ if the support of the prior is contained in $\mathcal{H}$. Interestingly, if $\mathcal{H}$ is the class of regularized NNs, these bounds can also be interpreted as an approximation of credible bounds for $\alpha < 1$ with respect to a Gaussian prior on the parameters of a NN (see Appendix A.2. for a derivation).[11]

In applications like BO, when optimizing an acquisition function based on these pointwise-defined bounds, we require the UBs for *all* $x \in X$. Thus, numerically optimizing such an acquisition function is practically infeasible, as it would require solving the optimization problems from (5) millions of times. In NOMU, we circumvent this problem by constructing the UBs for *all* $x \in X$ simultaneously. We can do so by only solving a *single* optimization problem, i.e., minimizing the NOMU loss from (4).

In the following theorem, we show that these pointwise-defined UBs can be computed by solving a single optimization problem under the following assumption.

Assumption 1 (UPWARDS DIRECTED) *For every $f_1, f_2 \in \mathcal{H}_{D^{train}}$ there exists an $f \in \mathcal{H}_{D^{train}}$ such that $f(x) \geq \max(f_1(x), f_2(x))$ for all $x \in X$.*

Theorem A.1 (SINGLE OPTIMIZATION PROBLEM) *Let $X = \prod_{i=1}^{d}[a_i, b_i] \subset \mathbb{R}^d$, $a_i < b_i$, let $Y = \mathbb{R}$, and let D^{train} be a nonempty set of training points. Furthermore, let $\mathcal{H}_{D^{train}} \subset (C(X, Y), \|\cdot\|_\infty)$ be compact and* upwards directed *and $\hat{f} \in \mathcal{H}_{D^{train}}$. Then, for every strictly-increasing and continuous $u : \mathbb{R} \to \mathbb{R}$, it holds that*

$$\overline{UB}_{pw} = \operatorname*{arg\,max}_{h \in \mathcal{H}_{D^{train}}} \int_X u(h(x) - \hat{f}(x))\, dx. \tag{7}$$

In words, $\overline{UB}_{\text{pw}}$ can be calculated via the single optimization problem (7) on $\mathcal{H}_{D^{\text{train}}}$.[12]

Proof. See Appendix A.4. □

In practice, Assumption 1 can be violated such that a straightforward calculation of the r.h.s. of (7) for an *arbitrary* u would result in unreasonable UBs. However, for a sensible choice of u, NOMU's UBs based on the r.h.s. of (7) still satisfy our Desiderata D1–D5, similar to $\overline{UB}_{\text{pw}}$ (see Appendix A.3 for a discussion).

The connection of NOMU's UBs to the pointwise worst-case bounds $\underline{UB}_{\text{pw}}, \overline{UB}_{\text{pw}}$ becomes clear by observing that minimizing NOMU's loss function L^π (Equation (4)) can be interpreted as solving the r.h.s of (7) for a specific choice of u, when $\mathcal{H}$ is the class of regularized NNs. In detail:

- Term (a) of the NOMU-loss (4) implements that $\hat{f}$ solves the regression task and thus $\hat{f} \in \mathcal{H}_{D^{\text{train}}}$ up to numerical precision (if the regularization λ is small enough).
- Term (b) enforces $\hat{r}_f(x_i^{\text{train}}) \approx 0$ and thus when defining $h := \hat{f} + \hat{r}_f$, we directly obtain $h(x_i^{\text{train}}) \approx y_i^{\text{train}}$ corresponding to the constraint $h \in \mathcal{H}_{D^{\text{train}}}$ in (7).
- While terms (a) and (b) enforce the constraints of (7), term (c) is the objective function of (7) for the specific choice of $u(z) := -e^{-c_{\text{exp}} z}, c_{\text{exp}} \in \mathbb{R}_{\geq 0}$.

[11] Standard BNNs also aim to approximate the posterior coming from exactly this prior.

[12] We formulate Theorem A.1 for upper UBs $\overline{UB}_{\text{pw}}$. The analogous statement also holds for lower UBs $\underline{UB}_{\text{pw}}$.

A.2. Bayesian Interpretation of Pointwise Worst-Case Uncertainty Bounds

In this Section, we provide a Bayesian interpretation of the pointwise worst-case UBs $\underline{UB}_{\text{pw}}, \overline{UB}_{\text{pw}}$ and elaborate on *relative uncertainties*.

In the following, we denote by $\mathcal{NN}^f_\theta : X \to Y$ a (standard) NN for model predictions. Note that $\mathcal{NN}^f_\theta$ does not represent the whole NOMU architecture but can be used as $\hat{f}$-sub-network in the NOMU architecture $\mathcal{NN}_\theta$. Furthermore, we consider the hypothesis class of regularized NNs, i.e., $\mathcal{H} := \left\{\mathcal{NN}^f_\theta : \|\theta\|_2 \leq \gamma\right\}$. Recall that one needs to assume that the prior is fully concentrated on $\mathcal{H}$ in order to interpret the pointwise UBs $\underline{UB}_{\text{pw}}, \overline{UB}_{\text{pw}}$ as $\alpha{=}1$ CBs. In the following, we will present an alternative Bayesian interpretation of $\overline{UB}_{\text{pw}}$.

Many other approaches (MC dropout, BNNs) assume a Gaussian prior on the parameters of the NNs, i.e, $\theta \sim \mathcal{N}(0, \sigma_\theta^2 I)$, and try to approximate the corresponding posterior CBs. Interestingly, $\underline{UB}_{\text{pw}}$ and $\overline{UB}_{\text{pw}}$ can also be seen as approximations of $\alpha{<}1$ CBs in the case of such a Gaussian prior on the parameters. This can be seen as follows:

Let the data generating process be given as $y = \mathcal{NN}^f_\theta + \varepsilon$, with $\varepsilon \sim \mathcal{N}(0, \sigma_{\text{n}}^2)$.[13] For a Gaussian prior on the parameters $\theta \sim \mathcal{N}(0, \sigma_\theta^2 I)$ the negative log posterior can be written as

$$-\log(p(\theta|D^{\text{train}})) = \overbrace{\frac{1}{2\sigma_{\text{n}}^2}\sum_{i=1}^{n^{\text{train}}}\left(\mathcal{NN}^f_\theta(x_i^{\text{train}}) - y_i^{\text{train}}\right)^2}^{=:L(\theta)} + \frac{\|\theta\|_2^2}{2\sigma_\theta^2} + n^{\text{train}}\log(\sigma_{\text{n}}) + C_{n^{\text{train}},\sigma_\theta}. \tag{8}$$

for a constant $C_{n^{\text{train}},\sigma_\theta} := \frac{n^{\text{train}}}{2}\log(2\pi) + \frac{1}{2}\log(2\pi\sigma_\theta^2)$. Then the pointwise upper UBs can be reformulated to

$$\overline{UB}_{\text{pw}}(x) \overset{\text{def.}}{=} \sup_{f \in \mathcal{H}_{D^{\text{train}}}} f(x) = \lim_{\sigma_{\text{n}} \to 0} \sup_{f \in \tilde{\mathcal{H}}^{\sigma_{\text{n}}}_{D^{\text{train}}}} f(x) \tag{9}$$

with

$$\tilde{\mathcal{H}}^{\sigma_{\text{n}}}_{D^{\text{train}}} := \left\{\mathcal{NN}^f_\theta : \frac{\sigma_\theta^2}{\sigma_{\text{n}}^2}L(\theta) + \|\theta\|_2^2 \leq \gamma\right\} \tag{10a}$$

$$= \left\{\mathcal{NN}^f_\theta : \log(p(\theta|D^{\text{train}})) \geq \tilde{\gamma}_{\sigma_{\text{n}}}\right\} \tag{10b}$$

where $\tilde{\gamma}_{\sigma_{\text{n}}} := -\frac{\gamma}{2\sigma_\theta^2} - n^{\text{train}}\log(\sigma_{\text{n}}) - C_{n^{\text{train}},\sigma_\theta}$.

Therefore, for small data noise $\sigma_{\text{n}} \approx 0$ we obtain from Equations (9) and (10b) that

$$\overline{UB}_{\text{pw}}(x) \approx \sup_\theta \left(\mathcal{NN}^f_\theta(x) : p(\theta|D^{\text{train}}) > e^{\tilde{\gamma}_{\sigma_{\text{n}}}}\right). \tag{11}$$

In words, from a Bayesian point of view, we seek the highest value $\mathcal{NN}^f_\theta(x)$ provided that the posterior density $p(\theta|D^{\text{train}}) > e^{\tilde{\gamma}_{\sigma_{\text{n}}}}$, which can be seen as a heuristic to approximate CBs similarly as the MAP from the parameter-space is a commonly used heuristic to approximate the posterior mean of y.

Remark A.2 *Note, that if $p(\mathcal{NN}^f_\theta(x)|D^{train}) = p(\theta|D^{train})$ and this posterior is unimodal, (11) corresponds to a classical* highest posterior density interval (HPDI).

Thus, pointwise worst-case UBs can be seen as an alternative method to approximate BNN-based CBs with the typically used Gaussian prior.

A.2.1. From Absolute to Relative Model Uncertainty

If the prior scale (e.g., γ or σ_θ in the above mentioned approaches) is known, in theory no further calibration is needed and one can interpret the resulting UBs in *absolute* terms (Kuleshov et al., 2018). However, typically, the prior scale is unknown and the resulting UBs can only be interpreted in terms of *relative* model uncertainty (i.e., how much more model uncertainty does one have at one point x compared to any other point x'?) and in a second step careful calibration of the resulting UBs is required.

Thus, there are two (almost) independent problems: First, the fundamental concept of how to estimate relative model uncertainty (such as MC dropout, deep ensembles, hyper deep ensembles or NOMU) and second, the calibration of these UBs. In this paper, we do not mix up these two challenges and only focus on the first one. Furthermore, desiderata D1–D4 and our metrics AUC and NLL_{min} do not depend on the scaling of the uncertainty. Whenever we use NLL as a metric, we make sure to calibrate the uncertainties by a scalar c to ensure that our evaluations do not depend on the scaling of the uncertainty predictions.

A.3. A Note on Theorem A.1

In practice, the set $\mathcal{H}_{D^{\text{train}}}$ often is not upwards directed for typical NN-architectures and Equation (7) of Theorem A.1 is not fulfilled in general. Indeed, $h^* := \arg\max_{h \in \mathcal{H}_{D^{\text{train}}}} \int_X u(h(x) - \hat{f}(x))\, d(x)$ can be much more overconfident than $\overline{UB}_{\text{pw}}$. However, in the following we will motivate why for a suitably chosen u the relative model uncertainty of h^* still fulfills the desiderata D1 (Non-Negativity), D2 (In-Sample) and D3 (Out-of-Sample).

Remark A.3 *In our proposed final algorithm from Section 3 NOMU, we incorporated D4 (Metric Learning) by modify-*

[13] For simplicity we assume homoskedastic noise in this section.

ing the network architecture as presented in Section 3.2.

Problem First, we give an example of a specific NN-architecture, where Theorem A.1 is not fulfilled due to a violation of the upwards directed assumption. Note that from a Lagrangian perspective it is equivalent to add a term $-\lambda \|\theta\|_2^2, \lambda \geq 0$ to the target function of (7) instead of bounding $\|\theta\|_2^2 \leq \gamma^2$ as a constraint in $\mathcal{H}_{D^{\text{train}}}$. Moreover, for a specific NN-architecture it is shown that regularizing $\|\theta\|_2^2$ is equivalent to regularizing the L_2-norm of the second derivative of the function $f = \mathcal{NN}_\theta^f$ (Heiss et al., 2019), i.e., regularizing $\int_X f''(x)^2\, dx$. If we choose for example $u = id : \mathbb{R} \to \mathbb{R}$, then increasing h in-between the two close points in the middle of Figure 7, would improve the target function of (7) $\int u(h(x) - \hat{f}(x))\, dx$ less than the additional regularization cost resulting from the additional second derivative when increasing h. Therefore, h^* will be below the mean prediction $\hat{f}$ in this region, break D1 (Non-Negativity), and $h^* \neq \overline{UB}_{\text{pw}}$.

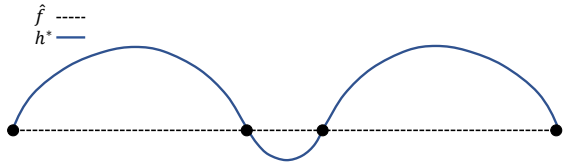


Figure 7. h^* would not fulfill D1 (Non-Negativity), if u was chosen as the identity and the architecture from Heiss et al. (2019) was used.

Solution However, if e.g., we choose $u : x \mapsto -e^{-c_{\text{exp}}x}$ with c_{exp} large enough, we highly penalize $h < \hat{f}$, and thus $\hat{f} \leq h^*$ D1 (Non-Negativity) is fulfilled. Since $h^* \in \mathcal{H}_{D^{\text{train}}}$, it follows that $h^*(x_i^{\text{train}}) - \hat{f}(x_i^{\text{train}}) = 0$ for all training points. This together with $\hat{f} \leq h^*$ implies that $\nabla \left(h^* - \hat{f}\right)(x_i^{\text{train}}) = 0$ at the training points (in the interior of X), which subsequently interrupts the downwards trend of $h^* - \hat{f}$ from one side of a point to the other side of a point as depicted in Figure 8.

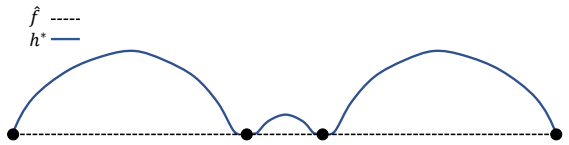


Figure 8. h^* fulfills D1–D3 if u prevents it from getting negative. See Appendix D for a detailed discussion how NOMU fulfills all five desiderata.

A.4. Proof of Theorem A.1

In the following, we prove Theorem A.4, an even more general version of Theorem A.1. The statement of Theorem A.1 follows from Theorem A.4 by setting

1. $X = \prod_{i=1}^d [a_i, b_i] \subset \mathbb{R}^d$ compact, $Y := \mathbb{R}$
2. $\mu = \lambda_d$, where λ_d is the Lebesgue measure on $\mathbb{R}^d$ (which is finite on every compact set and has full support).

Theorem A.4 *Let X be a nonempty topological space equipped with a finite measure μ with full support[14], let Y be a normed and totally ordered space and let D^{train} denote a set of observations. Moreover, let $\mathcal{H}_{D^{\text{train}}} \subset (C(X,Y), \|\cdot\|_\infty)$ be compact, and $\hat{f} \in \mathcal{H}_{D^{\text{train}}}$. Assume further that $\mathcal{H}_{D^{\text{train}}}$ is upwards directed (see Assumption 1). Then, for every strictly-increasing and continuous $u : Y \to \mathbb{R}$ it holds that*

$$\overline{UB}_{\text{pw}} = \operatorname*{arg\,max}_{h \in \mathcal{H}_{D^{\text{train}}}} \int_X u(h(x) - \hat{f}(x))\, d\mu(x). \tag{12}$$

Proof. First note that since μ is finite and h, $\hat{f}$ and $\overline{UB}_{\text{pw}}$ are bounded since $\mathcal{H}_{D^{\text{train}}}$ is assumed to be compact with respect to the $\|\cdot\|_\infty$-topology, it holds that the integral in (12) is finite.

Let $h^* \in \mathcal{H}_{D^{\text{train}}}$ denote an optimizer of (12), which exists since $\mathcal{H}_{D^{\text{train}}}$ is assumed to be compact and the operator $h \mapsto \int_X u(h(x) - \hat{f}(x))\, d\mu(x)$ is continuous on the $\|\cdot\|_\infty$-topology.

We want to show that $h^*(x) = \overline{UB}_{\text{pw}}(x)$ for every $x \in X$.

Note that per definition for all $x \in X$ and $h \in \mathcal{H}_{D^{\text{train}}}$ it holds that

$$\overline{UB}_{\text{pw}}(x) = \sup_{f \in \mathcal{H}_{D^{\text{train}}}} f(x) \geq h(x). \tag{13}$$

Thus $\overline{UB}_{\text{pw}}(x) \geq h^*(x)$ for all $x \in X$.

For the reverse inequality assume on the contrary that there exists an $x' \in X$ such that

$$\overline{UB}_{\text{pw}}(x') > h^*(x'). \tag{14}$$

Then, we define $f_{x'} := \arg\max_{f \in \mathcal{H}_{D^{\text{train}}}} f(x')$, which exists because of compactness and continuity. Since $f_{x'}$ and h^* are both continuous and $f_{x'}(x') = \overline{UB}_{\text{pw}}(x') > h^*(x')$ there exists a neighbourhood $U_{x'}$ of x' such that

$$f_{x'}(x) > h^*(x) \text{ for all } x \in U_{x'}, \tag{15}$$

Using the *upwards directed* property with $f_1 := f_{x'}$ and $f_2 := h^*$, it follows that there exists a $\tilde{h} \in \mathcal{H}_{D^{\text{train}}}$ with

$$\tilde{h}(x) \geq \max(f_{x'}(x), h^*(x)) \text{ for all } x \in X. \tag{16}$$

[14] I.e. every nonempty open set has non-zero measure.

Using (15) together with (16) implies further that

$$\tilde{h}(x) \geq h^*(x) \text{ for all } x \in X \text{ and} \tag{17}$$
$$\tilde{h}(x) > h^*(x) \text{ for all } x \in U_{x'}. \tag{18}$$

However, since u is strictly increasing and $\mu(U_{x'}) > 0$ by the full support assumption, we get that

$$\int_X u(\tilde{h}(x) - \hat{f}(x))\, d\mu(x) > \int_X u(h^*(x) - \hat{f}(x))\, d\mu(x), \tag{19}$$

which is a contradiction to the assumption that h^* is the optimizer of (12). Therefore, it holds that $\overline{UB}_{\text{pw}}(x) \leq h^*(x)$ for all $x \in X$.

In total we get that $\overline{UB}_{\text{pw}}(x) = h^*(x)$ for all $x \in X$, which concludes the proof. □

Remark A.5 *For our algorithm we select* $\mathcal{H} := \left\{\mathcal{NN}_\theta^f : \|\theta\|_2 \leq \gamma\right\}$ *to be the class of regularized NNs with a continuous activation function. Thus, the assumption of* $\mathcal{H}_{D^{train}}$ *being compact and a subset of* $C(X, Y)$ *is fulfilled.*

B. Experiments

B.1. Benchmark Methods

In this Section, we give a brief overview of each considered benchmark algorithm.

B.1.1. Gaussian Process (GP)

A GP defines a distribution over a set of functions $\{f : X \to Y\}$ where every finite collection of function evaluations follows a Gaussian distribution (see Williams & Rasmussen (2006) for a survey on GPs). A GP is completely specified by a *mean* $m : X \to Y$ and a *kernel* function $k_\pi : X \times X \to Y$, where π denotes a tuple of hyper-parameters. More formally, for any finite set of $k \in \mathbb{N}$ input points $\boldsymbol{x} := \{x_1, \ldots, x_k\}$, $x_i \in X$ and for $\boldsymbol{f} := (f_1, \ldots, f_k) \in Y^k$ with $f_i := f(x_i)$ it holds that

$$\boldsymbol{f} \sim \mathcal{N}_k\left(\boldsymbol{m}(\boldsymbol{x}), \boldsymbol{K}_\pi(\boldsymbol{x}, \boldsymbol{x})\right), \tag{20}$$

i.e., it follows a k-dimensional Gaussian distribution with covariance (or Gramian) matrix $[\boldsymbol{K}_\pi(\boldsymbol{x}, \boldsymbol{x})]_{i,j} := k_\pi(x_i, x_j)$, and mean vector $\boldsymbol{m}(\boldsymbol{x}) := (m(x_1), \ldots, m(x_k))$. Let

$$f \sim \mathcal{GP}(m(\cdot), k_\pi(\cdot, \cdot)), \tag{21}$$

denote a GP with mean function m and kernel k_π.

In the following, we summarize the main steps of GP regression for a 1D-setting and $m \equiv 0$.

1. *Define probabilistic model:*

$$y = f(x) + \varepsilon. \tag{22}$$

2. *Specify prior (kernel and data noise):*

$$f \sim \mathcal{GP}(0, k_\pi(\cdot, \cdot)), \tag{23}$$
$$\varepsilon | x \sim \mathcal{N}\left(0, \sigma_{\text{n}}^2(x)\right). \tag{24}$$

3. Calculate *likelihood* for training points $\boldsymbol{x}$ and $\boldsymbol{y} := \{y_1, \ldots, y_k\}$:

$$\boldsymbol{y} | \boldsymbol{x}, \pi \sim \mathcal{N}_k\left(\boldsymbol{0}, \boldsymbol{K}_\pi(\boldsymbol{x}, \boldsymbol{x}) + diag(\sigma_{\text{n}}^2(\boldsymbol{x}))\right), \tag{25}$$

with $\sigma_{\text{n}}^2(\boldsymbol{x}) := \left(\sigma_{\text{n}}^2(x_1), \ldots, \sigma_{\text{n}}^2(x_k)\right)$.

4. *Optimize kernel hyper-parameters (optional):*

$$\hat{\pi} \in \arg\max_\pi p(\boldsymbol{y} | \boldsymbol{x}, \pi). \tag{26}$$

5. Calculate *posterior predictive distribution* for new point (x^*):

$$f(x^*) | x^*, \boldsymbol{y}, \boldsymbol{x}, \hat{\pi} \sim \mathcal{N}\left(\hat{\mu}(x^*), \hat{\sigma}_f^2(x^*)\right), \tag{27}$$

where for $A := \left(\boldsymbol{K}_{\hat{\pi}}(\boldsymbol{x}, \boldsymbol{x}) + diag(\sigma_{\text{n}}^2(\boldsymbol{x}))\right)$ and $\boldsymbol{k}_{\hat{\pi}}(x^*, \boldsymbol{x}) := (k_{\hat{\pi}}(x^*, x_1), \ldots, k_{\hat{\pi}}(x^*, x_k))$ the parameters are given as

$$\hat{\mu}(x^*) := \boldsymbol{k}_{\hat{\pi}}(x^*, \boldsymbol{x}) A^{-1} f(\boldsymbol{x}) \tag{28}$$
$$\hat{\sigma}_f^2(x^*) := k_{\hat{\pi}}(x^*, x^*) - \boldsymbol{k}_{\hat{\pi}}(x^*, \boldsymbol{x}) A^{-1} \boldsymbol{k}_{\hat{\pi}}(x^*, \boldsymbol{x})^T. \tag{29}$$

Setting the data noise to zero $\sigma_{\text{n}} \equiv 0$ in (28) and (29) yields the mean prediction $\hat{f}$ and the model uncertainty prediction $\hat{\sigma}_f^2$ as

$$\hat{f}(x^*) = \boldsymbol{k}_{\hat{\pi}}(x^*, \boldsymbol{x}) \left(\boldsymbol{K}_{\hat{\pi}}(\boldsymbol{x}, \boldsymbol{x})\right)^{-1} f(\boldsymbol{x}), \tag{30}$$
$$\hat{\sigma}_f^2(x^*) = k_{\hat{\pi}}(x^*, x^*) - \boldsymbol{k}_{\hat{\pi}}(x^*, \boldsymbol{x}) \left(\boldsymbol{K}_{\hat{\pi}}(\boldsymbol{x}, \boldsymbol{x})\right)^{-1} \boldsymbol{k}_{\hat{\pi}}(x^*, \boldsymbol{x})^T, \tag{31}$$

which are then used to define the Gaussian process's UBs by $\left(\hat{f}(x) \mp c\, \hat{\sigma}_f(x)\right)$ with a calibration parameter $c \in \mathbb{R}_{\geq 0}$.

B.1.2. Monte Carlo Dropout (MCDO)

Let $\mathcal{NN}_\theta^f = W^K \circ \phi \circ W^{K-1} \circ \ldots \circ \phi \circ W^1(x)$ be an NN with $K - 1$ hidden layers, activation function ϕ, and fixed parameters $\theta = \{W^1, \ldots, W^K\}$, which have been trained with added dropout regularization. Furthermore, let $(p_1, \ldots, p_K)$ denote the dropout probability vector used when training $\mathcal{NN}_\theta^f$, i.e, p_i determines the probability for a single node in the i^{th} hidden layer W^i to be dropped in each backpropagation step.[15]

To obtain model uncertainty one draws M different NNs according to the dropout probability vector and represents model uncertainty using sample estimates of the mean and

[15] One could also use different probabilities p_{ij} for each node within a hidden layer. The equations extend straightforwardly.

variance of the model predictions.[16] These predictions are frequently termed *stochastic forward passes*. More formally, given a dropout probability vector $(p_1, \dots, p_K)$, one draws M realisations $\{\theta^{(1)}, \dots, \theta^{(M)}\}$ of parameters θ, where $\theta^{(m)} := \{W^{1,(m)}, \dots, W^{K,(m)}\}$. $W^{k,(m)}$ is obtained from the original hidden layer W^k by dropping each column with probability p_i, i.e., for $W^k \in \mathbb{R}^{d_{row} \times d_{col}}$ set

$$W^{k,(m)} = W^k \left[z_1^{(m)}, \dots, z_{d_{col}}^{(m)}\right], \; z_j^{(m)} \in \mathbb{R}^{d_{row}} \tag{32}$$

$$\text{where } z_j^{(m)} := \begin{cases} \mathbf{0}, \text{ with probability } p_i \\ \mathbf{1}, \text{ with probability } 1 - p_i. \end{cases} \tag{33}$$

UBs that represent model uncertainty and *known* data noise σ_{n}^2 are then estimated for each $x \in X$ as

$$\hat{f}(x) := \frac{1}{M} \sum_{m=1}^{M} \mathcal{NN}_{\theta^{(m)}}^f(x), \tag{34}$$

$$\hat{\sigma}^2(x) := \underbrace{\frac{1}{M} \sum_{m=1}^{M} \left(\mathcal{NN}_{\theta^{(m)}}^f(x) - \hat{f}(x)\right)^2}_{\text{model uncertainty}} + \underbrace{\sigma_{\text{n}}^2(x)}_{\text{data noise}}. \tag{35}$$

The model uncertainty prediction $\hat{\sigma}_f^2$ is then given as

$$\hat{\sigma}_f^2(x) := \frac{1}{M} \sum_{m=1}^{M} \left(\mathcal{NN}_{\theta^{(m)}}^f(x) - \hat{f}(x)\right)^2 \tag{36}$$

which defines Mc dropout's UBs as $\left(\hat{f}(x) \mp c\, \hat{\sigma}_f(x)\right)$ with a calibration parameter $c \in \mathbb{R}_{\geq 0}$.

B.1.3. Deep Ensembles (DE)

Deep ensembles consists of the following two steps:[17]

1. Use a NN to define a predictive distribution $p_\theta(y|x)$, select a pointwise loss function (proper scoring rule) $\ell(p_\theta, (x, y))$, which measures the quality of the predictive distribution $p_\theta(y|x)$ for an observation (x, y) and define the empirical loss used for training as

$$L(\theta) := \sum_{(x,y) \in D^{\text{train}}} \ell(p_\theta, (x, y)). \tag{37}$$

2. Use an ensemble of NNs, each with different randomly initialized parameters to represent model uncertainty.

[16] Alternatively, one could also determine the UBs using empirical upper and lower quantiles of the different model predictions.

[17] Lakshminarayanan et al. (2017) also considered in their paper a third step: *adversarial training*. However, as the authors point out, the effectiveness of adversarial training drops quickly as the number of networks in the ensemble increases. Therefore, we do not consider adversarial training in this paper.

Concretely, for regression, Lakshminarayanan et al. (2017) use a NN $\mathcal{NN}_\theta^f$ with two outputs: $\hat{\mu}^\theta$ (mean prediction) and $\hat{\sigma}_{\text{n}}^\theta$ (data noise prediction) and train it using as pointwise loss function the Gaussian negative log-likelihood, i.e,

$$p_\theta(y|x) := \mathcal{N}\left(y; \hat{\mu}^\theta(x), \left(\hat{\sigma}_{\text{n}}^\theta(x)\right)^2\right), \tag{38}$$

$$\ell(p_\theta, (x, y)) := \frac{\log\left(\left(\hat{\sigma}_{\text{n}}^\theta(x)\right)^2\right)}{2} + \frac{(\hat{\mu}^\theta(x) - y)^2}{2\left(\hat{\sigma}_{\text{n}}^\theta(x)\right)^2}. \tag{39}$$

To add model uncertainty, Lakshminarayanan et al. (2017) use an ensemble of M NNs $\{\mathcal{NN}_{\theta^{(1)}}^f, \dots, \mathcal{NN}_{\theta^{(M)}}^f\}$, where each NN outputs a mean and data noise prediction, i.e, for $x \in X$ and $m \in \{1, \dots, M\}$

$$\mathcal{NN}_{\theta^{(m)}}^f(x) := \left(\hat{\mu}^{\theta^{(m)}}(x), \hat{\sigma}_{\text{n}}^{\theta^{(m)}}(x)\right). \tag{40}$$

This then defines the learned predictive distribution for each NN in regression as

$$p_{\theta^{(m)}}(y|x) = \mathcal{N}\left(y; \hat{\mu}^{\theta^{(m)}}(x), \left(\hat{\sigma}_{\text{n}}^{\theta^{(m)}}(x)\right)^2\right). \tag{41}$$

Finally, the ensemble is treated as uniformly-weighted Gaussian mixture model, i.e.,

$$\frac{1}{M} \sum_{m=1}^{M} p_{\theta^{(m)}}(y|x), \tag{42}$$

which is further approximated using a single Gaussian by matching the first and second moments. The deep ensemble predictions for the mean $\hat{f}$ and for the combined model uncertainty and data noise $\hat{\sigma}^2$ are given as

$$\hat{f}(x) := \frac{1}{M} \sum_{m=1}^{M} \hat{\mu}^{\theta^{(m)}}(x), \tag{43}$$

$$\hat{\sigma}^2(x) := \underbrace{\frac{1}{M} \sum_{m=1}^{M} \left(\hat{\sigma}_{\text{n}}^{\theta^{(m)}}(x)\right)^2}_{\text{data noise}} + \underbrace{\frac{1}{M} \sum_{m=1}^{M} \left[\hat{\mu}^{\theta^{(m)}}(x) - \hat{f}(x)\right]^2}_{\text{model uncertainty}}. \tag{44}$$

The model uncertainty prediction $\hat{\sigma}_f^2$ is then given as

$$\hat{\sigma}_f^2(x) := \frac{1}{M} \sum_{m=1}^{M} \left[\hat{\mu}^{\theta^{(m)}}(x) - \hat{f}(x)\right]^2. \tag{45}$$

which defines deep ensembles' UBs as $\left(\hat{f}(x) \mp c\, \hat{\sigma}_f(x)\right)$ with a calibration parameter $c \in \mathbb{R}_{\geq 0}$.

Remark B.1 *For known data noise σ_n, no estimation is required and one can use an NN $\mathcal{NN}_\theta^f$ with only one output $\hat{\mu}^\theta$. If additionally the data noise σ_n is assumed to be homoskedastic, one can train $\mathcal{NN}_\theta^f$ using the mean squared error (MSE) with suitably chosen L2-regularization parameter λ. To obtain predictive bounds instead of credible bounds, one can add σ_n^2 to (45) at the end.*

B.1.4. HYPER DEEP ENSEMBLES (HDE)

Hyper deep ensembles (HDE) (Wenzel et al., 2020b) is a simple extension of DE. In HDE, the ensemble is designed by varying not only weight initializations, but also hyperparameters. HDE involves (i) a random search over different hyperparameters and (ii) stratification across different random initializations. HDE inherits all the components (e.g., the architecture, or the loss function) from DE, which are presented in detail in Appendix B.1.3. Specifically, formulas (43) and (44) for the mean prediction $\hat{f}$ and the model uncertainty prediction $\hat{\sigma}_f^2(x)$ are the same for HDE.

Let $\mathcal{NN}^f_{\theta^{(m)},\pi^{(m)}}$ denote a NN for model predictions with weights $\theta^{(m)}$ and hyperparameters $\pi^{(m)}$ (e.g. the dropout rate). Furthermore, let $\texttt{rand\_search}(\kappa)$ denote the random search algorithm from (Bergstra & Bengio, 2012), where κ is the desired number of different NNs with randomly sampled hyperparameters.

The only difference of HDE compared to DE is the procedure how the ensemble is built, which we recall in Algorithm 2. Algorithm 2 uses as subprocedure Algorithm 1 from (Caruana et al., 2004), which we present first.

Algorithm 1 greedily grows an ensemble among a given pre-defined set of models $\mathcal{M}$, until some target size M is met, by selecting *with-replacement* the NN leading to the best improvement of a certain score $\mathcal{S}$ on a validation set.[18]

Algorithm 1 `hyper_ens` (Caruana et al., 2004)

input $\mathcal{M}$,M
  Ensemble $\mathcal{E} := \{\}$; Score $\mathcal{S}(\cdot)$; $s_{best} = +\infty$
  while $|\mathcal{E}.\text{unique}()| \leq M$ **do**
    $\mathcal{NN}^f_{\theta^*} \in \arg\min_{\mathcal{NN}^f_\theta \in \mathcal{M}} \mathcal{S}\left(\mathcal{E} \cup \{\mathcal{NN}^f_\theta\}\right)$
    if $\mathcal{S}\left(\mathcal{E} \cup \{\mathcal{NN}^f_\theta\}\right) < s_{best}$ **then**
      $\mathcal{E} = \mathcal{E} \cup \{\mathcal{NN}^f_\theta\}$
    else
output $\mathcal{E}$
    end if
  end while
output $\mathcal{E}$

Note that the Algorithm 1 does not require the NNs to have the same random weight initialization. However, Wenzel et al. (2020b) consider a fixed initialization for HDE to isolate the effect of just varying the hyperparameters.

Finally, Algorithm 2 builds an ensemble of at most M unique NNs which can exploit both sources of diversity: different *random initializations* and different choices of *hyperparameters*.

[18] Note that we use a slightly different notation than Wenzel et al. (2020a) here.

Algorithm 2 `hyper_deep_ens` (Wenzel et al., 2020b)

input M, κ
  $\mathcal{M}_0 = \{\mathcal{NN}^f_{\theta^{(j)},\lambda^{(j)}}\}_{j=1}^{\kappa} \leftarrow \texttt{rand\_search}(\kappa)$
  $\mathcal{E}_0 \leftarrow \texttt{hyper\_ens}(\mathcal{M}_0, M)$
  $\mathcal{E}_{strat} = \{\}$
  for all $\mathcal{NN}^f_{\theta,\lambda} \in \mathcal{E}_0$.unique() **do**
    for all $m \in \{1, \ldots, M\}$ **do**
      $\theta' \leftarrow$ random initialization (seed number m)
      $\mathcal{NN}^f_{\theta^{(m)},\lambda} \leftarrow$ train $\mathcal{NN}^f_{\theta',\lambda}$
      $\mathcal{E}_{strat} = \mathcal{E}_{strat} \cup \{\mathcal{NN}^f_{\theta^{(m)},\lambda}\}$
    end for
  end for
output $\texttt{hyper\_ens}(\mathcal{E}_{strat}, M)$

B.2. Regression

B.2.1. METRICS

In this Section, we provide details on the three metrics, which we use to assess the quality of UBs in the regression settings. In the following, let

$$\left(\underline{UB}_c(x), \overline{UB}_c(x)\right) := \left(\hat{f}(x) \mp c\, \hat{\sigma}_f(x)\right) \tag{46}$$

denote UBs obtained from any of the considered models via a model prediction $\hat{f}$, a model uncertainty prediction $\hat{\sigma}_f$, and a calibration parameter $c \in \mathbb{R}_{\geq 0}$. Furthermore, we use the following shorthand notation:

$$\mathcal{UB}_c(x) := \left(\underline{UB}_c(x), \overline{UB}_c(x)\right). \tag{47}$$

In the following, let $D := \{(x_i, y_i) \in X \times Y, i \in \{1, \ldots, n\}\}$, with $n \in \mathbb{N}$ denote a set of input-output points (depending on the specific purpose, D would typically refer to a validation or test set).

Mean Width vs. Coverage Probability We first formalize the concepts of mean width (MW) and coverage probability (CP). Then we define the metric AUC.

Definition B.2 (COVERAGE PROBABILITY) *Let D denote a set of input-output points, and let $\mathcal{UB}_c$ denote UBs for a calibration parameter c. Then the coverage probability is defined as*

$$CP\left(D \,|\, \mathcal{UB}_c\right) := \frac{1}{|D|} \sum_{(x,y)\in D} \mathbb{1}_{\{\underline{UB}_c(x) \leq y \leq \overline{UB}_c(x)\}}. \tag{48}$$

Definition B.3 (MEAN WIDTH) *Let D denote a set of input-output points, and let $\mathcal{UB}_c$ denote UBs for a calibration parameter c. Then the mean width is defined as*

$$MW\left(D \,|\, \mathcal{UB}_c\right) := \frac{1}{|D|} \sum_{(x,y)\in D} |\overline{UB}(x) - \underline{UB}(x)|. \tag{49}$$

Remark B.4 *Note that in Definition B.3, uncovered points at which UBs (for some fixed calibration parameter c) are narrow have a positive effect on overall mean width. In order not to reward such overconfident mispredictions, a possible remedy is to consider in (49) only the subset $D^{capt} \subset D$ of input-output points captured by the UBs, i.e.,*

$$D^{capt} := \left\{(x,y) \in D : \underline{UB}(x) \leq y \leq \overline{UB}(x)\right\}.$$

However, focusing on captured data points only punishes UBs that capture some points with large widths and almost cover others with small widths unnecessarily harshly compared to UBs for which the reverse is true. In other words, a slight change in calibration parameter c can lead to very diverse evaluations of UBs that have been assessed almost equal under the original c. Since ultimately we are interested in comparing UBs based on a range of calibration parameters (see Measure 1) we decided to include all points in the calculation of MW in our experiments.

Ideally, MW should be as small as possible, while CP should be close to its maximal value 1. Clearly, CP is counteracting MW. This naturally motivates considering ROC-like curves, plotting MW against CP for a range of calibration parameters c, and comparing different UBs based on their *area under the curves* (AUC).

Metric 1 (AUC) *Let D denote a set of input-output points. Define further c^* as the minimal calibration parameter achieving full coverage of D for given UBs $\mathcal{UB}_c$, i.e.,*

$$c^* := \arg\min_{c \geq 0} \{CP\left(D \,|\, \mathcal{UB}_c\right) = 1\}.$$

AUC is then defined as the integral of the following curve

$$\{(CP\left(D \,|\, \mathcal{UB}_c\right), MW\left(D \,|\, \mathcal{UB}_c\right)) \,:\, c \in [0, c^*]\}.$$[19]

Negative Log-likelihood Next, we first define the second metric: *average*[20] *negative log (Gaussian) likelihood (NLL)* and then present our third metric NLL_{min}.

Metric 2 (NLL) *Let D denote a set of input-output points, and let $\mathcal{UB}_c$ denote UBs for a calibration parameter c with corresponding model prediction $\hat{f} : X \to \mathbb{R}$ and model uncertainty prediction $\hat{\sigma}_f : X \to \mathbb{R}_{\geq 0}$. Then NLL is defined as*

$$NLL(D|\mathcal{UB}_c) := \tag{50}$$
$$\frac{1}{|D|} \sum_{(x,y)\in D} \left[\frac{\left(y - \hat{f}(x)\right)^2}{2\left(c\hat{\sigma}_f(x)\right)^2} + \ln\left(c\hat{\sigma}_f(x)\right) \right] + \ln(2\pi)/2$$

The first term in NLL measures the error of the model prediction, where large errors can be attenuated by large uncertainties while the second term penalizes logarithmically larger uncertainties. Thus, NLL penalizes both over-confident ($c\hat{\sigma}_f(x) \approx 0$) wrong predictions as well as under-confident ($c\hat{\sigma}_f(x) \gg 0$) predictions.

We define the third metric as the minimal value of the NLL when varying the calibration parameter c.

Metric 3 (MINIMUM NLL) *Let D denote a set of input-output points, and let $\mathcal{UB}_c$ denote UBs for a calibration parameter c. Then NLL_{min} is defined as*

$$NLL_{min} := \min_{c \in \mathbb{R}_{\geq 0}} NLL(D|\mathcal{UB}_c). \tag{51}$$

Remark B.5 (CALIBRATION) *Different approaches of obtaining UBs require rather different scaling to reach similar CP (as well as MW). For instance, as mentioned above, we found that the calibration parameter c required to achieve a certain value of CP typically is larger for DE than those of the remaining methods by a factor of approximately 10. Therefore, in practice it is important to find a well-calibrated parameter c. Standard calibration techniques can be applied to NOMU (e.g. methods based on isotonic regression Kuleshov et al. (2018), or when assuming Gaussian marginals of the posterior one can select c as the $\arg\min_{c \in \mathbb{R}_{\geq 0}} NLL(D|\mathcal{UB}_c)$ on a validation set D).*

B.2.2. TOY REGRESSION: CONFIGURATION DETAILS

All NN-based methods are fully connected feed forward NNs with ReLU activation functions, implemented in TENSORFLOW.KERAS and trained for 2^{10} epochs of TENSORFLOW.KERAS' Adam stochastic gradient descent with standard learning rate 0.001 and full batch size of all training points. Moreover, weights and biases are initialized uniformly in the interval $[-0.05, 0.05]$.

NOMU Setup For the MC-integration of the integral in the NOMU loss (4), i.e., term (c), we use $l = 128$ and $l = 256$ artificial input points (we draw new artificial input points in every gradient descent step) in 1D and 2D, respectively.[21] For numerical stability, we use the parameters θ

[19] In our experiments, we approximated this integral via the trapezoidal rule.

[20] We remark that NLL can be interpreted as average marginal predictive density over the set D, assuming these marginals are Gaussian with mean $\hat{f}$ and variance $c\hat{\sigma}_f$. In particular, we refrain from posing assumptions of posterior independence, that we believe are highly flawed, i.e., NLL should not be interpreted as joint negative log predictive density.

[21] In 1D, we tried MC-integration based on uniform samples and MC-integration based on a deterministic grid, where both approaches led to qualitatively similar results and the latter is presented in this paper.

Table 5. Detailed results of the AUC metric for NOMU, GP, MCDO, DE, HDE and HDE* (our HDE see Appendix B.2.2) for all ten 1D synthetic functions. Shown are the medians and a 95% bootstrap confidence interval over 500 runs. Winners are marked in grey.

	NOMU		GP		MCDO		DE		HDE		HDE*	
FUNCTION	AUC↓	95%-CI	AUC↓	95%-CI	AUC	95%-CI	AUC↓	95%-CI	AUC↓	95%-CI	AUC↓	95%-CI
ABS	0.07	[0.07, 0.08]	0.14	[0.13, 0.16]	0.16	[0.15, 0.17]	0.08	[0.07, 0.09]	1.02	[0.82, 1.20]	0.72	[0.65, 0.81]
STEP	0.29	[0.26, 0.31]	0.93	[0.85, 1.02]	0.25	[0.24, 0.27]	0.14	[0.13, 0.15]	1.90	[1.61, 2.67]	0.82	[0.66, 0.99]
KINK	0.08	[0.08, 0.09]	0.17	[0.15, 0.19]	0.22	[0.21, 0.23]	0.09	[0.07, 0.10]	1.81	[1.40, 2.31]	1.01	[0.9 , 1.24]
SQUARE	0.12	[0.11, 0.13]	0.16	[0.13, 0.19]	0.38	[0.36, 0.41]	0.14	[0.12, 0.15]	1.72	[1.50, 2.23]	1.86	[1.54, 2.53]
CUBIC	0.07	[0.07, 0.08]	0.00	[0.00, 0.00]	0.10	[0.10, 0.11]	0.10	[0.09, 0.11]	0.63	[0.54, 0.79]	0.52	[0.45, 0.66]
SINE 1	1.10	[1.03, 1.16]	1.14	[1.09, 1.2]	0.90	[0.87, 0.93]	1.21	[1.16, 1.26]	9.38	[7.42, 11.4]	9.75	[7.64, 12.1]
SINE 2	0.38	[0.37, 0.40]	0.42	[0.41, 0.44]	0.36	[0.35, 0.36]	0.50	[0.47, 0.53]	3.18	[2.82, 4.08]	2.43	[2.12, 3.38]
SINE 3	0.20	[0.19, 0.21]	0.31	[0.29, 0.34]	0.28	[0.26, 0.29]	0.20	[0.19, 0.21]	1.43	[1.19, 1.73]	1.33	[1.10, 1.57]
FORRESTER	0.19	[0.18, 0.20]	0.18	[0.17, 0.19]	0.26	[0.24, 0.28]	0.25	[0.24, 0.27]	1.28	[1.11, 1.51]	1.58	[1.35, 1.93]
LEVY	0.40	[0.39, 0.42]	0.53	[0.51, 0.56]	0.38	[0.36, 0.39]	0.45	[0.43, 0.48]	3.43	[2.67, 4.42]	3.58	[2.83, 4.18]

Table 6. Detailed results of the NLL_{min} metric (without constant $\ln(2\pi)/2$) for NOMU, GP, MCDO, DE, HDE and HDE* (our HDE see Appendix B.2.2) for all ten 1D synthetic functions. Shown are the medians and a 95% bootstrap confidence interval over 500 runs. Winners are marked in grey.

	NOMU		GP		MCDO		DE		HDE		HDE*	
FUNCTION	NLL_{MIN}↓	95%-CI	NLL_{MIN}↓	95%-CI	NLL_{MIN}↓	95%-CI	NLL_{MIN}↓	95%-CI	NLL_{MIN}↓	95%-CI	NLL_{MIN}↓	95%-CI
ABS	−2.85	[−2.95, −2.80]	−2.82	[−2.90, −2.71]	−1.76	[−1.80, −1.69]	−2.91	[−3.03, −2.76]	−0.19	[−0.36, 0.00]	−0.72	[−0.82, −0.56]
STEP	−1.04	[−1.12, −0.97]	−0.61	[−0.81, −0.49]	−0.78	[−0.85, −0.69]	−2.49	[−2.59, −2.36]	0.61	[0.38, 0.83]	−1.18	[−1.42, −0.84]
KINK	−2.74	[−2.80, −2.69]	−2.64	[−2.75, −2.54]	−1.28	[−1.32, −1.22]	−2.81	[−2.93, −2.68]	0.56	[0.26, 0.83]	−0.35	[−0.46, −0.17]
SQUARE	−2.45	[−2.51, −2.41]	−3.80	[−4.21, −3.38]	−0.56	[−0.65, −0.52]	−2.27	[−2.37, −2.16]	0.53	[0.31, 0.73]	0.07	[−0.17, 0.22]
CUBIC	−2.98	[−3.04, −2.94]	−6.03	[−6.15, −5.92]	−2.41	[−2.47, −2.33]	−2.65	[−2.70, −2.59]	−0.77	[−0.93, −0.64]	−1.42	[−1.63, −1.15]
SINE 1	0.10	[0.04, 0.18]	−0.06	[−0.13, 0.01]	0.09	[0.07, 0.13]	0.09	[0.04, 0.16]	1.97	[1.71, 2.28]	1.65	[1.45, 1.84]
SINE 2	−1.35	[−1.38, −1.32]	−1.48	[−1.52, −1.42]	−1.03	[−1.06, −1.00]	−0.92	[−0.96, −0.88]	1.21	[0.95, 1.65]	0.45	[0.26, 0.73]
SINE 3	−1.55	[−1.61, −1.47]	−1.69	[−1.85, −1.55]	−1.07	[−1.12, −1.01]	−1.66	[−1.75, −1.58]	0.26	[0.10, 0.40]	−0.02	[−0.21, 0.21]
FORRESTER	−1.98	[−2.03, −1.91]	−2.68	[−2.86, −2.49]	−1.10	[−1.19, −1.03]	−1.75	[−1.79, −1.69]	−0.11	[−0.23, 0.17]	−0.50	[−0.65, −0.28]
LEVY	−1.12	[−1.14, −1.07]	−0.93	[−0.98, −0.90]	−0.76	[−0.79, −0.73]	−1.04	[−1.08, −0.97]	1.35	[0.99, 1.53]	0.74	[0.53, 0.96]

that gave the best training loss during the training which are not necessarily the ones after the last gradient descent step.

Deep Ensembles Setup We consider the proposed number of five ensembles (Lakshminarayanan et al., 2017) with a single output $\hat{\mu}^\theta$ only (accounting for zero data noise), each with three hidden layers consisting of $2^8, 2^{10}$ and 2^9 nodes (resulting in ≈ 4 million parameters). We train them on standard regularized mean squared error (MSE) with regularization parameter $\lambda = 10^{-8}/n^{\text{train}}$ chosen to represent the same data noise assumptions as NOMU.

MC Dropout Setup The MC dropout network is set up with three hidden layers as well, with $2^{10}, 2^{11}$ and 2^{10} nodes (resulting in ≈ 4 million parameters). Both training and prediction of this model is performed with constant dropout probability $p := p_i = 0.2$, proposed in (Gal & Ghahramani, 2016). We perform 100 stochastic forward passes. To represent the same data noise assumptions as NOMU and deep ensembles we set the regularization parameter to $\lambda = (1-p) \cdot 10^{-8}/n^{\text{train}} = (1-0.2) \cdot 10^{-8}/n^{\text{train}}$ (based on equation (7) from (Gal & Ghahramani, 2016)).

Gaussian Process Setup Finally, we compare to a Gaussian process with RBF-kernel,

$$k_\pi(x, x') := \kappa \cdot e^{-\left(\frac{\|x-x'\|_2^2}{h^2}\right)}, \tag{52}$$

with hyperparameters $\pi := (\kappa, h)$, where both the prior variance parameter κ and the length scale parameter h (initialized as $\kappa = h = 1$) are optimized in the default range $[10^{-5}, 10^5]$ using 10 restarts and the data noise level is set to 10^{-7},[22].[23]

Hyper Deep Ensembles Setup We consider the same network architectures as for DE with a single output $\hat{\mu}^\theta$ only (accounting for zero data noise), each with three hidden layers consisting of $2^8, 2^{10}$ and 2^9 nodes (resulting in ≈ 4 million parameters). We train them on standard regularized mean squared error (MSE).

Furthermore, we use the following proposed parameter values from Wenzel et al. (2020b) top build the ensembles: we consider $M := 5$ networks (termed K in (Wenzel et al., 2020b)) in the final ensemble and search over $\kappa = 50$ networks resulting from `random_search`. Moreover, as random hyperparameters we use the proposed *L2 regulariza-*

[22] This is mainly used for numerical stability.

[23] We use *GaussianProcessRegressor* from SCIKIT-LEARN.

Table 7. Aggregate results for NOMU, GP, MCDO, DE, HDE and HDE* (our HDE see Appendix B.2.2) for the set of all ten 1D synthetic functions. Shown are the medians and a 95% bootstrap CI of AUC and NLL_{min} (without constant $\ln(2\pi)/2$) across all runs. Winners are marked in grey.

METHOD	AUC↓	95%-CI	NLL_{MIN}↓	95%-CI
NOMU	0.21	[0.21, 0.22]	-1.76	[-1.81, -1.72]
GP	0.33	[0.32, 0.35]	-1.74	[-1.81, -1.69]
MCDO	0.30	[0.29, 0.30]	-1.02	[-1.05, -0.99]
DE	0.22	[0.21, 0.23]	-1.75	[-1.79, -1.70]
HDE	2.04	[1.90, 2.20]	0.59	[0.51, 0.66]
HDE*	1.73	[1.63, 1.84]	0.04	[-0.03, 0.10]

Table 8. Aggregate results for NOMU, GP, MCDO, DE, HDE and HDE* (our HDE see Appendix B.2.2) for the set of all eleven 2D synthetic functions. Shown are the medians and a 95% bootstrap CI of of AUC and NLL_{min} (without constant $\ln(2\pi)/2$) across all runs. Winners are marked in grey.

METHOD	AUC↓	95%-CI	NLL_{MIN}↓	95%-CI
NOMU	0.42	[0.41, 0.42]	-0.99	[-1.01, -0.96]
GP	0.36	[0.36, 0.37]	-1.07	[-1.09, -1.05]
MCDO	0.45	[0.44, 0.45]	-0.61	[-0.62, -0.60]
DE	0.42	[0.41, 0.43]	-0.95	[-0.97, -0.93]
HDE	5.06	[4.75, 5.49]	1.74	[1.66, 1.84]
HDE*	6.33	[5.87, 6.75]	1.58	[1.51, 1.67]

tion and the `dropout_rate`, which we draw as in (Wenzel et al., 2020b) log-uniform from $(0.001, 0.9)$ and $(10^{-8} \cdot 10^{-3}, 10^{-8} \cdot 10^{3})$, respectively. Furthermore, we use the proposed train/validation split of 80%/20% and as score $\mathcal{S}$ the NLLFinally, as for MCDO and DE we scale the realized L2 regularization by $(1-\texttt{dropout\_rate})/\texttt{floor}(0.8 \cdot n^{\text{train}})$ chosen to represent the same data noise assumptions as NOMU.

For our HDE version termed **HDE***, we use a train/validation split of 70%/30%, draw the dropout probability from $(0.001, 0.5)$ and continue training the final ensemble again on *all* n^{train} training points (rescaling the realized L2 regularization again by $\texttt{floor}(0.8 \cdot n^{\text{train}})/n^{\text{train}}$).

B.2.3. Toy Regression: Detailed Results

Results 1D In Table 5 (AUC) and Table 6 (NLL_{min}), we present detailed results, which correspond to the presented ranks from Table 1 in the main paper. In Table 7, we present aggregate results, i.e., median values (incl. a 95% bootstrap confidence interval (CI)) of AUC and NLL_{min} across *all* 5000 runs (500 runs for 10 test functions) for each algorithm. We see that NOMU performs well on both metrics and yields numbers similar to those of DE. The GP is only competitive on NLL_{min}. MCDO performs in both metrics worse than NOMU and DE. Not surprisingly, HDE, shown to be the state-of-the-art ensemble algorithm (Wenzel et al., 2020b), fails to produce reliable model uncertainty estimates in such a scarce and noiseless regression setting (see Appendix B.2.4 for a discussion).

Results 2D Test functions for 2D input are taken from the same library as in the 1D setting.[24] Specifically, we select the following 11 different test functions: *Sphere*, *G-Function*, *Goldstein-Price*, *Levy*, *Bukin N.6*, *Rosenbrock*, *Beale*, *Camel*, *Perm*, *Branin*, *Styblisnki-Tang*.

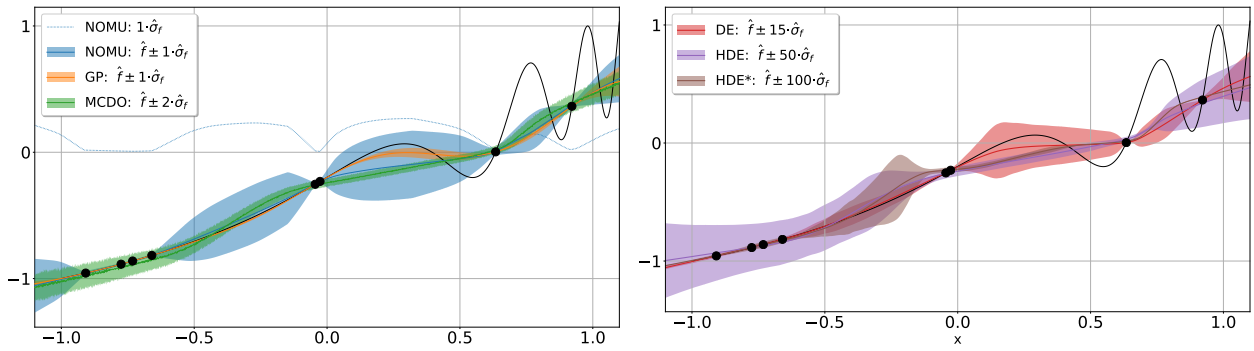
In Table 8, we present median values of AUC and NLL_{min} across *all* 5500 runs (500 runs for 11 test functions) for each algorithm. We also provide a 95% bootstrap confidence interval (CI) to assess statistical significance. We observe a similar ranking of the different algorithms as in 1D, i.e., NOMU and DE are ranked first in AUC; and the GP is ranked first in NLL_{min} followed by NOMU and DE, who share the second rank.

[24] sfu.ca/ ssurjano/optimization.html All test functions are scaled to $X = [-1, 1]^2$ and $f(X) = [-1, 1]$.

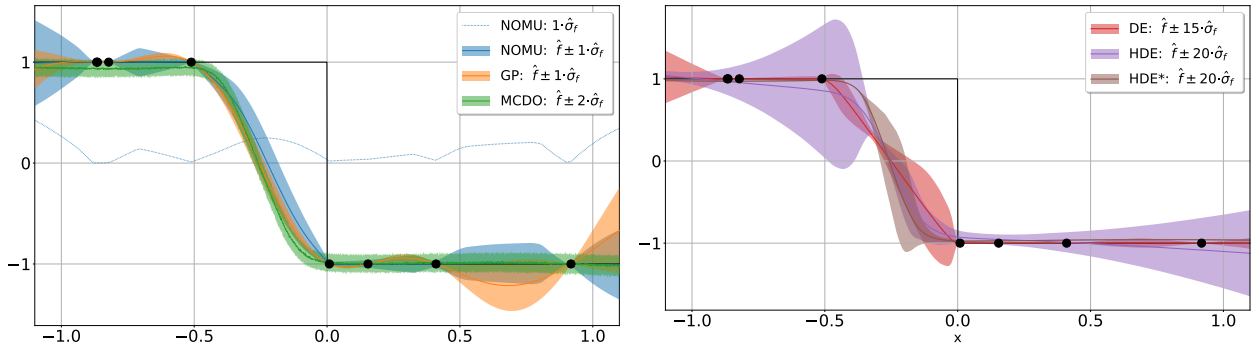
B.2.4. Toy Regression: Detailed Characteristics of Uncertainty Bounds

UB Characteristics in 1D Within the one-dimensional setting, characteristics of NOMU UBs can be nicely compared to those of the benchmark approaches. Figure 9 exemplarily shows typical outcomes of the UBs on a number of selected test functions as obtained in one run.

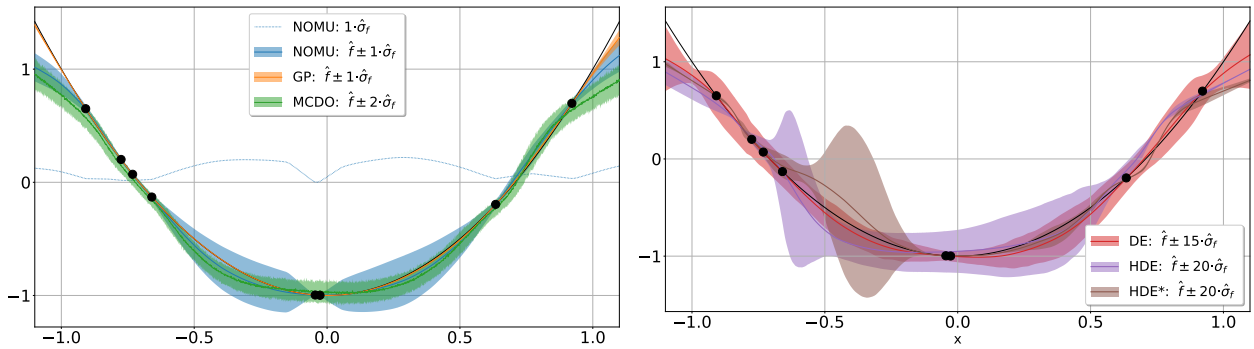
- **Deep Ensembles UBs:** Throughout the experiment, we observe that deep ensembles, while sometimes nicely capturing increased uncertainty in regions of high second derivative (e.g. around the kink in Figure 9e; pursuant in some form desiderata D4), still at times leads to UBs of somewhat arbitrary shapes. This can be seen most prominently in Figure 9b around $x \approx -0.25$, in Figure 9f around $x \approx -0.8$ or in Figure 9d around $x \approx 0.2$. Moreover, deep ensembles' UBs are very different in width, with no clear justification. This can be seen in Figure 9b when comparing the UBs for $x \geq 0$ against the UBs for $x \leq 0$ and at the edges of the input range in Figures 9a and 9f In addition, we frequently observe that deep ensembles UBs do not get narrow at training points, as for instance depicted in Figure 9c for $x < -0.5$, in Figure 9f for $x > 0.2$, or in Figure 9d for $x \in [-1, -0.7]$ and thus are not able to handle well small or zero data noise.
- **Hyper Deep Ensembles UBs:** Throughout our experiments, HDE produces less accurate UBs, which vary more from one run to another, than DE (recall that here we consider the setting of noiseless scarce regression). Possible reasons for this are:
  1. The scarcity of training/validation data. HDE trains



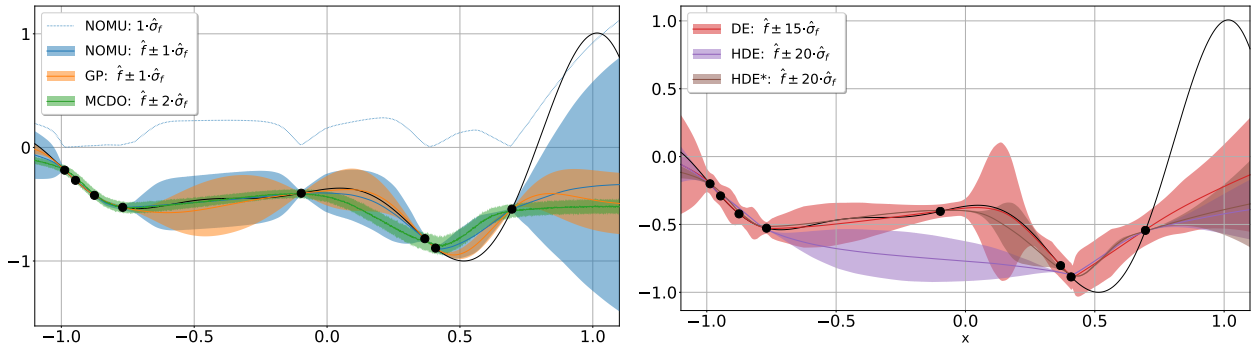
(a) Sine 3



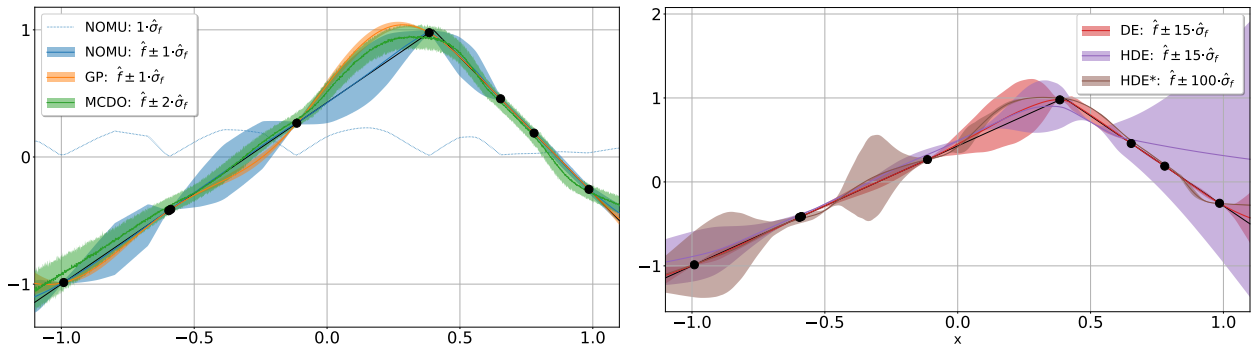
(b) Step



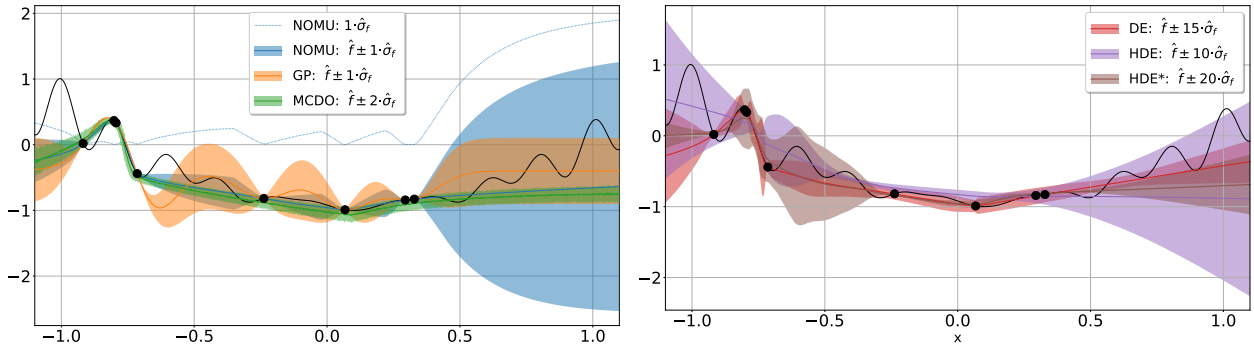
(c) Squared (GP would not behave reasonable for standard hyper-parameters in this instance, so we changed the range for the prior variance parameter κ to be $[10^{-5}, 4]$.)



(d) Forrester



(e) Kink



(f) Levy

Figure 9. Visualisations of resulting UBs for NOMU, GP, MCDO, DE, HDE and HDE* for a selection of test functions.

its NNs based on 80% of the training points and uses the remaining 20% to build an ensemble based on a score, whilst the other methods can use 100% of the training points for training. In a scarce data setting this implies that first, the mean prediction of HDE does not fit through all the training points, and second, the scoring rule is less reliable.

2. In a noiseless setting one already knows that the L2 regularization should be small, and thus optimizing this parameter is less useful here.

Therefore, we believe HDE is less well suited in a noiseless scarce regression setting to reliably capture model uncertainty. This manifests in Figure 9. HDE's uncertainty bounds often do not narrow at training points (Figures 9a–9c and 9e), while they suggest unreasonably overconfident mean predictions in some regions (Figures 9b and 9d). While our HDE* (see Appendix B.2.2 Hyper Deep Ensembles) manages to better narrow at training data, it tends to more frequently result in unnecessarily narrow bounds (Figures 9a, 9c and 9d). Overall, both HDE and HDE* vary a lot from run to run and thus tend to yield bounds of seemingly arbitrary shapes, where in each run the desiderata are captured in different regions.

- **MC Dropout UBs:** MCDO consistently yields tube-like UBs that do not narrow at training points. Interestingly, we remark that throughout the experiment MCDO samples frequently exhibit stepfunction-like shapes (see e.g. Figure 9d at $x \approx 0.5$ or Figure 9b for $x \in [-0.5, 0]$). This effect intensifies with increasing training time.
- **NOMU UBs:** In contrast, NOMU displays the behaviour it is designed to show. Its UBs nicely tighten at training points and expand in between and thus NOMU fulfills D1–D3 across all considered synthetic test functions.
- **Gaussian Process UBs:** The quality of the RBF-Gaussian process' UBs (as judged in this simulated setting) naturally varies with the true underlying function. While the UBs nicely capture parts of the true function with low widths in Figures 9c and 9d they have a hard time accurately enclosing true functions that are not as conformant with the prior belief induced by the choice of the RBF kernel (e.g. Figures 9b and 9e). Nonetheless, we also observe instances in which the training points are misleading to the GP's mean predictions even when considering ground truth functions for which this choice of kernel is very suitable. This manifests in over-confident mean predictions far away from the data generating true function (Figure 9a) or over-swinging behavior of the fitted mean (Figure 9f). It is true that one could find better

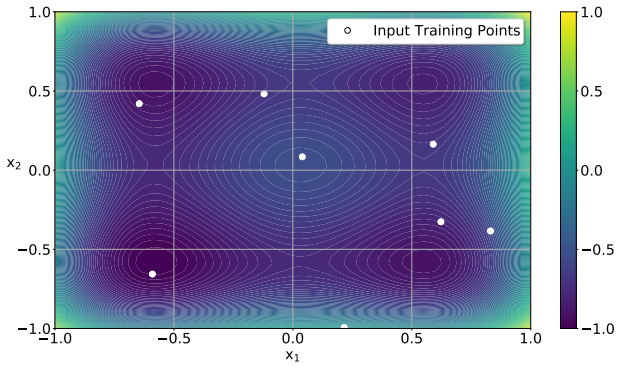


Figure 10. Contour plot of the 2D Styblinski test function.

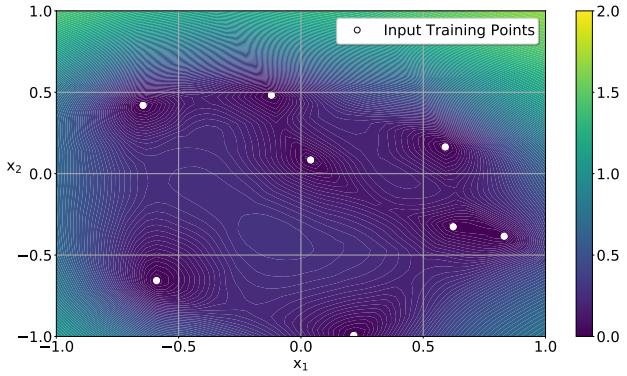
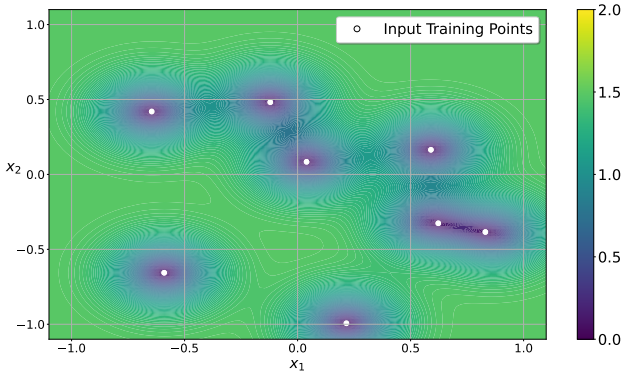


Figure 11. Estimated model uncertainty $\hat{\sigma}_f$ of NOMU for a single run of the Styblinski test function.

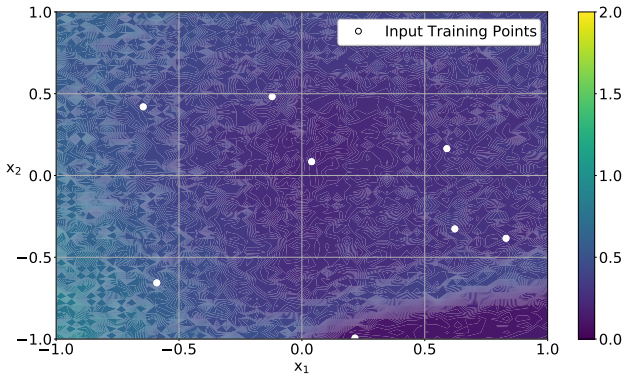
function-specific kernels. However, the RBF kernel is a good generic choice when evaluated across many different test functions without any specific prior information, which is why we choose this kernel for our comparison.

UB Characteristics in 2D To visualize the UBs in 2D, we select as the ground truth the *Styblinski* test function depicted in Figure 10. In Figure 11 (NOMU) and Figure 12 (benchmarks), we visualize the estimated model uncertainty as obtained in one run for this *Styblinski* test function.

- **NOMU (Figure 11):** As in 1D, we observe that NOMU's UBs nicely tighten at input training points and expand in-between, with overall steady shapes. Specifically, NOMU's UBs are larger for extrapolation, e.g., $[0, 1] \times [0.5, 1]$, compared to regions which are *surrounded* by input training data points, e.g., $[-0.25, 0.25] \times [-0.25, -0.75]$, even though the distance to the closest input training point is approximately the same. Thus, NOMU's UBs do not only depend on the distance to the closest training point but rather on the arrangement of all surrounding training points.
- **Benchmarks (Figure 12):** As in 1D, we see that the RBF-based GP's UBs do have the expected smooth and isotropic shape with zero model uncertainty at input training points. Moreover, as in 1D, MCDO's UBs exhibit a tubular shape of equal size ($\approx$ 0–0.25) across the whole input space. Whilst DE nicely captures zero model uncertainty at input training points, it again exhibits the somehow arbitrary behaviour in areas with few input training points. Both HDE and HDE* yield model uncertainty estimates that are small at input training points, except for HDE for one input training point $(x_1, x_2) \approx (0.6, 0.2)$, where the estimated model uncertainty is only small when continuing training on all training points (see HDE*). However, these estimates drastically fail to capture increased uncertainty in out-of-sample regions. For both algorithms, the model uncertainty estimate is unreasonably low in most regions of the input space and inexplicably high in a tiny fraction of the input space.

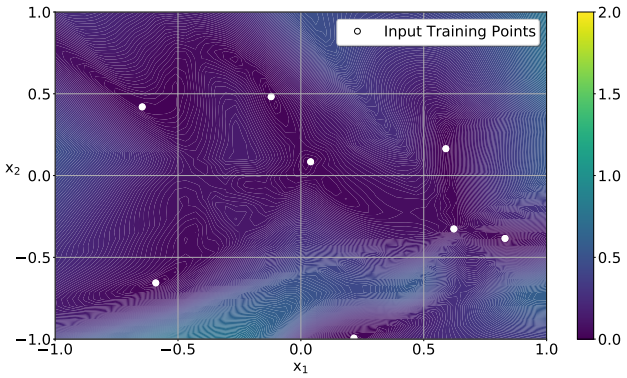


(a) GP with calibration constant $c = 1$.

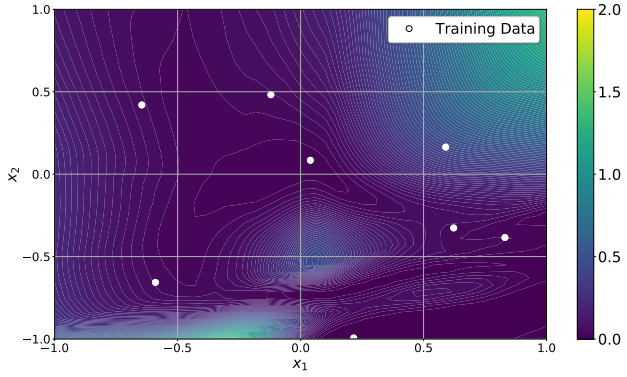


(b) MCDO with calibration constant $c = 10$.

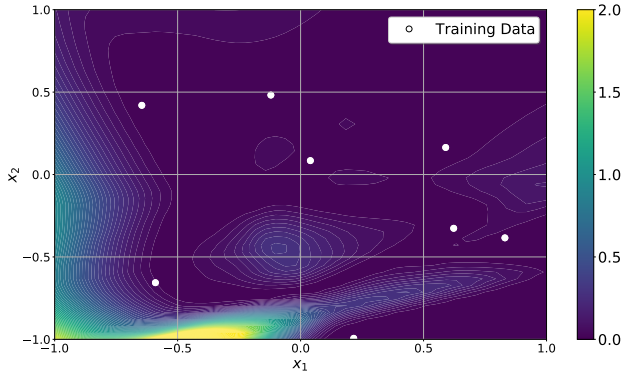
Figure 12. Estimated model uncertainty of Gaussian process (GP) and MC dropout (MCDO) for a single run of the Styblinski test function.



(c) DE with calibration constant $c = 15$.



(d) HDE with calibration constant $c = 30$.



(e) HDE* with calibration constant $c = 30$.

Figure 12. (cont.) Estimated model uncertainty of deep ensembles (DE), hyper deep ensembles (HDE) and HDE* for a single run of the Styblinski test function.

B.2.5. Generative Test-Bed: Details

In this section, we give a detailed description of the generative test-bed setting and provide further results for up to 20 dimensional input.

Detailed Setting

1. We sample a test-function (i.e. the ground truth function) $f : [-1, 1]^d \to \mathbb{R}$ from a BNN with i.i.d Gaussian weights. This means we sample i.i.d random parameters $\theta_i \sim \mathcal{N}(0, \sigma_{\text{prior}})$ and set $f = \mathcal{NN}_\theta$. The BNN is set up with three hidden layers with $2^{10}, 2^{11}$ and 2^{10} nodes, respectively. We choose σ_{prior} such that $\mathbb{E}_\theta \left[\sqrt{\mathbb{V}_x[\mathcal{NN}_\theta(x) \mid \theta]} \right] \approx 1$.[25] Resulting values for σ_{prior} are shown in Table 9.

Table 9. σ_{prior} depending on the input dimension d.

d	σ_{PRIOR}
1	0.114
2	0.102
5	0.092
10	0.084
20	0.070

2. We sample $n^{\text{train}} = 8 \cdot d$ input training points uniformly at random from the input space $[-1, 1]^d$ for Tables 2 and 10. In the case of Table 11, we sample $n^{\text{train}} = 8 \cdot d$ input training points from d-dimensional centered normal distribution with a covariance with $\min(5, d)$ eigenvalues of value 0.15 and the remaining eigenvalues only have the value 0.001. In both cases we get for input training point x_i^{train} the corresponding noiseless output-training point $y_i^{\text{train}} = f(x_i^{\text{train}}) = \mathcal{NN}_\theta(x_i^{\text{train}})$. And in both cases we create $n^{\text{test}} = 100 \cdot d$ test-data points from the same distribution as the training data points.[26]
3. We calculate UBs of all considered algorithms including NOMU. For this, we use for all algorithms the same configuration as in the 1D and 2D toy regression settings (see Section 4.1.1 and Appendix B.2.2) except for NOMU, where we set $\ell_{\text{min}} = 0.1$, $\ell_{\text{max}} = 1$ and use $l = 100 \cdot d$ artificial data points, where d denotes the respective dimension.
4. We measure the quality of the UBs via NLL for a fine grid of different values for the calibration parameter c.

We repeat these four steps $m_{\text{seeds}} = 200$-times. Then we choose for each method the value c where the average NLL is minimal. Importantly, the chosen calibration constant c *only* depends on the input-dimension d and on the algorithm but not on the randomly chosen test-function f. In Table 10 and Table 11 we report the mean and a 95% CI over these 200 runs for the chosen calibration constant c.

[25] $\mathbb{E}_\theta \left[\sqrt{\mathbb{V}_x[\mathcal{NN}_\theta(x) \mid \theta]} \right] \approx 1$ holds true for $x \sim \text{Unif}\left([-1, 1]^d\right)$. This might deviate (slightly) for Gaussian x.

[26] Note that we sample the artificial data points for NOMU always uniformly from $[-1, 1]^d$, since we assume that the low-dimensional manifold is often unknown in practice. NOMU could probably be improved if one tries to learn the distribution of the input data or if one already had some prior information about the distribution of the input data points. It is good to see how robustly NOMU can handle a different distribution of the input data points without the need to adapt the distribution of the artificial data points.

Table 10. Uniform data generation. Shown are average NLL (without constant $\ln(2\pi)/2$) and a 95% CI over 200 BNN samples.

FUNCTION	NOMU	GP	MCDO	DE	HDE	TUBE
BNN1D	-1.65±0.10	-1.08±0.22	-0.34±0.23	-0.38±0.36	8.47±1.00	-0.86±0.19
BNN2D	-1.16±0.05	-0.52±0.11	-0.33±0.13	-0.77±0.07	9.10±0.39	-0.81±0.06
BNN5D	-0.37±0.02	-0.33±0.02	-0.05±0.04	-0.13±0.03	8.41±1.00	-0.33±0.02
BNN10D	-0.09±0.01	-0.11±0.01	0.06±0.02	0.10±0.01	6.44±1.36	-0.12±0.01
BNN20D	0.15±0.01	0.06±0.01	0.18±0.01	0.30±0.01	3.58±0.81	0.07±0.01

Table 11. 5D-Gaussian data generation. Shown are average NLL (without constant $\ln(2\pi)/2$) and a 95% CI over 200 BNN samples.

FUNCTION	NOMU	GP	MCDO	DE	HDE	TUBE
BNN1D	-1.91±0.13	-1.13±0.27	-0.49±0.16	-0.27±0.54	8.95±1.18	-0.85±0.13
BNN2D	-1.55±0.05	-0.85±0.12	-0.62±0.13	-1.15±0.08	7.92±0.42	-0.95±0.11
BNN5D	-0.77±0.02	-0.72±0.02	-0.47±0.03	-0.55±0.03	9.68±1.08	-0.68±0.02
BNN10D	-1.36±0.01	-1.31±0.01	-1.08±0.03	-1.14±0.02	5.16±1.55	-1.25±0.01
BNN20D	-1.70±0.01	-1.72±0.01	-1.52±0.02	-1.53±0.01	0.99±0.39	-1.68±0.01

Discussion of the Results in Table 10 We see that for $d \leq 5$ NOMU significantly outperforms all other considered benchmarks. For dimensions $d \geq 10$ it gets harder to find a good metric for measuring the quality of the uncertainty bounds: For high dimensions almost no test data point sampled uniformly from $[-1, 1]^d$ is close to the training data points, so they all have quite high uncertainty. We further verified this for dimensions $d = 10$ and $d = 20$ by introducing another algorithm we call TUBE. TUBE uses NOMU's mean prediction and assigns the same (calibrated) constant uncertainty c to all test data points. As we can see in Table 10, TUBE is ranked first (en par with GP) in $d = 10$ and $d = 20$, highlighting that the metric NLL is flawed in this setting for dimensions $d \geq 10$. However, for dimensions $d \leq 5$ the trivial TUBE algorithm is significantly outperformed by other methods. Thus, the NLL-metric for dimensions $d \geq 10$ and uniformly distributed data on $[-1, 1]^d$ mainly measures if D3 (Out-of-Sample) is reliably fulfilled. D3 (Out-of-Sample) is particularly reliably fulfilled for TUBE and for GP. For BO however, especially D2 (In-Sample) is important to not get stuck in local optima, because in BO the test data-points are not uniformly distributed, and the predicted uncertainty close to the training points is particularly important. Therefore, in this setting, we only include results for input dimensions 1-5D (see Table 2) in the main part of the paper.

Discussion of the Results in Table 11 Another approach to circumvent the problem that the NLL metric lose their relevance for scenarios where the input test data points are all far away from the input training data points is to use a different distribution for X. The well known manifold-hypothesis (Cayton, 2005) states that in typical real world data-sets relevant input data points lie close to a low-dimensional manifold. Therefore, we run a further experiment where we sample from our training and test data from a Gaussian distribution that concentrates mainly close to a $\min(d, 5)$-dimensional flat manifold as described in Item 2 and report the results in Table 11. For this experiment we see that dimension $d \geq 10$ are interesting too, since TUBE is significantly beaten by NOMU and GP. In Table 11 one can see that NOMU is among the best methods for all considered dimensions.

Our Generative Test-Bed vs. Osband et al. (2021) Osband et al. (2021) measures the Kullback–Leibler divergence between the posterior of any method to the posterior of shallow GP with a fixed (deep) neural tangent kernel (NTK). Thus, they evaluate methods by their similarity to a NTK-GP posterior. Because of the fixed kernel the posterior of the NTK GP does not fulfill D4 (Metric Learning) at all. This implies that their evaluation metric does not reward D4 (Metric Learning) at all. However, we think that D4 (Metric Learning) is very important for real-world deep learning applications (see Footnote 4 and Appendix D.4). The posterior of a finite-width BNN fulfills D4 (Metric Learning). Therefore, approximating such a posterior is more desirable. However, in contrast to the NTK-GP posterior in (Osband et al., 2021), there is no closed formula for a finite-width BNN posterior. Thus, at first sight one cannot straightforwardly evaluate methods based on the Kullback–Leibler divergence between their posterior and this intractable BNN posterior as in Osband et al. (2021).

Nonetheless, in the following, we proof in Theorem B.7 that the metric we use in Tables 2, 10 and 11 indeed converges (up to a constant)[27] to the average Kullback–Leibler divergence $\bar{d}_{\text{KL}}$ to the posterior of a finite width BNN as m_{seeds} tends to infinity. First, we define the average Kullback–Leibler divergence $\bar{d}_{\text{KL}}$ and introduce some notation.

Definition B.6 (AVG KULLBACK–LEIBLER DIVERGENCE) *Let D^{train} be a finite set of training points and consider a prior distribution*[28] *$\mathbb{P}[f \in \cdot]$ on the function space $\{f : X \to Y\}$ and the corresponding posterior $\mathbb{P}[f \in \cdot \mid D^{train}]$ on the function space. Then the marginal of the posterior $\mathbb{P}[f(x) \in \cdot \mid D^{train}, x]$ is a measure on $\mathbb{R}$ for every given input data point $x \in X$ and every given training data set D^{train}. Let $\mathbb{Q}[\cdot \mid D^{train}, x]$ also be a measure on $\mathbb{R}$ for every given input data point $x \in X$ and every given training data set D^{train}.*[29] *The average Kullback–Leibler divergence is then defined as*

$$\bar{d}_{KL} = \mathbb{E}_{D^{train},x}\left[d_{KL}\left(\mathbb{P}[f(x) \in \cdot \mid D^{train}, x] \;||\; \mathbb{Q}[\cdot \mid D^{train}, x]\right)\right],$$

[27] If we are just interested in the relative performance of different methods compared to each other, the constant does not matter.

[28] In our generative test-bed the prior distribution is a BNN prior with i.i.d centered Gaussian weights.

[29] In our context $\mathbb{Q}[\cdot \mid D^{\text{train}}, x]$ can be the approximation of the marginal posterior at $x \in X$ given training data D^{train} obtained from any method such as NOMU, GP, MCDO, DE, HDE etc.

where the expectation is taken over x and D^{train} according to $\mathbb{P}$, and d_{KL} is the classical Kullback–Leibler divergence between two probability measures on $\mathbb{R}$. This is equivalent to Equation (1) from (Osband et al., 2021).

Theorem B.7 *Using the notation from Definition B.6, let $q(\cdot \mid D^{train}, x)$ be the density of $\mathbb{Q}[\cdot \mid D^{train}, x]$ on $\mathbb{R}$. Let $(f_j(x_j), D_j^{train}, x_j)_{j\in\{1,\dots,m_{seeds}\}}$ be i.i.d samples of $\mathbb{P}[(f(x), D^{train}, x) \in \cdot]$.*[30] *Then, the average NLL*

$$\lim_{m_{seeds}\to\infty} \frac{1}{m_{seeds}} \sum_{j=1}^{m_{seeds}} -\log\left(q(f_j(x_j) \mid D_j^{train}, x_j)\right) = \bar{d}_{KL} + C_{\mathbb{P}}$$

converges in probability to $\bar{d}_{KL}$ up to an additive constant $C_{\mathbb{P}}$, where $C_{\mathbb{P}}$ only[31] *depends on $\mathbb{P}$ and not on $\mathbb{Q}$.*

Proof. Let p be the density of $\mathbb{P}[(f(x), D^{\text{train}}, x) \in \cdot]$ and $p(\cdot \mid D^{\text{train}}, x)$ the density on $\mathbb{R}$ of the marginal posterior $\mathbb{P}[f(x) \in \cdot \mid D^{\text{train}}, x]$. Further we write $p(D^{\text{train}}, x)$ for the density of of the marginal $\mathbb{P}[(D^{\text{train}}, x) \in \cdot]$ evaluated at (D^{train}, x). Since

$$\lim_{m_{\text{seeds}}\to\infty} \frac{1}{m_{\text{seeds}}} \sum_{j=1}^{m_{\text{seeds}}} -\log\left(q(f_j(x_j) \mid D_j^{\text{train}}, x_j)\right)$$

can be seen as a Monte-Carlo approximation, it converges in probability to

$$\begin{aligned}
&\mathbb{E}_{f(x),D^{\text{train}},x}[-\log\left(q(f(x) \mid D^{\text{train}}, x)\right)] = \\
&= \int -\log\left(q(f(x) \mid D^{\text{train}}, x)\right) p(f(x), D^{\text{train}}, x) d(f(x), D^{\text{train}}, x) = \\
&= \int -\log\left(q(f(x) \mid D^{\text{train}}, x)\right) p(f(x) \mid D^{\text{train}}, x) p(D^{\text{train}}, x) d(f(x), D^{\text{train}}, x).
\end{aligned}$$

By Fubini this is equal to

$$\begin{aligned}
&\int\int -\log\left(q(f(x) \mid D^{\text{train}}, x)\right) p(f(x) \mid D^{\text{train}}, x) d(f(x)) p(D^{\text{train}}, x) d(D^{\text{train}}, x) = \\
&= \mathbb{E}_{D^{\text{train}},x}\Big[\underbrace{\int -\log\left(q(f(x) \mid D^{\text{train}}, x)\right) p(f(x) \mid D^{\text{train}}, x) d(f(x))}_{H(\mathbb{P}[f(x)\in\cdot \mid D^{\text{train}},x], \mathbb{Q}[\cdot \mid D^{\text{train}},x])}\Big],
\end{aligned}$$

where H is the cross-entropy. So this is further equal to

$$\begin{aligned}
&\mathbb{E}_{D^{\text{train}},x}\left[d_{\text{KL}}\left(\mathbb{P}[f(x) \in \cdot \mid D^{\text{train}}, x] \;||\; \mathbb{Q}[\cdot \mid D^{\text{train}}, x]\right)\right] + \\
&\mathbb{E}_{D^{\text{train}},x}\left[H\left(\mathbb{P}[f(x) \in \cdot \mid D^{\text{train}}, x]\right)\right] = \\
&= \bar{d}_{\text{KL}} + C_{\mathbb{P}},
\end{aligned}$$

where $C_{\mathbb{P}} = \mathbb{E}_{D^{\text{train}},x}\left[H\left(\mathbb{P}[f(x) \in \cdot \mid D^{\text{train}}, x]\right)\right]$ only depends on $\mathbb{P}$ and not on $\mathbb{Q}$ or q. □

To apply Theorem B.7 to the metrics reported in Tables 2, 10 and 11 one has to apply Footnote 30 and has to set $q(\cdot \mid D^{\text{train}}, x)$ to be the density corresponding to the already calibrated uncertainty at x obtained from any method trained on D^{train}, i.e., for example let c be a fitted calibration parameter, and $\hat{f}$ and $\hat{\sigma}$ be the fitted model and model uncertainty prediction of NOMU then $q(\cdot \mid D^{\text{train}}, x) := \mathcal{N}(\cdot; \hat{f}(x), c \cdot \hat{\sigma}(x))$ for $x \in X$.[32] In our setting, we made sure that the correct calibration constant c is chosen in the limit $m_{\text{seeds}} \to \infty$, since we chose one fixed value for c per dimension and method and do not over-fit on specific seeds.

B.2.6. Solar Irradiance Time Series: Details

Configuration Details We use for all algorithms the same configuration as in the 1D and 2D toy regression setting (see Section 4.1.1 and Appendix B.2.2) except that we now train all algorithms for 2^{14} epochs and set as L2-regularization $\lambda = 10^{-19}$ for NOMU, $\lambda = 10^{-19}/n^{\text{train}}$ for DE, and $\lambda = (1-0.2)\cdot 10^{-19}/n^{\text{train}}$ for MCDO. For HDE we accordingly draw the L2 regularization parameter log-uniform from $(10^{-19}\cdot 10^{-3}, 10^{-19}\cdot 10^{+3})$ (which affects the L2-regularization parameter of HDE* in the same way as in Appendix B.2.2).

Results Figure 13 visualizes the UBs from NOMU and the benchmark algorithms. As becomes evident, NOMU manages to best fit the training data[33], while keeping model uncertainty in-between training points. The corresponding metrics for this specific run are given in Table 12.

B.2.7. UCI Data Sets: Details

To get a feeling for how well NOMU (without data noise extensions or hyperparameter tuning) performs in noisy regression with high-dimensional input, we test our algo-

[30] We formulate the theorem for the hardest case of $n^{\text{test}} = 1$. The convergence in probability of

$$\lim_{m_{\text{seeds}}\to\infty} \frac{1}{m_{\text{seeds}}} \sum_{j=1}^{m_{\text{seeds}}} \frac{1}{n^{\text{test}}} \sum_{i=1}^{n^{\text{test}}} -\log\left(q(f_j(x_{j,i}) \mid D_j^{\text{train}}, x_{j,i})\right) = \bar{d}_{\text{KL}} + C_{\mathbb{P}}$$

is obviously even faster for $n^{\text{test}} > 1$. For the proof it is only important that $m_{\text{seeds}} \to \infty$. If one formulates the theorem for general $n^{\text{test}} > 1$, one has to formulate that $(D_j^{\text{train}})_{j\in\{1,\dots,m_{\text{seeds}}\}}$ are i.i.d, but $(f_j(x_{j,i}), x_{j,i})_{i\in\{1,\dots,n^{\text{test}}\}}$ are only conditionally independent conditioned on D_j^{train}, which would make the presentation unnecessarily technical.

[31] $\mathbb{P}$ does not depend on the chosen method. Only $\mathbb{Q}$ (and accordingly q) differs among the methods. Thus, $C_{\mathbb{P}}$ does not change the ranking amongst different methods.

[32] In theory, more general posteriors than Gaussians could be used too, but within this paper we always assumed Gaussian marginals of the posterior as all the considered benchmarks also output Gaussian distributed approximations of the marginal posteriors.

[33] (Gal & Ghahramani, 2016) consider the same data set to compare UBs of MC dropout and GPs. In this work however, NNs are trained for 10^6 epochs possibly explaining why MC dropout more nicely fits the training data in their case.

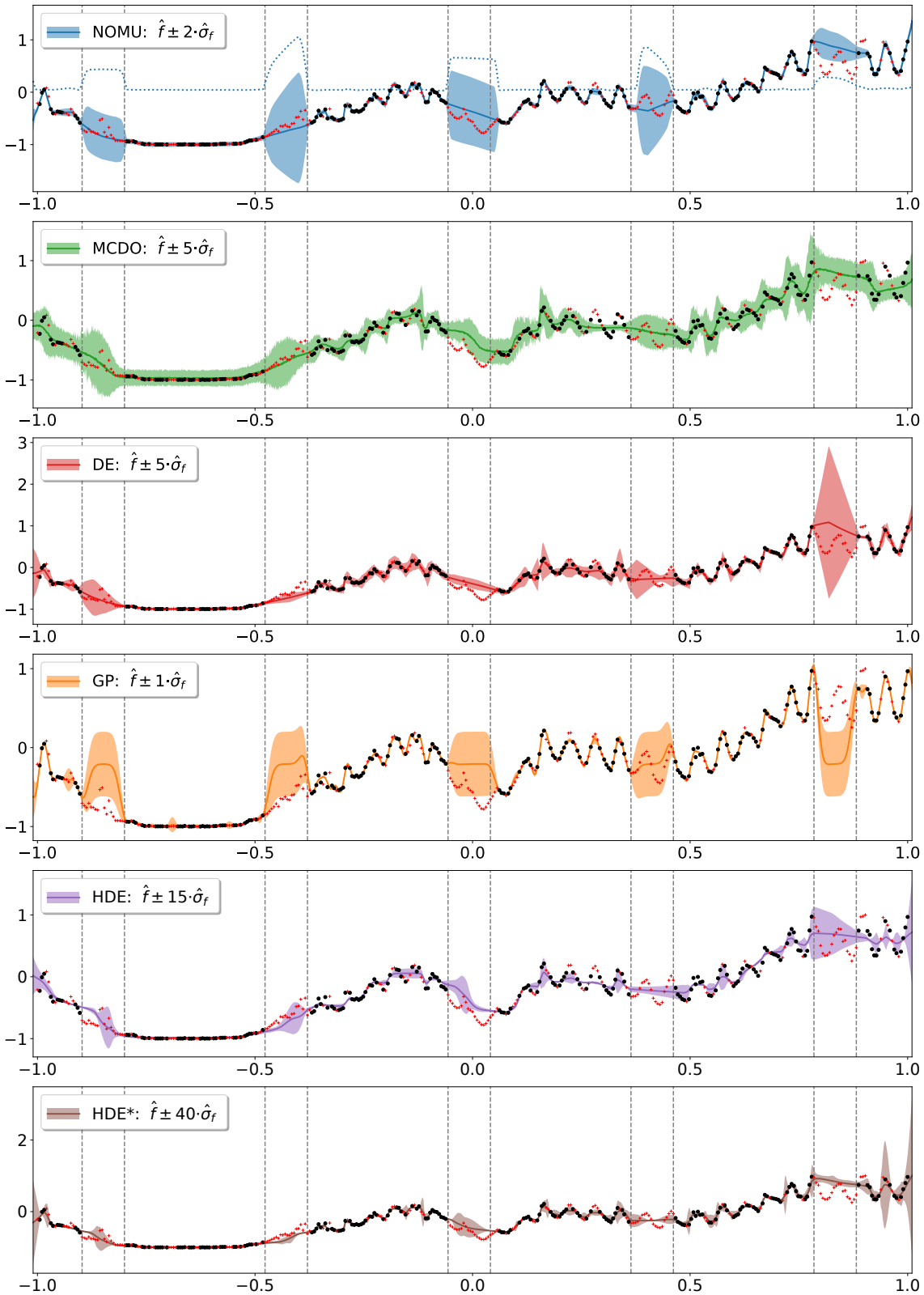


Figure 13. Visualization of each algorithm's model prediction (solid lines) and UBs (shaded areas) on the solar irradiance data set. Training and test points are shown as black dots and red crosses, respectively. We present UBs obtained by NOMU (c=2) compared to the benchmarks MC dropout (MCDO) (c=5), deep ensembles (DE) (c=5), Gaussian process (GP) (c=1) and hyper deep ensembles (HDE) (c=15) and (HDE*) (c=40). Additionally, NOMU's estimated and scaled model uncertainty $2\hat{\sigma}_f$ is shown as a dotted line.

Table 12. Metrics for solar irradiance time series.

METHOD	AUC↓	NLL_{MIN}↓
NOMU	0.32	−1.44
GPR	0.46	−1.24
MCDO	0.30	−1.11
DE	0.35	−1.28
HDE	0.47	−0.98
HDE*	0.59	−1.08

rithm on the popular UCI benchmark (Hernández-Lobato & Adams, 2015) and the UCI gap data sets introduced in (Foong et al., 2019).

The UCI gap data sets were designed to capture whether an algorithm's uncertainty estimate increases in between separated clusters of observations where epistemic uncertainty should be higher compared to regions with many data points. This is also required by D3. However, in Foong et al. (2019) only those data points that had been removed to create gaps in the training data were used as test data for calculating NLL. Thus, NLL on UCI gap data fails by construction to capture the uncertainty quality outside of the gap: First, D2 is not properly measured by this evaluation, since there are not many test-points in the test data-set that are close to training points. Second, also D3 is not properly measured, since the smaller gaps in between data points outside of the gap are not part of the test data-set. Note that D3 also concerns these kinds of gaps. To provide better evaluation of uncertainty, future work should focus on mixed test data-sets (moving some input points outside of the gap from the training set to the test set) similar to our train/test split from the experimental setting in Appendix B.2.6.

Nonetheless, we test NOMU on the UCI and UCI gap data sets using the same experiment setup as in the respective works.

NOMU Setup In line with the literature, NOMU's main and side architectures were chosen as a single-hidden-layer NN with 50 hidden nodes (except for the large protein data set, where the number of hidden nodes was 100). NOMU was trained for 400 (UCI) respectively 40 (UCI gap) epochs, with L2-reg $1e^{-09}$ and $1e^{-04}$ on the main- and side architectures respectively, using a batch size of 100 in Adam stochastic gradient descent with learning rate 0.01. NOMU used $\ell = 100$ artificial data points randomly sampled on each batch and was refit on validation data after the constant c had been calibrated on these. For the standard UCI data sets, we used the same loss parameters as in the remaining regression experiments, namely $\pi_{\text{exp}} = 0.01$, $\pi_{\text{sqr}} = 0.1$, and $c_{\text{exp}} = 30$. For the UCI gap data set, where uncertainty is only evaluated in gap regions, we chose $\pi_{\text{exp}} = 0.1$ and $\pi_{\text{sqr}} = 0.01$, relaxing the requirement of small uncertainty at observed data points and strengthening the pull upward on our raw uncertainty output. Moreover, for numerical stability we use the following slight adaptation of the NOMU loss from Definition 3.3:

Table 13. UCI-GAP average NLL and a 95% normal-CI.

DATASET	NOMU	LL	NLM-HPO	NLM
BOSTON	2.80 ±0.11	2.79 ±0.09	2.81 ±0.15	3.60 ±0.21
CONCRETE	3.38 ±0.08	3.53 ±0.13	3.72 ±0.11	3.89 ±0.21
ENERGY	3.71 ±1.62	6.49 ±5.50	3.78 ±2.27	3.40 ±1.99
KIN8NM	-0.97 ±0.02	-1.14 ±0.03	-1.06 ±0.03	-1.04 ±0.03
NAVAL	31.81 ±28.0	15.66 ±8.34	1.17 ±1.56	1.48 ±1.64
CCPP	2.91 ±0.04	2.89 ±0.03	2.96 ±0.09	2.90 ±0.05
PROTEIN	3.21 ±0.10	3.09 ±0.05	3.22 ±0.09	3.35 ±0.11
WINE	0.99 ±0.02	0.96 ±0.01	0.98 ±0.01	1.75 ±0.07
YACHT	2.15 ±0.30	1.33 ±0.29	1.72 ±0.33	1.44 ±0.21

$$L^{\pi}_{\text{stable}}(\mathcal{NN}_\theta) := \underbrace{\sum_{i=1}^{n^{\text{train}}} (\hat{f}(x_i^{\text{train}}) - y_i^{\text{train}})^2}_{(a)} + \pi_{\text{sqr}} \cdot \underbrace{\sum_{i=1}^{n^{\text{train}}} \rho_{\text{stable}}\left(\hat{r}_f(x_i^{\text{train}})\right)}_{\text{stable version of } (b)} + \pi_{\text{exp}} \cdot \underbrace{\frac{1}{\lambda_d(X)} \int_X u_{\text{stable}}\left(\hat{r}_f(x)\right)\, dx}_{\text{stable version of } (c)}, \tag{53}$$

where

$$\rho_{\text{stable}}(r) = \begin{cases} r^2 & , |r| \le 1 \\ 2|r| - 1 & , |r| > 1 \end{cases} \tag{54}$$

is the Huber-loss and

$$u_{\text{stable}}(r) = \begin{cases} u(r) = e^{-c_{\text{exp}} \cdot r} & , r \ge 0 \\ -c_{\text{exp}} \cdot r & , r < 0. \end{cases} \tag{55}$$

This stable loss L^{π}_{stable} behaves exactly the same as the standard NOMU-loss L^{π} as long as the outputs of the NN stay in a reasonable range. Only if the gradient descent gets unstable, such that the NN outputs very extreme value, this stabilized loss assures a bounded gradient of the loss L^{π}_{stable} with respect to the outputs of the NN $\mathcal{NN}_\theta$.

Results UCI Table 3 in the main paper reports NLL on test data averaged over 20 runs (as in Hernández-Lobato & Adams (2015)) for the UCI data set. It includes the following benchmark models:

- NLM-HPO and NLM correspond to BN(BO)-1 NL in Section D.1 respectively Section D.2. of (Ober & Rasmussen, 2019),
- LL corresponds to LL 1HL TANH in (Foong et al., 2019),
- MCDO2 represents Dropout (Convergence) in (Mukhoti et al., 2018),
- MCDO corresponds to Dropout in (Gal & Ghahramani, 2016) and DE to Deep Ensembles in (Lakshminarayanan et al., 2017).

We used the standard deviations that were reported in these respective works to derive 95%-normal CIs.

Results UCI Gap Table 13 lists Gaussian 95%-confidence intervals for NLL on the UCI gap data sets for NOMU as well as the values reported for benchmarks NLM, NLM-HPO and LL (Linearized Laplace) as listed above. These values result from different runs where in each run the gap were introduced in a different input dimension (see Foong et al. (2019)). We see that NOMU performs comparably, and confirm the observation that LL tends to best capture uncertainty in the large gaps artificially introduced in the training data.

B.3. Bayesian Optimization

B.3.1. BO: Configuration Details

In the following, we describe the detailed hyperparameter setup of all algorithms.

NOMU Setup For NOMU, we set $\pi_{\text{sqr}} = 1$, $\ell_{\text{min}} = 1\text{e}{-6}$ and use $l = 500$ artificial input points for 5D, 10D and 20D. Otherwise we use the exact same hyperparameters as in 1D and 2D regression (see in Section 4.1 the paragraph **Algorithm setup**).

Deep Ensembles Setup For deep ensembles, we use the exact same hyperparameters as in 1D and 2D regression (see Appendix B.2.2).

MC Dropout Setup For MC dropout, due to computational constraints, we were only able to use 10 stochastic forward passes (instead of 100 in the regression setting) However, this still results in an increase in compute time by a factor 10 of every single acquisition function evaluation compared to NOMU. Otherwise, we use the exact same hyperparameters as in 1D and 2D regression (see Appendix B.2.2) .

Gaussian Process Setup For Gaussian processes (pGP and GP), we use the exact same setup as in 1D and 2D regression (see Appendix B.2.2).

Hyper Deep Ensembles Setup For hyper deep ensembles, due to computational constraints, we were only able to use $\kappa = 20$ (instead of $\kappa = 50$ in the regression setting). However, this still results in an increase of training time by a factor of 5 in each single BO step compared to NOMU. Otherwise we use the exact same hyperparameters as in the 1D and 2D regression setting (see Appendix B.2.2).

Acquisition Function Optimization For each algorithm and each BO step, we determine the next BO input point by maximizing the corresponding upper-bound acquisition function, i.e., $(\hat{f}(x) + c\,\hat{\sigma}_f(x))$. We maximize this upper-bound acquisition function using the established DIRECT (Dividing Rectangles) algorithm.

Gaussian BNN Test Functions Note that for the Gaussian BNN test functions in BO (see Table 4 and Table 14, respectively) we do not have access to the ground truth global maximum. Thus, to determine the final regret, we use the established DIRECT (Dividing Rectangles) algorithm. With DIRECT, we calculate a point x^{direct} s.t. $f(x^{\text{direct}}) \approx \max_{x \in X} f(x)$ and report $|f(x^{\text{direct}}) - \max_{i \in \{1,\dots,72\}} f(x_i)|/f(x^{\text{direct}})$ in Table 4 and Table 14, respectively.

B.3.2. BO: Calibration

MW Scaling In general, having a good understanding of the prior scale is very hard. Nevertheless, often it is easier to have at least some intuition in which order of magnitude the posterior UBs should be on average. When function values approximately range in $[-1, 1]$, it is reasonable to require that after observing 8 initial points the mean width (MW) of the UBs lies within the order of magnitudes 0.05 and 0.5. Hence our motivation for choosing the calibration parameter c accordingly. An advantage of such a calibration is that it can be applied to every method equally, whereas there is in general no clear notion of setting the prior scale of two different methods (e.g. MC dropout and deep ensembles) to the same value. Note, that we only use MW to calibrate c directly after mean and variance predictions were fit based on the 8 initial data points. So MW is not fixed when further data points are observed in subsequent steps of the BO.

Dynamic c The initial choice of c can still be corrected in each BO step. Certainly, in the noiseless case it does not make sense to pick an input point $x_{i'}$ that is identical to an already observed input point $x_i, i < i'$, where nothing new is to be learned. Therefore, we want our NN-agent to get "bored" if its acquisition function optimization would suggest to pick an input point $x_{i'} \approx x_i, i < i'$. The idea of Dynamic c is to encourage the agent, in case it was "bored", to become more "curious" to explore something new instead of picking a "boring" input point. This can be achieved by iteratively increasing c until the acquisition function optimization suggests an input point $x_{i'} \not\approx x_i, \forall i < i'$. We then only apply the expensive function evaluation for $x_{i'} \not\approx x_i, \forall i < i'$ that we obtained after increasing c enough to not "bore" the NN. However, towards the last BO-steps, if we already know approximately where a global optimum is and we only want to fix the last digits of the optimizer, we have to evaluate the function closer to already observed points. In contrast, a very "young" NN (i.e., a network that

Table 14. Results for 15 different BO tasks. Shown are the average final regrets and a 95% normal-CI per dimension and for each individual function over 100 (5D) and 50 (10D and 20D) runs for an initial mean width (MW) of 0.05 and 0.5, respectively. Winners are marked in grey.

FUNCTION	NOMU MW 0.05	NOMU MW 0.5	GP MW 0.05	GP MW 0.5	MCDO MW 0.05	MCDO MW 0.5	DE MW 0.05	DE MW 0.5	HDE MW 0.05	HDE MW 0.5	PGP MW 0.05	RAND MW
LEVY5D	2.12e−3 ± 5.95e−4	1.52e−3 ± 4.34e−4	1.10e−3 ± 1.68e−4	8.84e−3 ± 1.15e−3	1.98e−2 ± 3.59e−3	1.27e−2 ± 2.57e−3	7.09e−3 ± 3.76e−3	9.09e−2 ± 9.16e−3	4.48e−3 ± 1.25e−3	3.76e−3 ± 1.22e−3	6.25e−3 ± 5.26e−4	5.44e−2 ± 4.22e−3
ROSENBROCK5D	1.75e−4 ± 2.76e−5	3.57e−4 ± 6.39e−5	3.62e−4 ± 6.52e−5	8.44e−4 ± 1.62e−4	2.83e−4 ± 4.64e−5	7.35e−4 ± 1.08e−4	7.93e−4 ± 7.74e−4	5.96e−3 ± 1.78e−3	1.11e−3 ± 7.25e−4	5.02e−4 ± 1.04e−4	7.56e−4 ± 8.13e−5	4.73e−3 ± 7.04e−4
G-FUNCTION5D	1.12e−1 ± 1.99e−2	1.06e−1 ± 1.97e−2	1.68e−1 ± 1.62e−2	2.60e−1 ± 1.72e−2	2.66e−2 ± 8.57e−3	7.80e−2 ± 8.82e−3	1.88e−1 ± 2.08e−2	2.03e−1 ± 2.20e−2	1.35e−1 ± 2.46e−2	1.37e−1 ± 2.59e−2	1.78e−1 ± 1.22e−2	3.63e−1 ± 1.09e−2
PERM5D	4.06e−4 ± 1.70e−4	2.58e−4 ± 1.13e−4	7.76e−5 ± 3.19e−5	1.80e−3 ± 4.44e−4	7.79e−5 ± 4.19e−5	7.11e−5 ± 1.45e−5	6.96e−4 ± 2.20e−4	1.73e−3 ± 4.55e−4	2.74e−3 ± 5.20e−4	2.50e−3 ± 5.58e−4	2.06e−4 ± 7.03e−5	4.62e−4 ± 1.20e−4
BNN5D	5.39e−2 ± 3.34e−2	3.56e−2 ± 1.87e−2	8.24e−2 ± 4.09e−2	4.41e−2 ± 3.04e−2	1.88e−1 ± 5.37e−2	1.51e−1 ± 5.12e−2	6.27e−2 ± 5.28e−2	6.32e−2 ± 3.61e−2	2.45e−1 ± 7.66e−2	2.13e−1 ± 6.85e−2	2.15e−2 ± 6.37e−3	5.42e−1 ± 9.21e−2
Average Regret 5D	3.38e−2 ± 7.77e−3	2.87e−2 ± 5.42e−3	5.03e−2 ± 8.80e−3	6.31e−2 ± 6.99e−3	4.70e−2 ± 1.09e−2	4.86e−2 ± 1.04e−2	5.18e−2 ± 1.14e−2	7.30e−2 ± 8.66e−3	7.75e−2 ± 1.61e−2	7.13e−2 ± 1.47e−2	4.14e−2 ± 2.75e−3	1.93e−1 ± 1.86e−2
LEVY10D	6.27e−3 ± 2.22e−3	6.27e−3 ± 2.22e−3	1.04e−2 ± 2.10e−3	1.04e−2 ± 2.10e−3	2.16e−2 ± 3.71e−3	2.17e−2 ± 5.30e−3	8.65e−2 ± 2.01e−2	1.46e−1 ± 1.41e−2	6.21e−3 ± 2.48e−3	5.81e−3 ± 1.99e−3	6.16e−3 ± 8.94e−4	1.06e−1 ± 7.64e−3
ROSENBROCK10D	2.34e−3 ± 1.43e−3	2.01e−3 ± 6.91e−4	9.14e−4 ± 1.25e−4	5.67e−3 ± 1.60e−3	2.25e−3 ± 2.58e−4	4.96e−3 ± 4.40e−4	1.42e−2 ± 4.43e−3	7.65e−2 ± 1.03e−2	5.10e−3 ± 1.44e−3	5.00e−3 ± 1.70e−3	3.09e−3 ± 5.77e−4	2.82e−2 ± 3.92e−3
G-FUNCTION10D	2.92e−1 ± 3.85e−2	3.53e−1 ± 2.74e−2	3.79e−1 ± 2.24e−2	4.21e−1 ± 2.35e−2	1.79e−1 ± 2.54e−2	4.15e−1 ± 1.12e−2	3.24e−1 ± 2.28e−2	3.33e−1 ± 2.06e−2	2.53e−1 ± 3.76e−2	2.55e−1 ± 3.72e−2	3.80e−1 ± 1.34e−2	4.50e−1 ± 6.48e−3
PERM10D	3.74e−4 ± 1.26e−4	4.27e−4 ± 1.34e−4	3.59e−4 ± 2.13e−4	6.52e−4 ± 2.04e−4	3.37e−4 ± 8.49e−5	5.53e−4 ± 1.33e−4	1.01e−3 ± 3.10e−4	1.48e−3 ± 3.92e−4	4.07e−4 ± 1.21e−4	4.08e−4 ± 1.48e−4	5.57e−4 ± 1.43e−4	1.81e−4 ± 5.44e−5
BNN10D	1.19e−1 ± 3.71e−2	9.72e−2 ± 3.40e−2	1.92e−1 ± 5.15e−2	1.64e−1 ± 4.10e−2	1.45e−1 ± 3.58e−2	1.58e−1 ± 2.86e−2	1.49e−1 ± 4.23e−2	1.10e−1 ± 3.90e−2	2.57e−1 ± 4.99e−2	2.00e−1 ± 3.75e−2	8.38e−2 ± 2.25e−2	5.92e−1 ± 5.34e−2
Average Regret 10D	8.40e−2 ± 1.07e−2	9.18e−2 ± 8.74e−3	1.17e−1 ± 1.12e−2	1.21e−1 ± 9.46e−3	6.96e−2 ± 8.80e−3	1.20e−1 ± 6.23e−3	1.15e−1 ± 1.05e−2	1.33e−1 ± 9.48e−3	1.04e−1 ± 1.25e−2	9.32e−2 ± 2.06e−2	9.46e−2 ± 5.23e−3	2.35e−1 ± 1.09e−2
LEVY20D	1.51e−2 ± 1.69e−3	1.40e−2 ± 1.63e−3	1.98e−2 ± 8.61e−3	2.64e−2 ± 1.12e−2	4.27e−2 ± 4.16e−3	6.91e−2 ± 9.00e−3	1.88e−1 ± 1.15e−2	2.01e−1 ± 1.07e−2	1.13e−2 ± 4.60e−3	1.61e−2 ± 5.47e−3	1.98e−2 ± 7.88e−3	1.48e−1 ± 6.58e−3
ROSENBROCK20D	3.47e−2 ± 7.08e−3	6.03e−3 ± 9.93e−4	8.94e−3 ± 4.00e−3	1.91e−2 ± 4.04e−3	7.41e−3 ± 1.14e−3	7.61e−3 ± 2.15e−3	6.46e−2 ± 1.22e−2	1.41e−1 ± 1.29e−2	2.96e−3 ± 1.36e−4	4.27e−3 ± 1.33e−3	1.09e−2 ± 1.52e−3	7.80e−2 ± 8.38e−3
G-FUNCTION20D	4.16e−1 ± 8.99e−3	4.18e−1 ± 1.70e−2	4.59e−1 ± 6.77e−3	4.86e−1 ± 3.95e−3	4.70e−1 ± 4.43e−3	4.92e−1 ± 1.58e−3	4.18e−1 ± 1.33e−2	4.02e−1 ± 1.61e−2	3.79e−1 ± 2.97e−2	3.73e−1 ± 2.56e−2	4.47e−1 ± 9.37e−3	4.86e−1 ± 2.44e−3
PERM20D	7.45e−5 ± 2.09e−5	7.46e−5 ± 2.50e−5	2.20e−4 ± 8.98e−5	1.57e−4 ± 7.50e−5	9.98e−5 ± 3.19e−5	2.03e−4 ± 8.86e−5	3.54e−5 ± 6.52e−6	1.85e−4 ± 6.90e−5	7.98e−5 ± 2.19e−5	9.28e−5 ± 3.89e−5	9.44e−5 ± 3.25e−5	1.90e−5 ± 4.67e−6
BNN20D	1.55e−1 ± 4.07e−2	1.20e−1 ± 3.08e−2	1.79e−1 ± 3.36e−2	1.38e−1 ± 2.67e−2	1.76e−1 ± 3.45e−2	2.43e−1 ± 4.40e−2	1.87e−1 ± 3.90e−2	3.01e−1 ± 7.09e−2	2.92e−1 ± 5.71e−2	3.04e−1 ± 4.99e−2	1.09e−1 ± 2.06e−2	6.85e−1 ± 3.29e−2
Average Regret 20D	1.24e−1 ± 8.46e−3	1.12e−1 ± 7.04e−3	1.33e−1 ± 7.11e−3	1.34e−1 ± 5.90e−3	1.39e−1 ± 7.02e−3	1.62e−1 ± 9.00e−3	1.71e−1 ± 8.89e−3	2.09e−1 ± 1.49e−2	1.37e−1 ± 1.29e−2	1.40e−1 ± 1.13e−2	1.17e−1 ± 4.80e−3	2.80e−1 ± 6.94e−3

has been trained on few training points) should get "bored" much more easily, since it is not sensible to exploit a given local optimum up to the last digits, when there is plenty of time to reach other, possibly better local optima.

Thus, it makes sense to first explore on a large scale where the good local optima are approximately, then to find out which of them are the best ones and finally to exploit the best one in greater detail at the very end.

Therefore, we want the threshold δ_i, that determines if $x_{i'} \overset{\delta_{i'}}{\approx} x_i \iff \|x_{i'} - x_i\|_2 \leq \delta_{i'}$ to decrease in each BO step. In our experiment, we choose an exponential decay of

$$\delta_i = \delta_{n^{\text{start}}} \cdot \left(\frac{\delta_{n^{\text{end}}}}{\delta_{n^{\text{start}}}} \right)^{(i-n^{\text{start}})/(n^{\text{end}}-n^{\text{start}})}, \qquad (56)$$

with $\delta_{n^{\text{start}}} = \frac{1}{16}$ and $\delta_{n^{\text{end}}} = 0.01$.

Concretely, we only evaluate f at $x_{i'}$ if $\|x_{i'} - x_i\|_2 > \delta_{i'} \forall i < i'$ is fulfilled. Otherwise we double c until it is fulfilled (With larger c more emphasis is put on exploration, so there is a tendency that $x_{i'}$ will be further away from the observed input points the larger we pick c, if D3 (Out-of-Sample) is fulfilled). After doubling c 15 times without success, we evaluate f at $x_{i'}$ no matter how close it is to the already observed input points (for methods that have severe troubles to fulfill D3 (Out-of-Sample), such as MCDO, even doubling c infinite times would not help if the maximal uncertainty is within an $\delta_{i'}$-ball around an already observed input point).

B.3.3. BO: Detailed Results

In Table 14, we present the mean final regrets, which correspond to the ranks shown in Table 4 in the main paper.

B.3.4. BO: Regret Plots

In Figures 14–16, we present the regret plots for each test function and both MW values.

C. Extensions

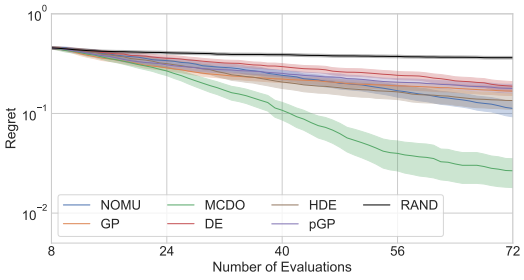
C.1. Incorporating Data Noise

We now discuss one way to incorporate data noise in NOMU. If σ_{n} is unknown, one option to learn it, is to add another output $\hat{\sigma}_{\text{n}}$ to its architecture and change the loss function to

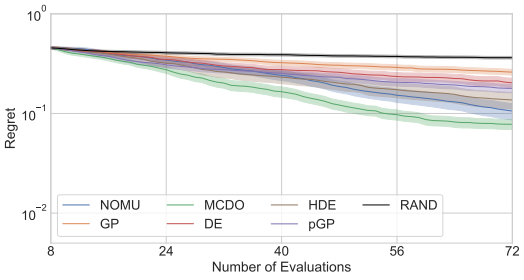
$$\begin{aligned} L^{\pi}(\mathcal{NN}_\theta) := & \sum_{i=1}^{n^{\text{train}}} \left(\frac{\left(\hat{f}(x_i^{\text{train}}) - y_i^{\text{train}} \right)^2}{2\left(\hat{\sigma}_{\text{n}}(x_i^{\text{train}}) \right)^2} + \ln\left(\hat{\sigma}_{\text{n}}(x_i^{\text{train}}) \right) \right) \\ & + \pi_{\text{sqr}} \cdot \sum_{i=1}^{n^{\text{train}}} \frac{\left(\hat{r}_f(x_i^{\text{train}}) \right)^2}{2\left(\hat{\sigma}_{\text{n}}(x_i^{\text{train}}) \right)^2} \\ & + \pi_{\text{exp}} \cdot \frac{1}{\lambda_d(X)} \int_X e^{-c_{\text{exp}} \cdot \hat{r}_f(x)} \, dx, \end{aligned} \qquad (57)$$

in the case of Gaussian data noise uncertainty. In this case we recommend to first train the model prediction $\hat{f}$ and the data noise $\hat{\sigma}_{\text{n}}$ only and then freeze their parameters and train the $\hat{r}_f$-network for capturing model uncertainty $\hat{\sigma}_f$. Note that in the NOMU loss (4) we implicitly assumed a constant very small and negligible data noise σ_{n}^2, absorbed as a factor into the hyperparameter π_{exp} and the regularization factor λ. Therefore, if one wishes to be consistent in these prior assumptions, the hyperparameters π_{exp} and λ need to be divided by $2\sigma_{\text{n}}^2$ in Equation (57). Thus, when using the loss (57) instead of the NOMU loss (4), π_{exp} and λ need to be chosen significantly larger.

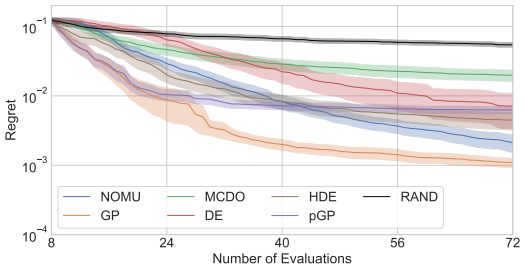
In the case of known (heteroscedastic) data noise $\sigma_{\text{n}}(x)$, (57)



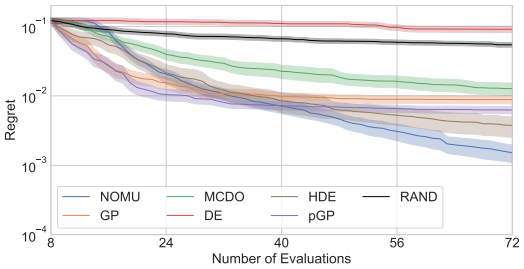
(a) Regret plot G-Function, 0.05 MW



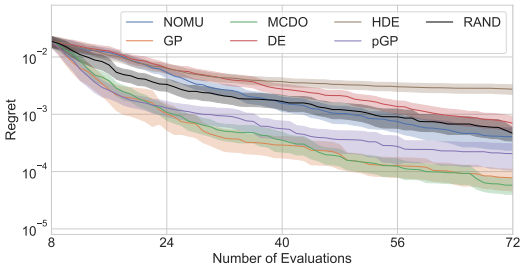
(b) Regret plot G-Function, 0.5 MW



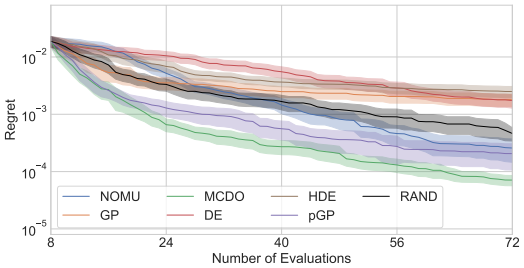
(c) Regret plot Levy, 0.05 MW



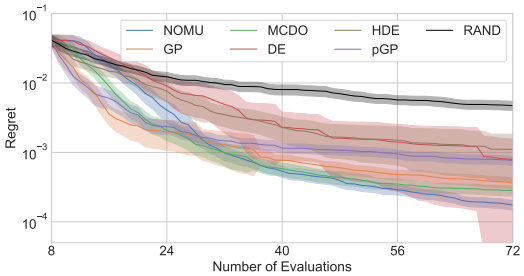
(d) Regret plot Levy, 0.5 MW



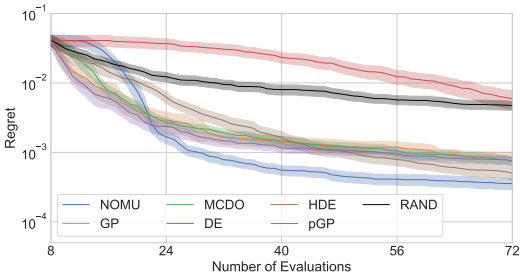
(e) Regret plot Perm, 0.05 MW



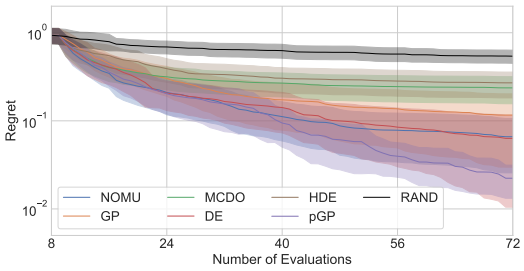
(f) Regret plot Perm, 0.5 MW



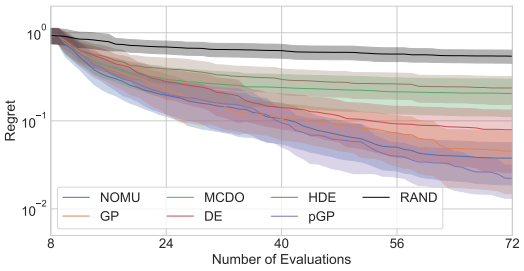
(g) Regret plot Rosenbrock, 0.05 MW



(h) Regret plot Rosenbrock, 0.5 MW

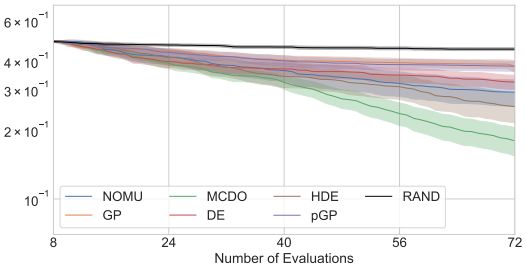


(i) Regret plot BNN, 0.05 MW

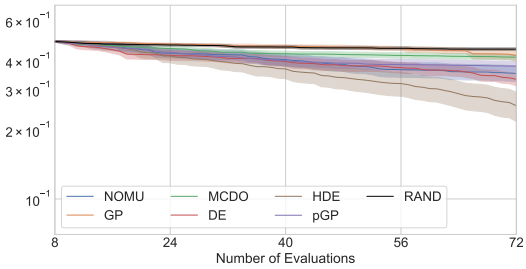


(j) Regret plot BNN, 0.5 MW

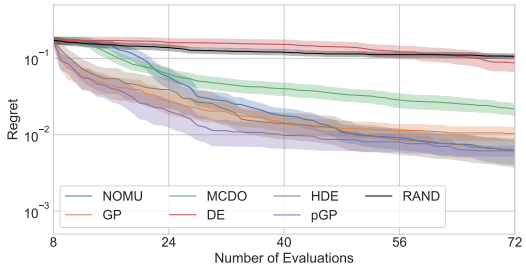
Figure 14. Regret plots for all 5D test functions and MWs of 0.05 and 0.5, respectively. We show regrets averaged over 100 runs (solid lines) with 95% CIs.



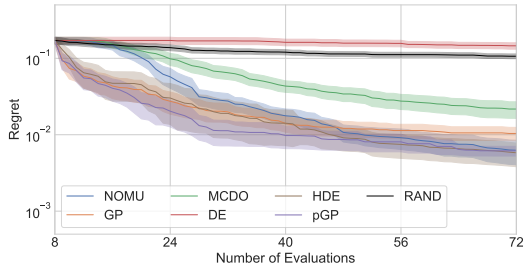
(a) Regret plot G-Function, 0.05 MW



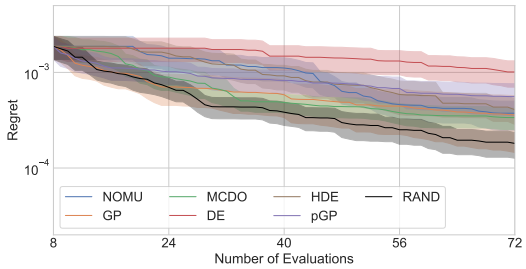
(b) Regret plot G-Function, 0.5 MW



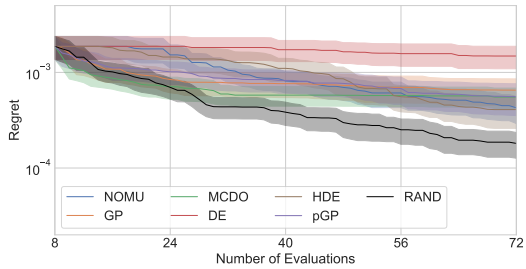
(c) Regret plot Levy, 0.05 MW



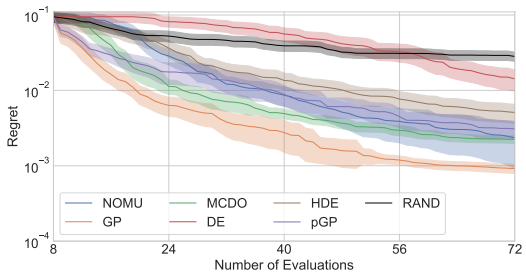
(d) Regret plot Levy, 0.5 MW



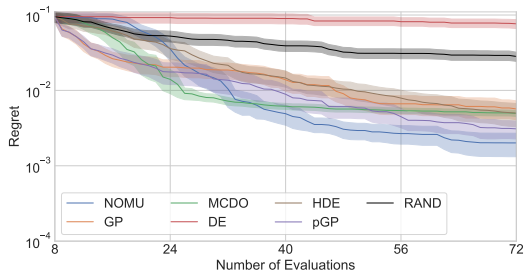
(e) Regret plot Perm, 0.05 MW



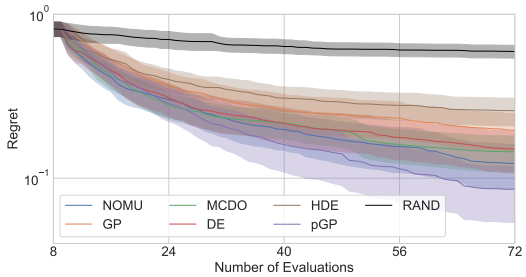
(f) Regret plot Perm, 0.5 MW



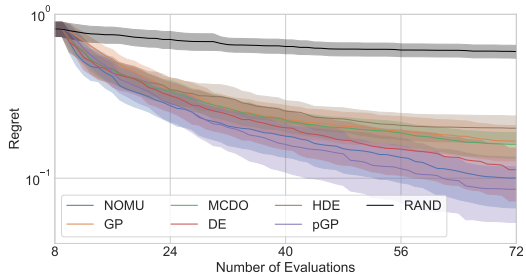
(g) Regret plot Rosenbrock, 0.05 MW



(h) Regret plot Rosenbrock, 0.5 MW

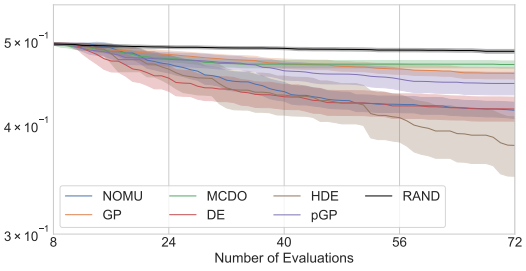


(i) Regret plot BNN, 0.05 MW

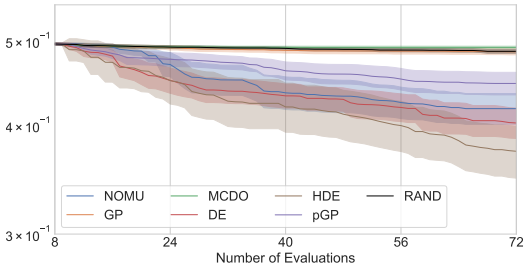


(j) Regret plot BNN, 0.5 MW

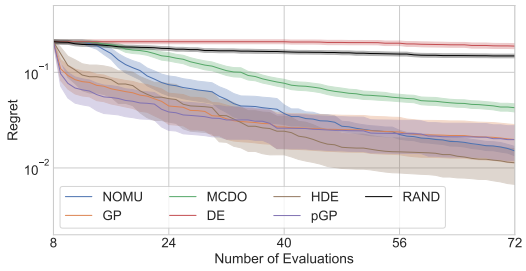
Figure 15. Regret plots for all 10D test functions and MWs of 0.05 and 0.5, respectively. We show regrets averaged over 100 runs (solid lines) with 95% CIs.



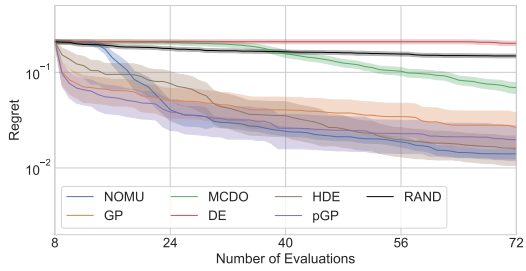
(a) Regret plot G-Function, 0.05 MW



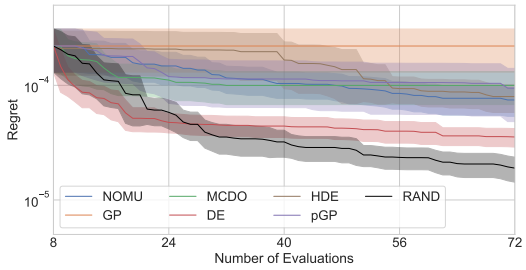
(b) Regret plot G-Function, 0.5 MW



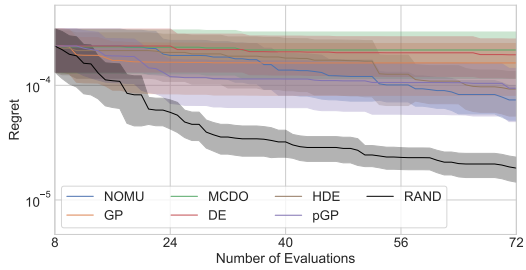
(c) Regret plot Levy, 0.05 MW



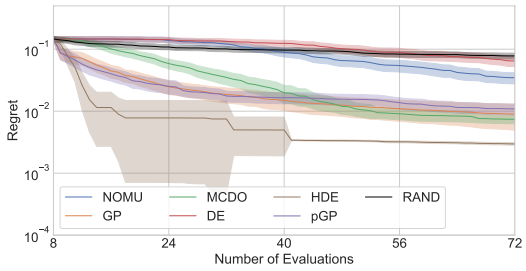
(d) Regret plot Levy, 0.5 MW



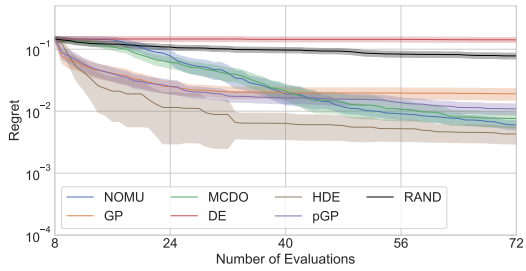
(e) Regret plot Perm, 0.05 MW



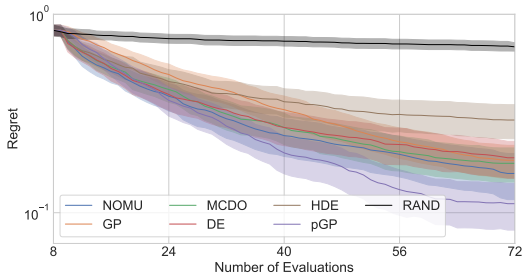
(f) Regret plot Perm, 0.5 MW



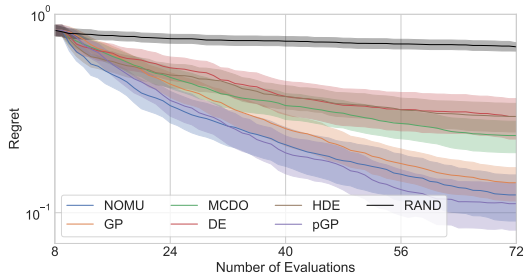
(g) Regret plot Rosenbrock, 0.05 MW



(h) Regret plot Rosenbrock, 0.5 MW



(i) Regret plot BNN, 0.05 MW



(j) Regret plot BNN, 0.5 MW

Figure 16. Regret plots for all 20D test functions and MWs of 0.05 and 0.5, respectively. We show regrets averaged over 100 runs (solid lines) with 95% CIs.

can be simplified, replacing $\hat{\sigma}_{\text{n}}$ by σ_{n} and dropping the term $\ln(\sigma_{\text{n}})$ (in this case, one can then also drop the $\hat{\sigma}_{\text{n}}$-output of the NOMU architecture).

Predictive UBs are then obtained as

$$\left(\hat{f}(x) \mp \sqrt{c_1\, \hat{\sigma}_f^2(x) + c_2\, \hat{\sigma}_{\text{n}}^2(x)}\right) \tag{58}$$

with suitable calibration parameters $c_1, c_2 \in \mathbb{R}_{\geq 0}$.

In the case of known normally distributed data noise (and under the assumption that the posterior of each $f(x)$ is Gaussian), it is sufficient to calibrate one calibration parameter $\tilde{c}$ to obtain approximate α predictive bounds

$$\left(\hat{f}(x) \mp \Phi^{-1}(1 - \frac{1-\alpha}{2})\sqrt{\tilde{c}\, \hat{\sigma}_f^2(x) + \sigma_{\text{n}}^2(x)}\right), \tag{59}$$

where $\tilde{c}$ relates to the typically unknown prior scale.

C.2. NOMU for Upwards and Downwards Directed Hypothesis Classes

As mentioned in Appendix A.1 and discussed in more detail in Appendix A.3, often the set $\mathcal{H}_{D^{\text{train}}}$ is not upwards directed for typical NN-architectures and Equation (7) of Theorem A.1 is not fulfilled in general. Therefore, we carefully designed our NOMU algorithm to be able to cope with settings where the set $\mathcal{H}_{D^{\text{train}}}$ is not upwards and/or downwards directed. The downwards directed property is defined analogously as follows:

Assumption 2 (DOWNWARDS DIRECTED) *For every $f_1, f_2 \in \mathcal{H}_{D^{train}}$ there exists an $f \in \mathcal{H}_{D^{train}}$ such that $f(x) \leq \min(f_1(x), f_2(x))$ for all $x \in X$.*

However, in the following, we discuss a modification of NOMU, which is specifically designed for the case if $\mathcal{H}_{D^{\text{train}}}$ is indeed upwards and/or downwards directed. In this case, by Theorem A.1, we can directly solve

$$\underset{h \in \mathcal{H}_{D^{\text{train}}}}{\arg\max} \int_X u(h(x) - \hat{f}(x))\, dx \tag{60}$$

to obtain an upper UB and/or

$$\underset{h \in \mathcal{H}_{D^{\text{train}}}}{\arg\min} \int_X u(h(x) - \hat{f}(x))\, dx \tag{61}$$

to obtain a lower UB, without the need for any modifications as used in the proposed NOMU algorithm (we do not need the dashed connections in the architecture from Figure 2, we do not need a specific choice of u and we do not need to introduce $\hat{r}_f$ and the final activation function φ.). The UBs obtained in this way *exactly* coincide then with the pointwise upper and lower UBs defined in (5), respectively.

Moreover in this case, $\hat{f}$ can be even removed from Equations (60) and (61) (as can be seen from the proof of Theorem A.4). Thus, in the following loss formulation, we will remove $\hat{f}$ in the respective term ($\tilde{c}$).

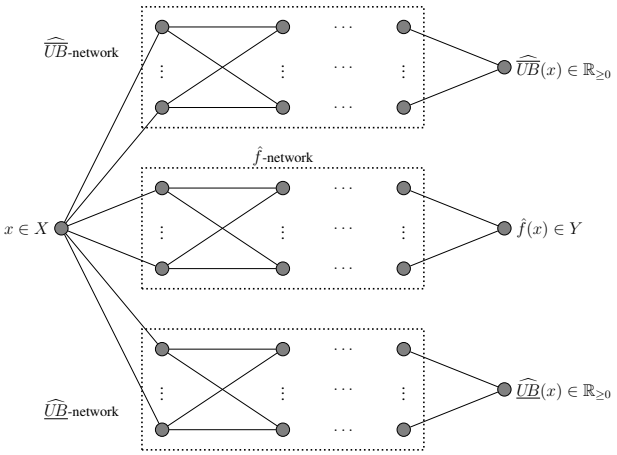


Figure 17. $\widetilde{\mathcal{NN}}_\theta$: a modification of NOMU's original network architecture for upwards and downwards directed hypothesis classes.

C.2.1. THE ARCHITECTURE

Under Assumption 1 (upwards directed) *and* Assumption 2 (downwards directed), we propose an architecture $\widetilde{\mathcal{NN}}_\theta$ consisting of three sub-networks with three outputs: the model prediction $\hat{f}$, the estimated lower UB $\widehat{\underline{UB}}$ and the estimated upper UB $\widehat{\overline{UB}}$. In Figure 17, we provide a schematic representation of $\widetilde{\mathcal{NN}}_\theta$.

C.2.2. THE LOSS FUNCTION

From Equations (60) and (61) we can then directly formulate the following modified NOMU loss function $\tilde{L}^\pi$.

Definition C.1 (NOMU LOSS UPWARDS AND DOWNWARDS DIRECTED) *Let $\pi := (\pi_{sqr}, \pi_{exp}, c_{exp}) \in \mathbb{R}^3_{\geq 0}$ denote a tuple of hyperparameters. Let λ_d denote the d-dimensional Lebesgue measure. Furthermore, let $u : Y \to \mathbb{R}$ be strictly-increasing and continuous. Given a training set D^{train}, the loss function $\tilde{L}^\pi$ is defined as*

$$\tilde{L}^\pi(\widetilde{\mathcal{NN}}_\theta) := \underbrace{\sum_{i=1}^{n^{train}} (\hat{f}(x_i^{train}) - y_i^{train})^2}_{(\tilde{a})} \tag{62}$$

$$+\ \pi_{sqr} \cdot \underbrace{\sum_{i=1}^{n^{train}} (\widehat{\overline{UB}}(x_i^{train}) - y_i^{train})^2 + (\widehat{\underline{UB}}(x_i^{train}) - y_i^{train})^2}_{(\tilde{b})} \tag{63}$$

$$- \pi_{exp} \cdot \underbrace{\frac{1}{\lambda_d(X)} \int_X u(-\widehat{\underline{UB}}(x)) + u(\widehat{\overline{UB}}(x))\, dx}_{(\tilde{c})}. \tag{64}$$

The interpretations of the three terms ($\tilde{a}$), ($\tilde{b}$) and ($\tilde{c}$) are

analogous to the ones in the original NOMU loss formulation.

Note that, the three sub-networks: the $\widehat{\overline{UB}}$-network, the $\widehat{\underline{UB}}$-network and the $\hat{f}$-network can also be trained independently using the corresponding terms in the loss function. Moreover, if one is only interested in the upper (lower) UB or $\mathcal{H}_{D^{\text{train}}}$ is only upwards (downwards) directed, i.e., fulfills only Assumption 1 (Assumption 2), then one can remove the respective sub-network from the architecture as well as the corresponding terms in the loss function. Furthermore, note that, now the obtained UBs can be asymmetric too.

D. Discussion of the Desiderata

In this section, we discuss in more detail the desiderata proposed in Section 3.1. Specifically, we discuss how NOMU fulfills them and thereby prove several propositions. First, we establish a relation of NOMU to the classical Bayesian approach.

The Bayesian point of view allows for mathematically rigorous estimation of uncertainty. However, in general a fully Bayesian approach for quantifying uncertainty is very challenging and involves to

i. formulate a realistic prior,
ii. use an algorithm to approximate the posterior (challenging to get a good approximation in feasible time for complex models such as NNs),
iii. use this approximation of the posterior to obtain UBs.

We follow a different approach by directly approximating iii. based on essential properties of the posterior, e.g. for zero data noise model uncertainty vanishes at, and becomes larger far away from training data. This can be a reasonable approach in applications since many Bayesian state-of-the-art methods even fail to fulfill these basic properties when implemented in practice (Malinin & Gales, 2018) (see Figure 1). Since especially for NNs ii. is very costly, we ask ourselves the question, whether in practice it is even true that one has more intuition about the important properties of the prior than about the important properties of the posterior? In other words, can i. and ii. be skipped by directly approximating iii.? In the case of mean predictions, many successful algorithms following this procedure already exist. These algorithms directly try to approximate the posterior mean by exploiting one's intuition how the posterior of a realistic prior should behave without the need of precisely specifying the prior, e.g.:

1. *Spline regression*: In many applications it is very intuitive that a good approximation of the posterior mean should not have unnecessarily large second derivative, without explicitly stating the corresponding prior. Even though spline regression can be formulated as the posterior mean of a limit of priors (Wahba, 1978), for a practitioner it can be much more intuitive to decide whether spline regression fits to one's prior beliefs by looking at the smoothness properties of the posterior mean than looking at such complex priors.
2. *Convolutional Neural Networks (CNNs)*: In image recognition, it is very intuitive to see that a good approximation of the posterior mean should fulfill typical properties, e.g. two pictures that are just slightly shifted should be mapped to similar outputs. CNNs fulfill such typical properties to a large extent and thus have proven to be very successful in practice. Nevertheless, from a Bayesian point of view these properties rely on a yet unknown prior.

D.1. Desideratum D1 (Non-Negativity)

Desideratum D1 (Non-Negativity) is trivial, since $\sigma_f \geq 0$ per definition. Credible bounds $\underline{CB}$ and $\overline{CB}$ are lower and upper bounds of an interval $[\underline{CB}, \overline{CB}]$, therefore $\underline{CB} \leq \overline{CB}$ holds by definition as well. To the best of our knowledge, every established method to estimate model uncertainty yields bounds that fulfil D1 (Non-Negativity). Furthermore, note that D1 (Non-Negativity) also holds in the presence of data noise uncertainty.

D.1.1. How Does NOMU Fulfill D1 (Non-Negativity)?

By definition, NOMU exactly fulfills D1 (Non-Negativity).

Proposition D.1.a *For NOMU, $\hat{\sigma}_f \geq 0$ and thus $\underline{UB}_c = \hat{f} - c\sigma_f \leq \hat{f} \leq \hat{f} + c\sigma_f = \overline{UB}_c$ for all $c \geq 0$.*

Proof. This holds since $\ell_{\text{min}} \geq 0$ in the readout map φ (see Equation (2)). □

D.2. Desideratum D2 (In-Sample)

In the case of zero data noise, Desideratum D2 (In-Sample) holds true exactly.

Proposition D.2.a (Zero Model Uncertainty at Training Points) *Let $\sigma_n \equiv 0$. Furthermore, let D^{train} be a finite set of training points and consider a prior distribution $\mathbb{P}[f \in \cdot]$ on the function space $\{f : X \to Y\}$ such that there exists a function in the support of $\mathbb{P}[f \in \cdot]$ that exactly fits through the training data. Then for the posterior distribution it holds that for all $(x^{train}, y^{train}) \in D^{train}$ that*

$$\mathbb{P}(f(x^{train}) = y^{train} | D^{train}) = 1 \tag{65}$$
$$\mathbb{P}(f(x^{train}) \neq y^{train} | D^{train}) = 0. \tag{66}$$

In words, there is no model uncertainty at input training points, i.e., $\sigma_f(x^{train}) = 0$ for all $x^{train} \in \{x^{train} : (x^{train}, y^{train}) \in D^{train}\}$.

Proof. Intuitively, if the noise is zero, the data generating process is $y^{\text{train}} = f(x^{\text{train}}) + 0$. Thus, if we observe $(x^{\text{train}}, y^{\text{train}}) \in D^{\text{train}}$, we know that $f(x^{\text{train}}) = y^{\text{train}}$ with zero uncertainty. More formally, let $(x^{\text{train}}, y^{\text{train}}) \in D^{\text{train}}$ and define for some $\epsilon > 0$

$$U_\epsilon(y^{\text{train}}) := (y^{\text{train}} - \epsilon, y^{\text{train}} + \epsilon)$$
$$U_\epsilon(D^{\text{train}}) := \bigcup_{(x,y)\in D^{\text{train}}} U_\epsilon(x, y),$$

where $U_\epsilon(x, y)$ denotes an ϵ-ball around $(x, y) \in X \times Y$. Furthermore, let D be a random variable describing the data generating process assuming zero noise. Then it holds that

$$\begin{aligned}&\mathbb{P}(f(x^{\text{train}}) \in U_\epsilon(y^{\text{train}})^c | D \in U_\epsilon(D^{\text{train}}))\\ =&\frac{\mathbb{P}(D \in U_\epsilon(D^{\text{train}}) \wedge f(x^{\text{train}}) \in U_\epsilon(y^{\text{train}})^c)}{\mathbb{P}(D \in U_\epsilon(D^{\text{train}}))}\\ =&\frac{0}{\mathbb{P}(D \in U_\epsilon(D^{\text{train}}))}.\end{aligned}$$

Note that $\mathbb{P}(D \in U_\epsilon(D^{\text{train}})) > 0$ for every $\epsilon > 0$, since by assumption there exists a function in the support of the prior that exactly fits through the training data.[34] Defining the posterior

$$\begin{aligned}&\mathbb{P}(f(x^{\text{train}}) \neq y^{\text{train}} | D^{\text{train}}) :=\\ &\lim_{\epsilon\to 0} \mathbb{P}(f(x^{\text{train}}) \in U_\epsilon(y^{\text{train}})^c | D \in U_\epsilon(D^{\text{train}})),\end{aligned}$$

in the canonical way concludes the proof. □

Even if theoretically, we know that $\sigma_f(x^{\text{train}}) = 0$ at all training points x^{train}, in practice $\hat{\sigma}_f(x^{\text{train}}) \approx 0$ can be acceptable (due to numerical reasons).

D.2.1. Non-Zero Data Noise

For non-zero data noise there is non-zero *data-noise induced model uncertainty* at input training points. However, also for non-zero but small data noise the model uncertainty at input training points should not be significantly larger than the data noise. In fact, for GPs one can rigorously show that $\sigma_f(x^{\text{train}}) \leq \sigma_{\text{n}}(x^{\text{train}})$ in the case of known σ_{n}.

Proposition D.2.b (GPs Model Uncertainty at training points) *Let D^{train} be a set of training points. For a prior $f \sim \mathcal{GP}(m(\cdot), k(\cdot,\cdot))$ and fixed σ_n it holds that*

$$\sigma_f(x^{train}) \leq \sigma_n(x^{train}), \tag{67}$$

for all input training points x^{train}.

[34] Formally, $\exists f^*$ that fits exactly through D^{train} with the property $\mathbb{P}[f \in U_\epsilon(f^*)] > 0$. Since, $0 < \mathbb{P}\left[f \in U_{\epsilon(f^*)}\right] < \mathbb{P}(D \in U_\epsilon(D^{\text{train}}))$ (for the canonical L_∞-topology) the claim follows. Given this one can also see that Proposition D.2.a still holds true with the even weaker assumption $\mathbb{P}(D \in U_\epsilon(D^{\text{train}})) > 0$.

Proof. We prove the proposition by induction over the number of training points n^{train}. For this let

$$A^{n^{\text{train}}} := \{x^{\text{train}} : (x^{\text{train}}, y^{\text{train}}) \in D^{\text{train}}\}.$$

- **Base case** $n^{\text{train}} = 1$: In this case $A^{n^{\text{train}}} = A^1 = \{x_1^{\text{train}}\}$. Let $k := k(x_1^{\text{train}}, x_1^{\text{train}})$. Since

$$\sigma_f^2(x_1^{\text{train}}) \overset{(29)}{=} k - k\frac{1}{k + \sigma_{\text{n}}^2(x_1^{\text{train}})} \cdot k \leq \sigma_{\text{n}}^2(x_1^{\text{train}}) \tag{68}$$
$$\iff k^2 \geq k^2 - \sigma_{\text{n}}^4(x_1^{\text{train}}), \tag{69}$$

  the claim follows.

- $n^{\text{train}} = m$: Let $K(A^m, A^m)$ be the Gram matrix and let

$$P := \left[K(A^m, A^m) + diag(\sigma_{\text{n}}(A^m))\right].$$

  We then assume for all $x^{\text{train}} \in A^m$ that

$$\sigma_f^2(x^{\text{train}} | A^m) \overset{(29)}{=} \tag{70}$$
$$k(x^{\text{train}}, x^{\text{train}}) - k(x^{\text{train}}, A^m)^T P^{-1} k(x^{\text{train}}, A^m) \leq \sigma_{\text{n}}^2(x^{\text{train}}) \tag{71}$$

- **Inductive step** $n^{\text{train}} = m + 1$: We now show that under the inductive assumption for any $x^{\text{train}} \in A^{m+1}$ we have

$$\begin{aligned}&\sigma_f^2(x^{\text{train}})\\ &\overset{(29)}{=} k(x^{\text{train}}, x^{\text{train}}) - \begin{pmatrix} k(x^{\text{train}}, A^m) \\ k(x^{\text{train}}, x_{m+1}^{\text{train}}) \end{pmatrix}^T \begin{pmatrix} P & Q \\ R & S \end{pmatrix}^{-1} \begin{pmatrix} k(x^{\text{train}}, A^m) \\ k(x^{\text{train}}, x_{m+1}^{\text{train}}) \end{pmatrix}\\ &\leq \sigma_{\text{n}}^2(x^{\text{train}})\end{aligned} \tag{72}$$

  with

$$\begin{aligned}R &:= (k(x_1^{\text{train}}, x_{m+1}^{\text{train}}), \dots k(x_m^{\text{train}}, x_{m+1}^{\text{train}})) = k(x_{m+1}^{\text{train}}, A^m)^T,\\ Q &:= (k(x_1^{\text{train}}, x_{m+1}^{\text{train}}), \dots k(x_m^{\text{train}}, x_{m+1}^{\text{train}}))^T = R^T,\\ S &:= k(x_{m+1}^{\text{train}}, x_{m+1}^{\text{train}}) + \sigma_{\text{n}}^2(x_{m+1}^{\text{train}}).\end{aligned}$$

  Setting $k := k(x^{\text{train}}, x^{\text{train}}), v := k(x^{\text{train}}, A^m)$, and $w := k(x^{\text{train}}, x_{m+1}^{\text{train}})$ (72) can be rewritten as

$$k - \begin{pmatrix} v \\ w \end{pmatrix}^T \begin{pmatrix} \tilde{P} & \tilde{Q} \\ \tilde{R} & \tilde{S} \end{pmatrix} \begin{pmatrix} v \\ w \end{pmatrix} \leq \sigma_{\text{n}}^2(x^{\text{train}}) \tag{73}$$

  with submatrices $\tilde{P}, \tilde{Q}, \tilde{R}, \tilde{S}$ as in (Williams & Rasmussen, 2006, (A.12)). Furthermore, with $M := (S - RP^{-1}Q)^{-1}$ as in (Williams & Rasmussen, 2006, (A.12)), we get that (73) is equivalent to

$$\begin{aligned}&k - (v^T\tilde{P}v + v^T\tilde{Q}w + w\tilde{R}v + w\tilde{S}w) \leq \sigma_{\text{n}}^2(x^{\text{train}})\\ \iff &k - (v^TP^{-1}v + v^TP^{-1}QMRP^{-1}v - v^TP^{-1}QMw\\ &\quad - wMRP^{-1}v + wMw) \leq \sigma_{\text{n}}^2(x^{\text{train}})\\ \iff &k - v^TP^{-1}v - (v^TP^{-1}Q - w)M(RP^{-1}v - w) \leq \sigma_{\text{n}}^2(x^{\text{train}})\\ \iff &\underbrace{k - v^TP^{-1}v}_{=\sigma_f^2(x^{\text{train}}|A^m)} - (v^TP^{-1}Q - w)^2M \leq \sigma_{\text{n}}^2(x^{\text{train}})\end{aligned} \tag{74}$$

where the last line follows since $R^T = Q$ and P symmetric. Note that

$$M = \Big(k(x^{\text{train}}_{m+1}, x^{\text{train}}_{m+1}) + \sigma_{\text{n}}^2(x^{\text{train}}_{m+1}) - k(x^{\text{train}}_{m+1}, A^m)^T P^{-1} k(x^{\text{train}}_{m+1}, A^m) \Big)^{-1} = \big(\sigma_{\text{n}}^2(x^{\text{train}}_{m+1}) + \sigma_f^2(x^{\text{train}}_{m+1}|A^m)\big)^{-1}.$$

With this, (74) can be further reformulated as

$$\sigma_f^2(x^{\text{train}}|A^m) - \underbrace{\frac{(k(x^{\text{train}}, A^m)^T P^{-1} k(x^{\text{train}}_{m+1}, A^m) - k(x^{\text{train}}, x^{\text{train}}_{m+1}))^2}{\sigma_{\text{n}}^2(x^{\text{train}}_{m+1}) + \sigma_f^2(x^{\text{train}}_{m+1}|A^m)}}_{\geq 0} \leq \sigma_{\text{n}}^2(x^{\text{train}}). \tag{75}$$

First, for $x^{\text{train}} \in A^m$, (75) holds true by assumption. Second, for $x^{\text{train}} = x^{\text{train}}_{m+1}$ we obtain

$$\begin{aligned} & \sigma_f^2(x^{\text{train}}_{m+1}|A^m) - \frac{(\sigma_f^2(x^{\text{train}}_{m+1}|A^m))^2}{\sigma_{\text{n}}^2(x^{\text{train}}_{m+1}) + \sigma_f^2(x^{\text{train}}_{m+1}|A^m)} \leq \sigma_{\text{n}}^2(x^{\text{train}}_{m+1}) \\ \iff & \sigma_f^4(x^{\text{train}}_{m+1}|A^m) - \sigma_{\text{n}}^4(x^{\text{train}}_{m+1}) \leq \sigma_f^4(x^{\text{train}}_{m+1}|A^m) \\ \iff & -\sigma_{\text{n}}^4(x^{\text{train}}_{m+1}) \leq 0. \end{aligned}$$

Thus, (73) holds true for any $x^{\text{train}} \in A^{m+1}$. □

In fact, Proposition D.2.b does not come as a surprise: Even if we only observe one training point $(x^{\text{train}}, y^{\text{train}})$ and ignore all our prior knowledge by using a flat "uninformative" improper prior $p(f(x^{\text{train}})) \propto 1$, this results in $\sigma_f(x^{\text{train}}) = \sigma_{\text{n}}(x^{\text{train}})$. Introducing additional information—e.g. observing more additional training points and introducing additional prior information (such as smoothness assumptions instead of a flat uninformative prior)— typically reduces model uncertainty further. Thus, we believe that $\sigma_f(x^{\text{train}}) \leq \sigma_{\text{n}}(x^{\text{train}})$ holds for most reasonable priors.

Finally, note that Proposition D.2.a and Proposition D.2.b hold true for *every* prior respectively *every Gaussian* process prior as long as there exists an f in the support of this prior which explains the observed training points (even if this prior is strongly misspecified). For example this assumption is obviously fulfilled for the prior of Gaussian distributed weights of an overparameterized NN (BNN).

D.2.2. Why Does MC Dropout Strongly Violate D2 (In-Sample) ?

In Figure 1, MC dropout (MCDO) predicts for every input training point $\hat{\sigma}_f(x_i^{\text{train}}) > 100\sigma_{\text{n}}$. Thus, if $\hat{\sigma}_f(x_i^{\text{train}})$ was correctly calculated as posterior model uncertainty, this would be an practically unobservable event as long as f actually comes from this prior ($\mathbb{P}\left[|y_i^{\text{train}} - f(x_i^{\text{train}})| > 100\sigma_{\text{n}}\right] < 10^{-2173}$). Therefore, this is clear statistical evidence that MCDO severely fails to estimate posterior model uncertainty at training points. This can have one of the following three reasons:

1. MCDO severely fails in correctly approximating the posterior given its prior (i.i.d. Gaussian on weights).
2. MCDO's prior does not fit to the data generating process at all.
3. During our experiments we very often observed very extreme events that should only happen with probabilities smaller than 10^{-2000}.

We agree with prior work (Gal & Ghahramani, 2016; Blundell et al., 2015) that a Gaussian prior on the weights of a NN, i.e., the prior mentioned in Item 2, is a very reasonable assumption. Note that NOMU can also be seen as a heuristic to approximate the posterior model uncertainty given exactly the same prior (see Appendix A.2). Therefore, since Item 3 can be ruled out, we can conclude that MCDO's problem is Item 1.

MC Dropout's Failure in Approximating the Posterior Table 10 and Table 11 show that even though we generate the ground truth function from the *same* prior assumed by MCDO (and also assumed by most BNN algorithms), NOMU significantly outperforms MCDO. This empirically shows (with the help of Theorem B.7) that (i) NOMU is able to better approximate posterior BNN-credible bounds than MCDO in terms of average Kullback-Leibler divergence $\bar{d}_{\text{KL}}$ (including further popular variational BNN approximations from (Graves, 2011; Blundell et al., 2015; Hernández-Lobato & Adams, 2015), which themselves are outperformed by MCDO) and (ii) MCDO's variational approximation algorithm severely fails in approximating the targeted posterior.

D.2.3. Importance of D2 (In-Sample) in BO

Especially in Bayesian optimization (BO) it is particularly important to fulfill D2 (In-Sample) as much as possible, since D2 (In-Sample) helps a lot to prevent the BO-algorithm from getting stuck in local maxima. For NNs we often observed that at the i'-th step the mean prediction is maximized/minimized either at the boundary or exactly at a training point with the largest/smallest function value observed so far $\max_{i\in\{1,\dots,i'\}}$/$\min_{i\in\{1,\dots,i'\}}$ (see Figure 9). In the latter case, without model uncertainty (or with almost constant model uncertainty as is the case in MC dropout), one would query all future function evaluations at exactly this point without learning anything new. E.g., consider the situation of Figure 9d when minimizing the *Forrester* function. Each new function evaluation of MC Dropout would be sampled at an already observed training point $x \approx 0.4$. This intuitively explains why estimating model uncertainty precisely at the training data points is especially important in

BO and why it can be very problematic in BO, if the model uncertainty does not decrease sufficiently at the training data points. To summarize, D2 (In-Sample) strongly influences the acquisition function in a direction that discourages the algorithm from choosing the same point again and D2 (In-Sample) together with D3 (Out-of-Sample) can prevent the BO-algorithm from getting stuck (see also Appendix B.3.2).

D.2.4. Dominating Data Noise

In the case of dominating data noise uncertainty $\sigma_{\text{n}} \gg 0$, the model uncertainty σ_f should not be small at input training points (only if one observes a very large amount of input training points very close to a input training point x^{train} the model uncertainty should become small.) However, in this paper, we do not focus on the case of large data noise uncertainty, but on the case of negligible or zero data noise. In particular, D2 (In-Sample) is only formulated for this case.

D.2.5. How Does NOMU Fulfill D2 (In-Sample)?

Recall, that we train NOMU by minimizing

$$L^{\pi}(\mathcal{NN}_{\theta}) + \lambda \left\|\theta\right\|_2^2, \tag{76}$$

where the NOMU loss $L^{\pi}(\mathcal{NN}_{\theta})$ is defined as:

$$L^{\pi}(\mathcal{NN}_{\theta}) := \underbrace{\sum_{i=1}^{n^{\text{train}}} (\hat{f}(x_i^{\text{train}}) - y_i^{\text{train}})^2}_{(a)} + \pi_{\text{sqr}} \cdot \underbrace{\sum_{i=1}^{n^{\text{train}}} (\hat{r}_f(x_i^{\text{train}}))^2}_{(b)} \tag{77}$$

$$+ \pi_{\text{exp}} \cdot \underbrace{\frac{1}{\lambda_d(X)} \int_X e^{-c_{\text{exp}} \cdot \hat{r}_f(x)} \, dx}_{(c)}. \tag{78}$$

Then, the following proposition holds:

Proposition D.2.c *Let $\lambda, \pi_{exp}, c_{exp} \in \mathbb{R}_{\geq 0}$ be fixed and let $\hat{\sigma}_f$ be NOMU's model uncertainty prediction. Then, it holds that $\hat{\sigma}_f(x_i^{train})$ converges to ℓ_{min} for $\pi_{sqr} \to \infty$ for all input training points x_i^{train}, where $\ell_{min} \geq 0$ is an arbitrarily small constant modelling a minimal model uncertainty used for numerical reasons.*

Proof. By the definition of $L^{\pi}(\mathcal{NN}_{\theta})$, i.e., since (b) dominates the loss function if $\pi_{\text{sqr}} \to \infty$, it follows that $\hat{r}_f(x_i^{\text{train}}) = 0$. More precisely, for the NN $\mathcal{NN}_{\theta^*} = (\hat{f}^*, \hat{r}_f^*)$ with parameters θ^* that minimize (76) it holds that

$$\begin{aligned}
&L^{\pi}(\mathcal{NN}_{\theta^*}) + \lambda \left\|\theta^*\right\|_2^2 \leq L^{\pi}(0) + \lambda \left\|0\right\|_2^2 \\
&\qquad = \sum_{i=1}^{n^{\text{train}}} (y_i^{\text{train}})^2 + \pi_{\text{exp}} \cdot 1 \\
\iff &\sum_{i=1}^{n^{\text{train}}} (\hat{f}^*(x_i^{\text{train}}) - y_i^{\text{train}})^2 + \pi_{\text{sqr}} \cdot \sum_{i=1}^{n^{\text{train}}} \left(\hat{r}_f^*(x_i^{\text{train}})\right)^2 + \\
&\pi_{\text{exp}} \cdot \frac{1}{\lambda_d(X)} \int_X e^{-c_{\text{exp}} \cdot \hat{r}_f^*(x)} \, dx + \lambda \left\|\theta^*\right\|_2^2 \leq \sum_{i=1}^{n^{\text{train}}} (y_i^{\text{train}})^2 + \pi_{\text{exp}} \\
\iff &\pi_{\text{sqr}} \cdot \sum_{i=1}^{n^{\text{train}}} \left(\hat{r}_f^*(x_i^{\text{train}})\right)^2 \leq \sum_{i=1}^{n^{\text{train}}} (y_i^{\text{train}})^2 + \pi_{\text{exp}} =: C
\end{aligned}$$

where for fixed parameters $\lambda, \pi_{\text{exp}}, c_{\text{exp}} \in \mathbb{R}_{\geq 0}$, $C > 0$ is a constant. Assume now that for $\hat{r}_f^*$ does not vanish at all training data points for π_{sqr} to infinity, i.e., that there exists an $\epsilon > 0$ such that for every π_{sqr} large enough $\sum_{i=1}^{n^{\text{train}}} \left(\hat{r}_f^*(x_i^{\text{train}})\right)^2 > \epsilon$. This however implies

$$\pi_{\text{sqr}} \cdot \epsilon < C \iff \pi_{\text{sqr}} < \frac{C}{\epsilon} \quad \forall \pi_{\text{sqr}} \text{ large enough},$$

which yields a contradiction. Thus, $\lim_{\pi_{\text{sqr}} \to \infty} \hat{r}_f^*(x^{\text{train}}) = 0$ for all training input points x^{train}. Finally, by Equation (2) it follows that $\hat{\sigma}_f(x_i^{\text{train}}) = \ell_{\text{min}}$. □

Note that even for a finite (sufficiently large) π_{sqr}, the raw model uncertainty $\hat{r}_f$ converges to zero as λ goes to zero ($\frac{\pi_{\text{sqr}}}{\lambda} \to \infty$), if the model is sufficiently over-parameterized. Empirically one can see in Figures 1, 4, 5, 9, 11, 13, 18 and 19 that NOMU fulfills D2 (In-Sample) with a high precision for our choice of hyper-parameters.

D.3. Desideratum D3 (Out-of-Sample)

We first consider the case of zero (or negligible) data noise $\sigma_{\text{n}} \approx 0$ and then discuss possible extensions to settings with non-zero data noise.

D.3.1. Zero Data Noise

The notion of distance used in D3 (Out-of-Sample) heavily depends on the specific application (i.e., on the prior used in this application). More concretely, there are the following two "hyperparameters".

1. First, the metric[35] $d : X \times X \to \mathbb{R}_{\geq 0}$ on X we use to measure distances can heavily depend on the prior for the specific application. For example, in the case of image recognition, two pictures that are only slightly shifted can be seen as very close to each other even if the Euclidean distance of their pixel-values is quite high.[36] If one uses a CNN-architecture in NOMU this prior belief on d is

[35] We use the term "metric" to describe a general pseudometric.

[36] We assume in this paper that if one sees a 1920×1080-pixel

approximately captured. The successful generalization properties of many different network architectures can be explained precisely by their use of application-dependent *non-Euclidean* metrics (Bronstein et al., 2017). (Additionally, instead of fixing d apriori, further aspects of the metric d can be learned from the training data as we discuss in detail in Appendix D.4.)

2. Second, even if we can precisely write down a metric $d : X \times X \to \mathbb{R}_{\geq 0}$, a priori it is not clear how to define the distance $\tilde{d} : X \times 2^X \to \mathbb{R}_{\geq 0}$ between a point x and the input training points from $D_X^{\text{train}} := \{x^{\text{train}} : (x^{\text{train}}, y^{\text{train}}) \in D^{\text{train}}\}$. Both common definitions $d(x, D_X^{\text{train}}) := \inf_{z \in D_X^{\text{train}}} d(x, z)$ and $d(x, \text{Conv}(D_X^{\text{train}}))$, where $\text{Conv}(\cdot)$ denotes the convex hull, are inappropriate choices for $\tilde{d}$.[37] In Section 3.1, we consider a point x closer to the input training points if it is "surrounded" by input training points in all directions, as opposed to a point x which only has close input training points in some directions and there is a large range of directions without any close input training points. This implies that, for example,

   (i) very close to noiseless input training points that are surrounded by many other noiseless input training points there is very little model uncertainty.

   (ii) for extrapolation one typically has more uncertainty than for interpolation.

   (iii) far away from the convex hull of the training points model uncertainty is very high.

   (iv) also within the convex hull of the training data, model uncertainty is high within big gaps in-between training points.

   Figure 11 shows how well NOMU fulfills these properties of $\tilde{d}$ similarly to a GP (see Figure 12a).

D.3.2. Non-Zero Homoscedastic Data Noise

If there is homoscedastic non-zero data noise $\sigma_{\text{n}}(x) \equiv \sigma_{\text{n}} > 0$, it is important that the "distance" $\tilde{d}$ of x to the input training points is not minimal if it exactly equals one of the input training points. Instead, one should use a notion of distance $\tilde{d}$ that can even get smaller if there are multiple input training points at x or very close to x.

D.3.3. Non-Zero Heteroscedastic Data Noise

One can also extend D3 (Out-of-Sample) to heteroscedastic settings. In that case, the used notion of "distance" $\tilde{d}$ of x to the input training points needs to be weighted by the precision of the input training points, i.e. if x is close to multiple input training points with low data noise $\sigma_{\text{n}}(\cdot)$ you consider x "closer" to the input training points than if x is close to multiple input training points with high data noise.

D.3.4. Example for $\tilde{d}$ Based on GPs

In this section, we give the concrete example of Gaussian process regression (GPR) from Appendix B.1.1 in which $\tilde{d}$ from D3 (Out-of-Sample) can be written down explicitly in closed form.

For any arbitrary metric d on X, consider for instance the kernel $k(x_i, x_j) = e^{-d(x_i, x_j)^2}$. Then, $\tilde{d}(x, D_X^{\text{train}}) = \hat{\sigma}_f(x | D_X^{\text{train}})$, with posterior model uncertainty $\hat{\sigma}_f$ from Equation (29). While this is one of the simplest possible ways to define $\tilde{d}$, alternatively one could also define it differently if it shares similar qualitative properties.[38]

Why do we still consider it interesting to formulate D3 (Out-of-Sample) *vaguely*, given that there is already such a precise formula as is the case for GPs?

1. The GP's formula only holds true for the specific prior of a GP. We however, want to formulate desiderata that capture the most essential properties of credible bounds that almost all reasonable priors have in common.

2. We want to provide some easy to understand intuition for D3 (Out-of-Sample): It might be challenging to see directly from the GP's formula (29) how the posterior model uncertainty qualitatively behaves as visualized in Figure 12a.

To summarize, both the exact notion of distances $d, \tilde{d}$ *and* the exact rate of how model uncertainty increases with increasing distance to the input training points depend on one's prior belief. However, Section 3.1 gives a qualitative description of properties that most reasonable (generic) priors have in common (see Items (i)–(iv)).

D.3.5. How Does NOMU Fulfill D3 (Out-of-Sample)?

Recall, that we train NOMU by minimizing

$$L^{\pi}(\mathcal{NN}_{\theta}) + \lambda \left\|\theta\right\|_2^2, \tag{79}$$

image, which is perfectly recognizable as a cat, every 10-pixel shift of this picture is also recognizable as a cat with almost no uncertainty (even though this cannot be proven mathematically). Thus, it is very desirable to predict very small model uncertainty $\hat{\sigma}_f(x)$ for every image $x \in X$ which is only shifted by less than 10 pixel from at least one noiselessly labeled training image x^{train}.

[37] E.g. for GPs, none of these two classical notions of distance between a point and a set is entirely applicable (see Equation (29)). An appropriate choice of $\tilde{d}$ should be a compromise between these two notions.

[38] For instance, $\tilde{d}$ could also be defined based on a kernel of the form $k(x_i, x_j) = g(d(x_i, x_j))$ with a monotonically decreasing function g, e.g. a Matérn-typed kernel.

where the NOMU loss $L^{\pi}(\mathcal{NN}_\theta)$ is defined as:

$$L^{\pi}(\mathcal{NN}_\theta) := \underbrace{\sum_{i=1}^{n^{\text{train}}} (\hat{f}(x_i^{\text{train}}) - y_i^{\text{train}})^2}_{(a)} + \pi_{\text{sqr}} \cdot \underbrace{\sum_{i=1}^{n^{\text{train}}} (\hat{r}_f(x_i^{\text{train}}))^2}_{(b)} \tag{80}$$

$$+ \pi_{\text{exp}} \cdot \underbrace{\frac{1}{\lambda_d(X)} \int_X e^{-c_{\text{exp}} \cdot \hat{r}_f(x)} \, dx}_{(c)} . \tag{81}$$

The interplay of (b), (c), and regularization promotes D3 (Out-of-Sample) (note that the behaviour of $\hat{r}_f$ directly translates to the behaviour of σ_f): Term (c) pushes $\hat{r}_f$ towards infinity across the whole input space X. However, due to the counteracting force of (b) as well as regularization, $\hat{r}_f$ increases continuously as you move away from the training data – see for example Figure 9 and Figure 11 (or any other plot showing NOMU, i.e. Figures 1, 4, 5, 13, 18 and 19). In Figure 11 one can see how NOMU fulfills the properties (i)–(iv) of $\tilde{d} : X \times 2^X \to \mathbb{R}_{\geq 0}$ mentioned in Appendix D.3.1. In Figures 18 and 19 one can observe how NOMU behaves when a non-stationary metric $d_{\text{Figure 18}}(x, x') \neq |x - x'|$ respectively non-stationary non-isotropic metric $d_{\text{Figure 19}}(x, x') \neq \|x - x'\|_2$ is used (because $d_{\text{Figure 18}}$ and $d_{\text{Figure 19}}$ were learned from the data as desired by D4 (Metric Learning) in these examples).

The hyperparameters π_{exp} and c_{exp} control the size and shape of the UBs. Concretely, the larger π_{exp}, the wider the UBs; the larger c_{exp}, the narrower the UBs at points x with large $\hat{\sigma}_f(x)$ and the wider the UBs at points x with small $\hat{\sigma}_f(x)$.

Finally, we give some intuition that if CNNs are used for the two sub-networks in NOMU's architecture, D3 (Out-of-Sample) will be fulfilled with respect to an almost shift-invariant metric d: In the noiseless setting, we can choose π_{sqr} large enough such that D2 (In-Sample) is fulfilled, so that we have $\hat{\sigma}_f(x^{\text{train}}) \approx 0$ at any training input point x^{train}. Regularized CNNs have the property that if you slightly shift the input the output barely changes. So if x can be obtained from x^{train} by slightly shifting it, the CNN-output $\hat{\sigma}_f(x) \approx \hat{\sigma}_f(x^{\text{train}}) \approx 0$ also does not move too far away from zero. Only if you move further away with respect to the almost shift-invariant metric d, the CNN-output $\hat{\sigma}_f$ is able to move further away from zero. The same principle can also be used for other geometric NNs (e.g. graph neural networks (GNNs)) which correspond to different (non-Euclidean) metrics Bronstein et al. (2017).

D.4. Desideratum D4 (Metric Learning)

A priori, it is often not obvious which metric d to choose in D3 to measure distances. In many applications, it is therefore best to *learn* this metric from the training data (as explained in Footnote 4 on Page 4).

In the following section, we present visualizations of D4 (Metric Learning) for all benchmark algorithms in easy to understand, low dimensional settings.

D.4.1. Visualization of D4 (Metric Learning)

1D In order to visualize D4 and show how for NOMU the mean prediction impacts it's model uncertainty prediction we conduct the following experiment. We sample 16 *equidistant* noiseless training points of a trend-adjusted version of *Sine 3*. We then train NOMU (hyperparameters are as in Appendix B.2.2 with $\pi_{\text{sqr}} = 0.5$, $\ell_{\text{min}} = 10^{-4}$, regularization parameter 10^{-4} on the $\hat{r}_f$-network, and number of training epochs 2^{12}) and compute the corresponding UBs. Figure 18 shows that NOMU UBs are wider (cp. the dotted blue line) in those areas of the input space where small changes of x lead to large variation in the target ($\approx x \geq 0$) compared to areas without large variation in the target ($\approx x \leq 0$). This effect is present even though the input training points are sampled from an *equidistant* grid, and thus isolates the effect of D4.

2D Analogously, we visualize D4 for two-dimensional input by training NOMU on 16 training points sampled on an *equidistant* 4×4-grid and evaluated at the two-dimensional extension of the *Step* function, i.e.,

$$f = \mathbb{R}^2 \to \mathbb{R} : (x_1, x_2) \mapsto \begin{cases} -1 & \text{if } x_1 < 0 \\ 1 & \text{if } x_1 \geq 0. \end{cases} \tag{82}$$

Here, D4 can be interpreted as follows: imagine we do not have any prior knowledge of whether x_1 or x_2 is more important for predicting the unknown function f. However, when NOMU observes the 16 training points it should be able to learn that x_1 is more important for the model prediction than x_2, *and* that the function is more regular/predictable far away from $\{x_1 \approx 0\}$. D4 requires in this example that feature x_1 should have a higher impact than feature x_2 also

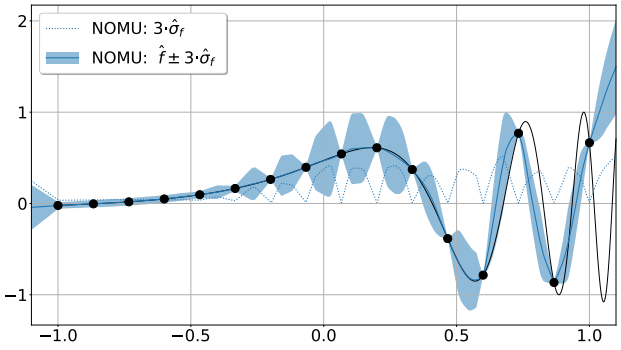


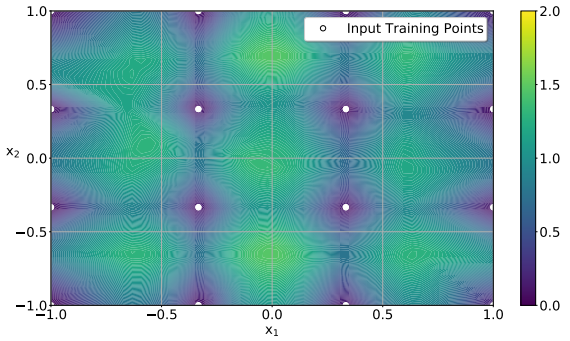
Figure 18. Visualisation of D4 (Metric Learning).

on the model uncertainty prediction. If a model for UBs did not incorporate D4, we would expect the uncertainty in this example to fulfill $\hat{\sigma}_f((x_1, x_2)) = \hat{\sigma}_f((x_2, x_1))$ because of the equidistant grid of the training points (this is indeed the case for GPs, see Figure 20a).

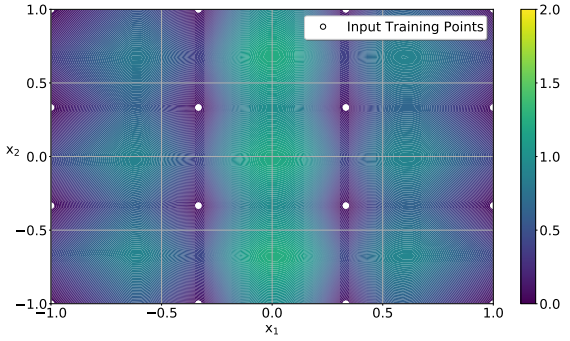
For NOMU however, we have very good control on how strongly we enforce D4, e.g., we can strengthen D4 by increasing the L2-regularization of the hidden layers in the $\hat{r}_f$-network and/or decreasing the size of the $\hat{r}_f$-network.

NOMU: Visualization of D4 in 2D In Figure 19, we present the estimated model uncertainty $\hat{\sigma}_f$ obtained for *different* hyperparameters of the $\hat{r}_f$-network with *fixed* $\hat{f}$-architecture among all four subplots. Thus, Figure 19 shows how D4 realizes in different magnitudes. In Figure 19a, we use the same hyperparameters for the $\hat{r}_f$-network as for the $\hat{f}$-network. In 19b, we only increase the L2-regularization of the $\hat{r}_f$-network. In Figure 19c, we only decrease the size of $\hat{r}_f$-networks. In Figure 19d, we combine both, i.e., we increase the L2-regularization of the $\hat{r}_f$-network *and* decrease the size of the $\hat{r}_f$-network. While D4 is barely visible in Figure 19a, it is clearly visible in Figures 19b–19d. In Figures 19b–19d, we observe that the estimated model uncertainty $\hat{\sigma}_f$ grows faster in horizontal directions (corresponding to changes in x_1) than in vertical directions. In Figures 19b–19d, we further observe that the estimated model uncertainty $\hat{\sigma}_f$ is larger around $\{x_1 \approx 0\}$ than far away from this region. The magnitude of both these effects increases from Figure 19b to Figure 19d. Both of these effects can also be observed for MC dropout (MCDO) and deep ensembles (DE) (see Figure 20b and Figure 20c).

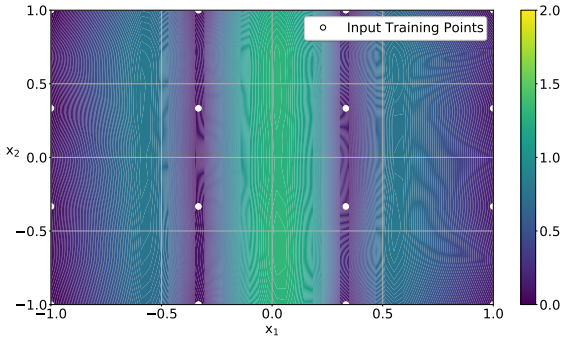
Benchmarks: Visualization of D4 in 2D In Figure 20, we present uncertainty plots of all benchmark methods. We can see that deep ensembles (DE) gives high preference to capturing D4, even though its estimated model uncertainty still is subject to some randomness with non-uniform patterns for $x_1 \in [-0.25, 0.25]$ (Figure 20c). Moreover, MC dropout (MCDO) also captures higher model uncertainty for $x_1 \in [-0.25, 0.25]$ as desired by D4, but it does not fulfill D2 (Figure 20b). The Gaussian process (GP) with RBF kernel does not account for D4 (Figure 20a), which directly follows from the definition. Similarly to deep ensembles (DE), hyper deep ensembles (HDE) and HDE* strongly capture D4 but show even more random behaviour. This randomness is visible most prominently along $x_1 = 0$ where one should observe large model uncertainty, whereas their estimated model uncertainty is surprisingly close to 0.



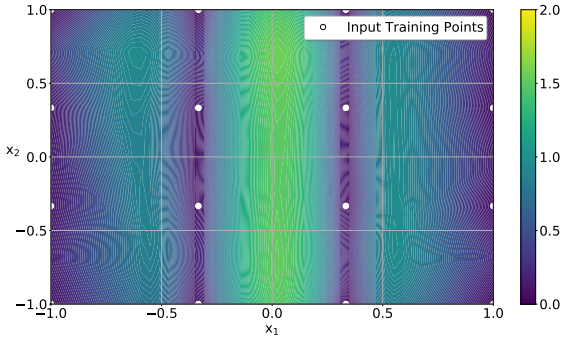
(a) Same L2-regularization on the $\hat{r}_f$-network and $\hat{f}$-network ($\lambda = 10^{-8}$).



(b) Larger L2-regularization on the $\hat{r}_f$-network ($\lambda = 10^{-4}$) than on the $\hat{f}$-network ($\lambda = 10^{-8}$).

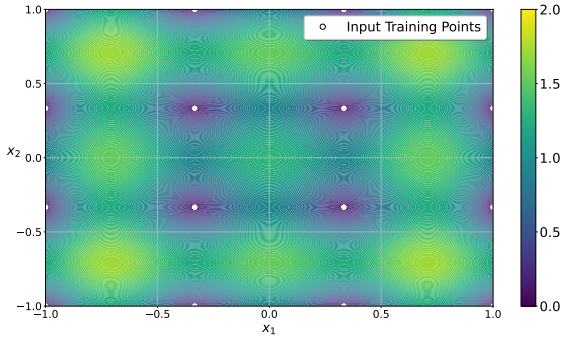


(c) Shallow $\hat{r}_f$-network consisting of 4 hidden nodes.

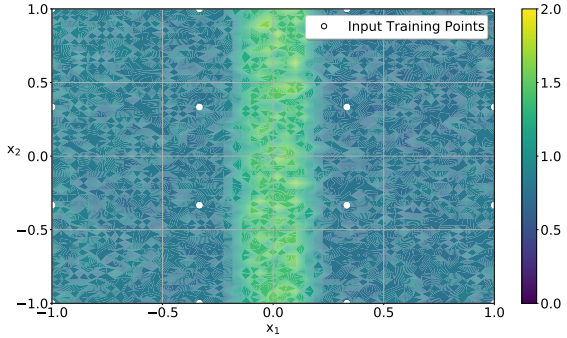


(d) Shallow $\hat{r}_f$-network consisting of 4 hidden nodes *and* larger regularization of $\lambda = 10^{-4}$ on $\hat{r}_f$-network.

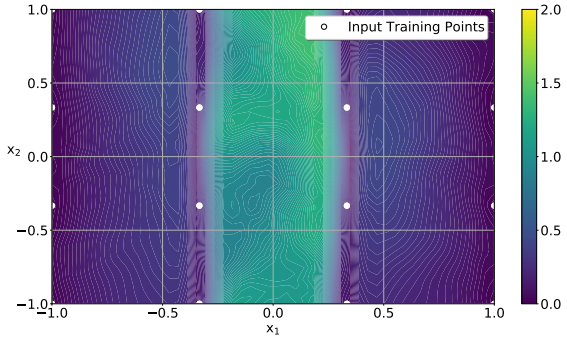
Figure 19. Estimated model uncertainty $\hat{\sigma}_f$ of NOMU: visualizing D4 in 2D.



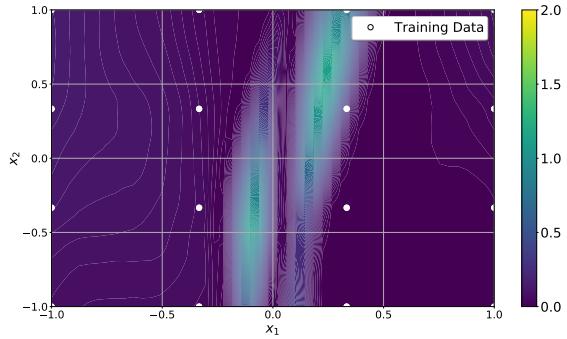
(a) GP (c=15)



(b) MCDO (c=20)

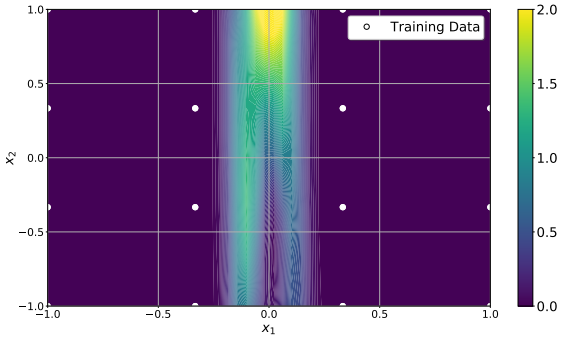


(c) DE (c=30)



(d) HDE (c=30)

Figure 20. Estimated model uncertainty of Gaussian process (GP), MC dropout (MCDO), deep ensembles (DE), and hyper deep ensembles (HDE)



(e) HDE* (c=30)

Figure 20. (cont.) Estimated model uncertainty of HDE*.

D.4.2. How Does NOMU Fulfill D4 (Metric Learning)?

Recall NOMU's architecture depicted in Figure 2 in the main paper.

The $\hat{r}_f$-network learns the raw model uncertainty and is connected with the $\hat{f}$-network through the last hidden layer (dashed lines in Figure 2). This connection enables $\hat{r}_f$ to re-use features that are important for the model prediction $\hat{f}$. This behaviour directly translates to $\hat{\sigma}_f$ (see Equation (2)), implementing Desideratum D4 (Metric Learning).

In Figures 18 and 19, one can observe that NOMU fulfills D4 (Metric Learning). Moreover, Figure 19 shows how to control the strength of D4 (Metric Learning) by varying the L2-regularization and the number of neurons of the $\hat{r}_f$-architecture.

D.5. Desideratum D5 (Vanishing)

First, note that D3 (Out-of-Sample) already suggest D5 (Vanishing), since in the limit $n^{\text{train}} \to \infty$ every point x in the support of the input data generating distribution is infinitely close to infinitely many other input training points x^{train} and therefore infinitely close to the input training set.

To the best of our knowledge, D5 (Vanishing) is fulfilled by most reasonable algorithms that estimate model uncertainty. Specifically, NOMU, GPs, DE and HDE capture D5 (Vanishing). This can be nicely observed in Figure 13, where all these algorithms result in zero model uncertainty in areas with many input training points (up to numerical precision).

MC Dropout can be seen as a variational algorithm for approximating BNNs. While in theory, BNNs should also fulfill D5 (Vanishing), MC Dropout often struggles to do so, as can be observed in Figure 13 (see also D.2.2 for a discussion why MC Dropout struggles to approximate BNNs).

Finally, D5 (Vanishing) is widely accepted (Kendall & Gal, 2017; Malinin & Gales, 2018) and most of the time

loosely stated along the lines of "While data noise uncertainty (aleatoric uncertainty) is irreducible, model uncertainty (epistemic uncertainty) vanishes with an increasing number of training observations." In other words, the width of credible bounds converges to zero whilst the width of predictive bounds does not converge to zero in the presence of data noise.

However, whilst such statements are qualitatively true, formally, D5 (Vanishing) only holds *in the limit* of the number of training points n^{train} to infinity and for $x \in X$ that are *in the support* of the input data generating distribution. Furthermore, note that D3 (Out-of-Sample) also holds in the presence of data noise uncertainty.

Moreover, note that while D1–D4 are statements on *relative* model uncertainty, i.e., statements that are independent of the calibration parameter $c \geq 0$ (see A.2.1), D5 (Vanishing) is a statement about *absolute* model uncertainty. Thus, D5 (Vanishing) only holds for a fixed calibration parameter $c \geq 0$ (if c increases sufficiently fast with increasing n^{train}, model uncertainty does not vanish).

D.5.1. Why does D5 (Vanishing) only hold in the limit?

1. Even for fixed $c \geq 0$, in the case of large *unknown* data noise uncertainty that is simultaneously learned by the algorithm, adding another input training point x close to existing input training points x^{train} whose corresponding target y is very far away from y^{train} could lead to an increase in model uncertainty, since this new training point (x, y) would increase the predicted data noise uncertainty and thus increase the data noise induced model uncertainty.

2. Even if there is no data noise uncertainty $\sigma_{\text{n}} \equiv 0$ and $c \geq 0$ is fixed, adding another input training point can increase the model uncertainty, when D4 (Metric Learning) is fulfilled. To see this, consider the following scenario: an already observed set of training points suggest that f is very flat/simple/predictable (e.g. linear) in a certain region. However, adding a new training point (x, y) shows that f is much more irregular in this region than expected. Then, the learned metric can drastically change resulting in increased model uncertainty in this region (outside an ϵ-ball around the new input training point x).

D.5.2. How Does NOMU Fulfill D5 (Vanishing)?

Recall, that we train NOMU by minimizing

$$L^{\pi}(\mathcal{NN}_{\theta}) + \lambda \left\|\theta\right\|_2^2, \tag{83}$$

where the NOMU loss $L^{\pi}(\mathcal{NN}_{\theta})$ is defined as:

$$L^{\pi}(\mathcal{NN}_{\theta}) := \underbrace{\sum_{i=1}^{n^{\text{train}}} (\hat{f}(x_i^{\text{train}}) - y_i^{\text{train}})^2}_{(a)} + \pi_{\text{sqr}} \cdot \underbrace{\sum_{i=1}^{n^{\text{train}}} (\hat{r}_f(x_i^{\text{train}}))^2}_{(b)} \tag{84}$$

$$+ \pi_{\text{exp}} \cdot \underbrace{\frac{1}{\lambda_d(X)} \int_X e^{-c_{\text{exp}} \cdot \hat{r}_f(x)} \, dx}_{(c)}. \tag{85}$$

Then, the following proposition holds:

Proposition D.5.a *Let $\lambda, \pi_{exp}, c_{exp}, \pi_{sqr} \in \mathbb{R}_{\geq 0}$ be fixed and let the activation-functions of $\mathcal{NN}_{\theta}$ be Lipschitz-continuous and let $\hat{\sigma}_f(x)$ be NOMU's model uncertainty prediction. Then, it holds that $\hat{\sigma}_f(x)$ converges in probability to ℓ_{min} for $n^{train} \to \infty$ for all input training points x in the support of the input data generating distribution $\mathbb{P}_X$, i.e., more formally $\forall x \in supp\left(\mathbb{P}_X\right), \forall \epsilon > 0, \forall \delta \geq 0 : \exists n^0 : \forall n \geq n^0 :$*

$$\mathbb{P}\left[|\hat{\sigma}_f(x) - \ell_{min}| \geq \epsilon\right] < \delta, \tag{86}$$

where $\ell_{min} \geq 0$ is an arbitrarily small constant modelling a minimal model uncertainty used for numerical reasons.

Proof. Let $x \in \text{supp}\left(\mathbb{P}_X\right)$, $\delta \geq 0$ and $\epsilon > 0$.

First, let $L^{\pi}(0) := c < \infty$, i.e., the value of the loss function when inserting the constant zero function. Then for the optimal solution $\mathcal{NN}_{\theta^*}$ of (83) it immediately follows that

$$\left\|\theta^*\right\|_2^2 \leq \frac{c}{\lambda}. \tag{87}$$

Using the fact that all activation-functions ϕ of $\mathcal{NN}_{\theta}^*$ are Lipschitz-continuous together with (87), one can show that there exists a constant $L := L(c, \lambda, \phi, \text{architecture})$ such that the raw model uncertainty prediction $\hat{r}_f^{\theta^*}$ is Lipschitz-continuous with constant L.

Next, let $U := U_{\frac{\epsilon}{4L}}(x)$ be an open ball with radius $\frac{\epsilon}{4L}$ around x. Given that the diameter of U is equal to $\frac{\epsilon}{2L}$ and the fact that $\hat{r}_f^{\theta^*}$ is Lipschitz-continuous with constant L, it follows that

$$\max_{z \in U} \hat{r}_f^{\theta^*}(z) - \min_{z \in U} \hat{r}_f^{\theta^*}(z) < \frac{\epsilon}{2}. \tag{88}$$

Given U let $p := \mathbb{P}\left[x \in U\right]$. Since $x \in \text{supp}\left(\mathbb{P}_X\right)$, per definition it holds that $p > 0$.

In the following let $D_{n,x}^{\text{train}} \sim (\mathbb{P}_X)^n$ denote the random variable representing a set of n input training points. Now, let n^0 be sufficiently large such that

$$\mathbb{P}\left[|D_{n^0,x}^{\text{train}} \cap U| > 4 \cdot \frac{c}{\epsilon^2}\right] > 1 - \delta. \tag{89}$$

Note that one can explicitly calculate the value of n^0, since $|D^{\text{train}}_{n^0,x} \cap U| \in \mathbb{N}_0$ is binomial distributed with $p > 0$ and $n^0 \in \mathbb{N}$.

Finally, we show that $\hat{r}_f^{\theta^*}(x) < \epsilon$ by contradiction. For this, assume on the contrary that

$$\hat{r}_f^{\theta^*}(x) \geq \epsilon. \tag{90}$$

Using (88) it follows that for all $z \in D^{\text{train}}_{n^0,x} \cap U$ it holds that $\hat{r}_f^{\theta^*}(z) > \frac{\epsilon}{2}$ with probability larger than $1-\delta$. This together with the fact that each summand in the term (b) in the NOMU loss function L^π is non-negative implies that

$$(b) \geq \sum_{x^{\text{train}} \in D^{\text{train}}_{n^0,x} \cap U} \left(\hat{r}_f(x^{\text{train}})\right)^2 > \left(\frac{\epsilon}{2}\right)^2 \cdot 4 \cdot \frac{c}{\epsilon^2} = c. \tag{91}$$

Putting everything together and using the fact that each term in the NOMU loss is non-negative implies that

$$L^\pi(\mathcal{NN}_{\theta^*}) + \lambda \left\|\theta^*\right\|_2^2 \geq (b) \overset{(91)}{>} c = L^\pi(0) + \lambda \left\|0\right\|_2^2,$$

which is a contradiction for $\mathcal{NN}_{\theta^*}$ being optimal in (83).

Therefore, we can conclude that $\hat{r}_f^{\theta^*}(x) < \epsilon$ with probability larger than $1-\delta$. By definition of $\hat{\sigma}_f$ this implies that $|\hat{\sigma}_f - \ell_{\text{min}}| < \epsilon$ with probability larger than $1-\delta$, which concludes the proof. □

Note that empirically one can see in Figure 13 (in the areas with many input training points) how well NOMU fulfills D5 (Vanishing) in real-world settings. Furthermore, on can see that the statement only holds true and is only desirable for x in the support of the input data generating distribution $\mathbb{P}_X$ (not in the gaps).

E. NOMU vs. Prior Networks

In this Section, we highlight several differences of NOMU compared to *prior regression networks* that were recently introduced in a working paper by Malinin et al. (2020a).

While the high level idea of introducing a separate loss term for the in-sample-distribution and *out-of-distribution (OOD)* distribution is related to NOMU, there are several important differences, which we discuss next:

1. Malinin et al. (2020a)'s approach focuses on estimating both model and data noise uncertainty. Thus, to properly compare it to NOMU, we consider their approach for known and negligible data noise uncertainty, e.g. for $\sigma_{\text{n}} = 10^{-10}$ we need to set in their paper $(L, \nu) = (I \cdot l^{-1}, \frac{1}{\sigma_{\text{n}}} l)$ with $l \to \infty$, such that the their corresponding model uncertainty prediction is given by $(\kappa(x)\Lambda(x))^{-1} \overset{l\to\infty}{=} \frac{\sigma_{\text{n}}}{\kappa(x)} \cdot I$. In the following, we will consider for simplicity a one-dimensional output, i.e. $\hat{\sigma}_f = \frac{\sigma_{\text{n}}}{\kappa(x)}$.
2. They explicitly define a prior "target" distribution independent of $x \in X$, which is parametrized by κ_0 (model uncertainty) and m_0 (mean prediction) for OOD input points. Specifically, for OOD input points their mean prediction $\hat{f}$ is pushed towards m_0. In many applications the success of classical mean predictions of deep NNs is evident. In none of these applications there was a term that pushed the mean prediction to a fixed predefined prior mean. Therefore, for NOMU we keep the mean prediction untouched.
3. Instead of our loss, their loss (derived from a reverse KL-divergence) is of the form:

$$\underbrace{\sum_{i=1}^{n^{\text{train}}} \frac{\left(\hat{f}(x_i^{\text{train}}) - y_i^{\text{train}}\right)^2}{2\left(\sigma_{\text{n}}\right)^2} + \sum_{i=1}^{n^{\text{train}}} \frac{1}{\kappa(x)}}_{\text{in-sample}} + \tag{92}$$

$$\underbrace{\int_X \frac{\kappa_0 \left(\hat{f}(x) - m_0\right)^2}{2\left(\sigma_{\text{n}}\right)^2} + \frac{\kappa_0}{\kappa(x)} - \log\frac{\kappa_0}{\kappa(x)} - 1 \, d\mu_{\text{OOD}}(x)}_{\text{out-of-distribution}},$$

   where m_0, κ_0 are the prior parameters for the mean and model uncertainty prediction and μ_{OOD} is an OOD measure. Specifically, they enforce zero model uncertainty at input training points via the linear term $\frac{1}{\kappa(x)}$, while we use a quadratic term. Moreover, they only use an OOD term and no *out-of-sample (OOS)* term (see below Item 4).
4. In their loss formulation in (92), they only use an *out-of-distribution* (OOD) term, while we use an *out-of-sample (OOS)* term. By OOD they refer to input training points *only* far away from the training data, e.g., in (Malinin et al., 2020a, Section 3) μ_{OOD} only has support far away from the convex hull of the input training points. Thus, they do not enforce model uncertainty in gaps between input training points. In contrast, by *out-of-sample (OOS)* we refer to a distribution with no mass on the input training points, i.e. we sample new input points that are not equal to the input training points but come from the same range (we recommend to use a distribution that is similar to the data generating process). Therefore, our loss explicitly also enforces model uncertainty in gaps between input training points.
5. They use a different architecture and train only *one* NN. This implies that their mean prediction $m(x)$ can be influenced in unwanted ways by the model uncertainty prediction $(\kappa(x)\Lambda(x))^{-1}$.
6. Their theoretical motivation substantially differs from ours: They only partially specify a prior belief by defining *marginal* distributions for $f(x)$ for each input point

$x \in X$, without specifying a joint prior distribution for f. However, given only marginals no joint distribution, which is the crucial aspect when defining a prior in regression, can be derived without further information (E.g. consider Gaussian processes (GPs); here all one-dimensional marginal distributions are simply given by $\mathcal{N}((m(x)), k(x,x))$. However, the crucial part is how to define $k(x, x')$ specifying the relation of x and x'. Only defining the marginals does not suffice to fully define GPs, leaving this most crucial part undefined).
However, for NOMU, we give in Appendix A a theoretical connection to BNNs, with Gaussian prior on the weights. This induces a prior on the function space, i.e., a distribution over f rather than separate *marginal* distributions over $f(x)$ for each $x \in X$.

7. Parametrizing the model precision instead of the model uncertainty can have negative effects due to (implicit) regularization of NNs in the case of negligible or zero data noise. To get uncertainties in gaps between the input training points (small $\kappa(x)$) while having almost zero uncertainty at these input training points ($\kappa(x) \to \infty$), would imply very high regularization costs for the function $\kappa(x)$ and thus is very hard to learn for a NN. For NOMU, we therefore parameterize directly the model uncertainty (which is always finite) instead of the model precision (that should be infinite at noiseless training data points).

8. Our experimental results suggest that NOMU clearly outperforms DE in BO, whilst DE outperforms *prior regression networks* in their set of experiments.

F. NOMU vs. Neural Processes

In this section, we discuss the differences between neural processes (NPs) introduced by Garnelo et al. (2018a;b) and NOMU. Specifically, we explain in the following why NPs and NOMU are solving very different problems in different settings.

For training NPs, one has to observe data from 1000s of realizations of f_k, sampled i.i.d. from the prior distribution (for each f_k one observes x_i and $f_k(x_i)$ to train the NP). This is often mentioned in Garnelo et al. (2018a;b), e.g., Garnelo et al. (2018a, Section 4.1): *"We generate [...] datasets that consist of functions generated from a GP [...]* ***At every training step we sample a curve from the GP*** *[...]."*. For NOMU we consider the **very different** task of estimating $p(y^{\text{test}}|x^{\text{test}}; D^{\text{train}})$ based on a **single dataset**, i.e, generated from a single realization $f = f_1$.

For example in the case of the Boston housing data set from Section 4.1.4, there is only one function $f = f_1$ involved that maps a (multidimensional) input data point x corresponding to a house in Boston to its price $f(x)$. For this data set, NPs would not be well suited, since it only contains data $(x_i, y_i) = (x_i, f(x_i) + \varepsilon_i)$ coming from this specific function f. One does not have access to data corresponding to another function f_2 that had been sampled from the same prior distribution.

The same is true for the other data sets we consider in this paper (e.g., for the solar irradiance data set we only use the data visible in Figure 5 and NOMU does not have access to any data coming from other time series to make its predictions). Thus, NPs cannot be applied to the tasks considered in this paper.

Summary. NOMU, GP, MCDO, DE and HDE are designed to be trained on data coming from *one* unknown function f without having access to data from other functions $f_2, f_3, \ldots$. In contrast, NPs are designed to be trained on *multiple* data sets generated from *multiple* functions $f_1, f_2, f_3, \ldots$.

G. Aleatoric Neural Networks: Aleatoric vs. Epistemic Uncertainty

In this section, we discuss the classical approach of a NN with two outputs, one output for a model prediction and another for aleatoric uncertainty, which is trained using the (scaled) Gaussian negative log-likelihood as introduced by Nix & Weigend (1994). We will use the terms aleatoric uncertainty and data noise as well as model uncertainty and epistemic uncertainty interchangeably.

In what follows, we call this method *aleatoric neural network (ANN)*. Within this section, we show that such an ANN does not explicitly estimate model uncertainty (in contrast to all other benchmark methods discussed in this paper), i.e., when using the aleatoric uncertainty output $\hat{\sigma}_{\text{n}}$ naively as $\hat{\sigma}_f$, the so obtained $\hat{\sigma}_f := \hat{\sigma}_{\text{n}}$ does not represent model uncertainty (epistemic uncertainty).[39] First, we give the definition of an ANN.

Definition G.1 (ANN) *An ANN is a fully-connected feed-forward NN $\mathcal{NN}^{\text{ANN}} : \mathbb{R}^d \to \mathbb{R} \times \mathbb{R}_+$ with two outputs: (i) the model prediction $\hat{f} \in \mathbb{R}$ and (ii) a model uncertainty prediction $\hat{\sigma}_n \in \mathbb{R}_+$ that is trained for a given set of training points D^{train} using the following loss function:*

$$L^{\text{ANN}}(\mathcal{NN}^{\text{ANN}}) := \frac{1}{|D^{\text{train}}|} \sum_{(x,y)\in D^{\text{train}}} \left[\frac{\left(y - \hat{f}(x)\right)^2}{2\left(\hat{\sigma}_n(x)\right)^2} + \ln\left(\hat{\sigma}_n(x)\right) \right] \tag{93}$$

The idea of ANN is to estimate the noise scale $\sigma_{\text{n}}(x) = \sqrt{\mathbb{V}[\varepsilon]} = \sqrt{\mathbb{V}[y|x, f(x)]}$. Figure 21 shows that, as we

[39] The reminder of this section only targets readers who do not directly see that substituting $\hat{\sigma}_f$ by $\hat{\sigma}_{\text{n}}$ is an extremely bad idea.

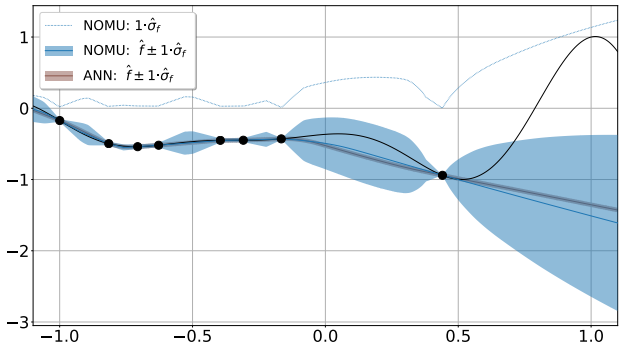


Figure 21. Comparison of UBs resulting from ANN ($c = 1$) (where $\hat{\sigma}_\text{n}$ is used as a substitute for $\hat{\sigma}_f$) vs. NOMU for the Forrester function (solid black line). For NOMU, we also show $\hat{\sigma}_f$ as a dotted blue line. Training points are shown as black dots.

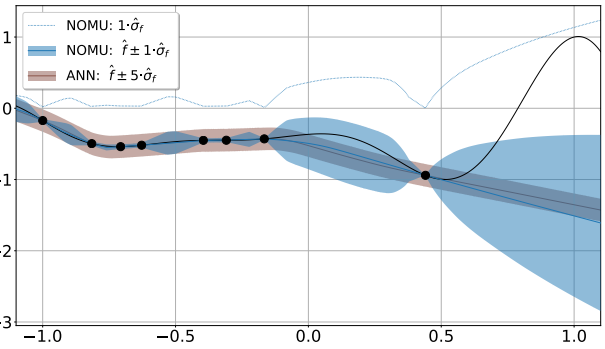


Figure 22. Comparison of UBs resulting from ANN ($c = 5$) (where $\hat{\sigma}_\text{n}$ is used as a substitute for $\hat{\sigma}_f$) vs. NOMU for the Forrester function (solid black line). For NOMU, we also show $\hat{\sigma}_f$ as a dotted blue line. Training points are shown as black dots.

would expect, the trained ANN has learned the true $\sigma_\text{n} \equiv 0 \approx \hat{\sigma}_\text{n}$ quite precisely. However, it as also becomes evident that the ANN does not learn any form of model uncertainty. Very far away from all observed training data points, the ANN does not express any uncertainty about its prediction (in Figure 22, to the right of $x = 0.5$, the predictions are very far away from the truth, but $\hat{\sigma}_\text{n}$ does not capture this uncertainty). Therefore, the ANN's aleatoric uncertainty output $\hat{\sigma}_\text{n}$ does not fulfill desideratum D3 (Out-of-Sample). The problem of misusing $\hat{\sigma}_\text{n}$ as substitute for σ_f is not that $\hat{\sigma}_\text{n}$ is too small (as we study relative uncertainty in this paper (see Appendix A.2.1), one can always scale the uncertainty by a factor c). However, Figure 22 shows that also the scaled uncertainty completely fails to capture the desiderata, i.e., the aleatoric uncertainty output $\hat{\sigma}_\text{n}$ is almost constant for all input points x. Thus, UBs of an ANN do not fulfill D2 (In-Sample) and D3 (Out-of-Sample). In Figure 22, we can see that $5\hat{\sigma}_\text{n}$ is way too underconfident at input training points and at the same time way too overconfident far away from the observed input training points. This would result in very bad NLL scores on a test set. (Moreover in noisy settings, i.e., $\sigma_\text{n} \neq 0$, the aleatoric uncertainty output $\hat{\sigma}_\text{n}$ of an ANN does not fulfill D5 (Vanishing) either: $\hat{\sigma}_\text{n}$ should converge to σ_n while $\hat{\sigma}_f$ should converge to zero as the amount of training data increases.)

Furthermore, especially in Bayesian optimization (BO) $c\hat{\sigma}_\text{n}$ would be a terribly bad substitute for σ_f: Maximizing the upper UB acquisition function $\hat{f} + c\hat{\sigma}_\text{n}$, would be almost equivalent to maximizing $\hat{f}$ since $c\hat{\sigma}_\text{n}$ is almost constant because of the lack of D2 (In-Sample) and D3 (Out-of-Sample). If one wants to maximize the function in Figure 22 on $[-1, 1]$, the next BO-step would propose to query an input training point at the left boundary $x = -1$ (even for large c). However, one does not learn anything new from evaluating at $x = -1$, because this input training point has already been evaluated in a previous BO-step. All subsequent BO steps would propose to query the same point $x = -1$ without exploring any other region of the input space and one would never find the true maximum at $x \approx 1$. In contrast, a reasonable model for estimating σ_f (such as NOMU), would directly (after scaling up c dynamically by a factor 2 as described in Appendix B.3.2) choose a point in the unexplored right region $x \approx 1$, because the left side $x \approx -1$ is already well explored.

Overall, aleatoric uncertainty σ_n and epistemic uncertainty σ_f are two very different objects. Thus, an estimator $\hat{\sigma}_\text{n}$ designed to estimate σ_n is usually a bad estimator for σ_f.

H. Hyperparameter Sensitivity Analysis

In this section, we provide a sensitivity analysis with respect to NOMU's loss hyperparameters, i.e., π_sqr, π_exp, c_exp, and D^art. First, we present a visual qualitative analysis in 1D showing how each of these hyperparameters affects the shape of NOMU's UBs (Appendix H.1). Second, we also present an extensive quantitative sensitivity analysis in the generative test-bed setting from Section 4.1.2, where in addition to the loss hyperparameters we also include the hyperparameters of the readout map ℓ_min, and ℓ_max in our analysis (Appendix H.2).

H.1. Qualitative Sensitivity Analysis

In this section, we consider the setting of Section 4.1.1, and visualize the effect of increasing or decreasing each of NOMU's loss hyperparameters π_sqr, π_exp, c_exp, and D^art on the example of the 1D Levy function. For reference, Figure 23 shows NOMU's UBs (with scaling factor $c = 2$) for the default loss hyperparameters $\pi_\text{sqr} = 0.1$, $\pi_\text{exp} = 0.01$, $c_\text{exp} = 30$, and $D^\text{art} = 128$ that are used in Section 4.1.1 in the main paper.

For each of D^art and c_exp, we fit two additional NOMU mod-

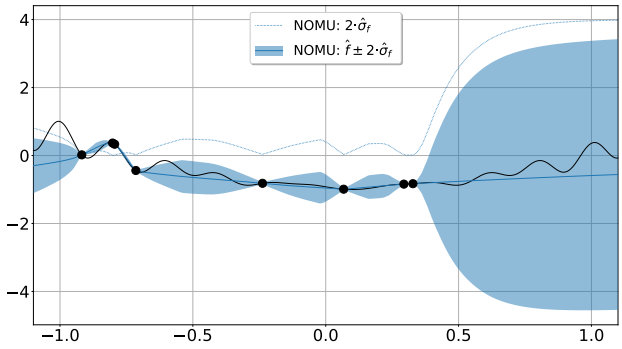


Figure 23. NOMU's UBs (c=2) for the generic loss hyperparameters from Section 4.1.1 in the main paper.

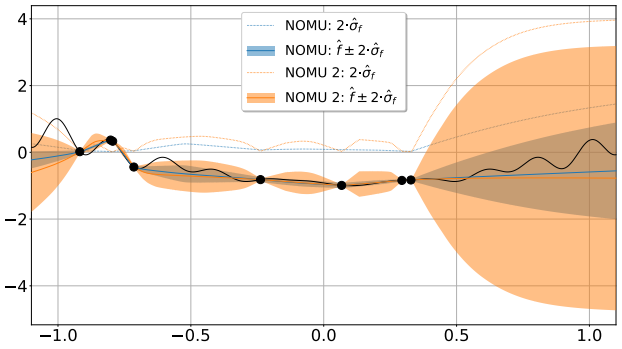


Figure 24. NOMU's UBs (c=2) for $\pi_{\text{sqr}} = 1e-4, \pi_{\text{exp}} = 1e-5$ (blue) and $\pi_{\text{sqr}} = 1, \pi_{\text{exp}} = 0.1$ (orange).

els, where we ceteris paribus de- and increase the hyperparameter's default value by factors $1/s$ and s, respectively. The multiplicative factors π_{sqr} and π_{exp} we treat jointly in our sensitivity analysis: First, we show the effect of ceteris paribus de- and increasing the default value of the product $\pi_{\text{exp}}\pi_{\text{sqr}}$ by factors $s_l = 0.001$ and $s_u = 10$. Second, we vary the ratio $\pi_{\text{exp}}/\pi_{\text{sqr}}$ in the same fashion, with scaling factors $s_l = 1/16$ and $s_u = 16$. Within all of the experiments in this section, we sample artificial input points D^{art} on an equidistant (deterministic) grid on $[-1.1, 1.1]$. This allows us to give a qualitative analysis of the hyperparameters' effects as follows.

Varying $\pi_{\text{exp}}\pi_{\text{sqr}}$ with Scaling Factors of $s_l = 0.001$ and $s_u = 10$. Decreasing $\pi_{\text{exp}}\pi_{\text{sqr}}$ by decreasing both π_{exp} and π_{sqr} leads to more tubular bounds by relaxing the desiderata D2 (In-Sample) and D3 (Out-of-Sample). This can be seen in Figure 24: NOMU's blue dashed uncertainty (corresponding to small $\pi_{\text{exp}}\pi_{\text{sqr}}$) is larger at data points than the orange one of NOMU 2 (corresponding to large $\pi_{\text{exp}}\pi_{\text{sqr}}$), and it is smaller further away from the training data points. NOMU's default hyperparameters are in a range where the loss is already at its limit[40] enforcing the desiderata D2 (In-Sample) and D3 (Out-of-Sample). Therefore, further increasing $\pi_{\text{exp}}\pi_{\text{sqr}}$ (while keeping their ratio and the other hyperparameters fixed) barely causes the UBs to change as can be seen for the orange UBs in Figure 24. Increasing $\pi_{\text{exp}}\pi_{\text{sqr}}$ too much, can lead to numerical instabilities.

Varying $\pi_{\text{exp}}/\pi_{\text{sqr}}$ with Scaling Factors of $s_l = 1/16$ and $s_u = 16$. Increasing the ratio of $\pi_{\text{exp}}/\pi_{\text{sqr}}$ (while keeping their product and all other hyperparameters fixed) simply causes NOMU's UBs to uniformly widen across the entire domain. Indeed, in Figure 25, the orange UBs of NOMU 2 (corresponding to large $\pi_{\text{exp}}/\pi_{\text{sqr}}$) are blown up and cover the blue UBs (corresponding to small $\pi_{\text{exp}}/\pi_{\text{sqr}}$) throughout the input space.

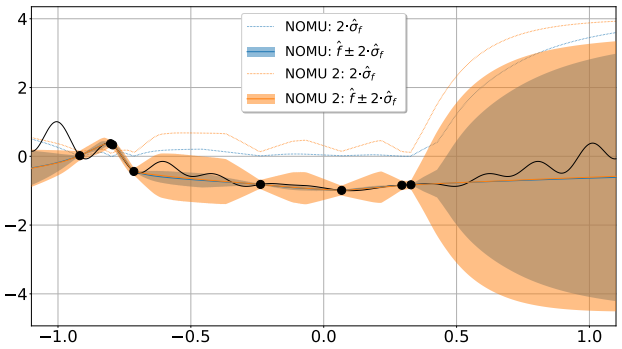


Figure 25. NOMU's UBs (c=2) for $\pi_{\text{exp}} = 0.0025, \pi_{\text{sqr}} = 0.4$ (blue) and $\pi_{\text{exp}} = 0.04, \pi_{\text{sqr}} = 0.025$ (orange).

Varying c_{exp} with a Scaling Factor of $s = 2$. Increasing the hyperparameter c_{exp} causes UBs to shrink in areas of large uncertainty and to widen in areas of small uncertainty. This effect is visualized in Figure 26: the orange dashed uncertainty line of NOMU 2 (large $c_{\text{exp}} = 60$) lies above the blue one of NOMU (small $c_{\text{exp}} = 15$) at data points; and in regions of large uncertainty, the orange UBs (corresponding to large c_{exp}) turn out to be more narrow than the blue ones (corresponding to small c_{exp}). Thus, increasing c_{exp} causes NOMU's UBs to be more tubular.

Varying D^{art} with a Scaling Factor of $s = 8$. Finally, we assess the effect of changing the number of artificial data

[40] For NOMU's default parameters the ratio $\pi_{\text{exp}}\pi_{\text{sqr}}/\lambda$ is already very large such that the explicit regularization via $\lambda \|\theta\|_2^2$ is almost negligible. Thus, the UBs are only prevented from having even larger curvature by implicit regularization, i.e., within a given number of epochs the training algorithm cannot reach a function with more curvature, because increasing the amplitude of the loss is (partially) compensated by the adaptivity of the ADAM algorithm. Only in ranges where $\pi_{\text{exp}}\pi_{\text{sqr}}/\lambda$ is small enough for the explicit regularization to actually matter, the UBs become sensitive to the ratio $\pi_{\text{exp}}\pi_{\text{sqr}}/\lambda$. Then the regularization of $\lambda \|\theta\|_2^2$ keeps the curvature of $\hat{\sigma}_f$ low, i.e., the UBs become more tubular.

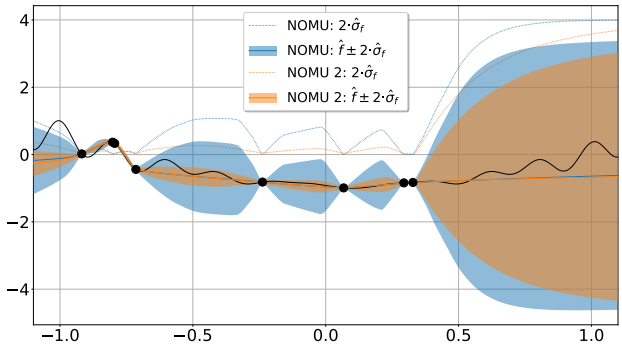


Figure 26. NOMU's UBs (c=2) for $c_{\text{exp}} = 15$ (blue) and $c_{\text{exp}} = 60$ (orange).

points D^{art} used to approximate the integral (c) of NOMU's loss function defined in Equation (4). As expected, the UBs behave overall very similarly. However, for very small gaps in between input training points, D^{art} can have an influence on the estimated UBs. For example, in the gap between the training input point at $x \approx -0.77$ and the one at $x \approx -0.73$, the $\hat{\sigma}_f$ obtained from the smaller D^{art} (blue) vanishes (because of a lack of artificial data points falling in this gap), while $\hat{\sigma}_f$ obtained from the larger D^{art} also estimates non-zero model uncertainty in this small gap.

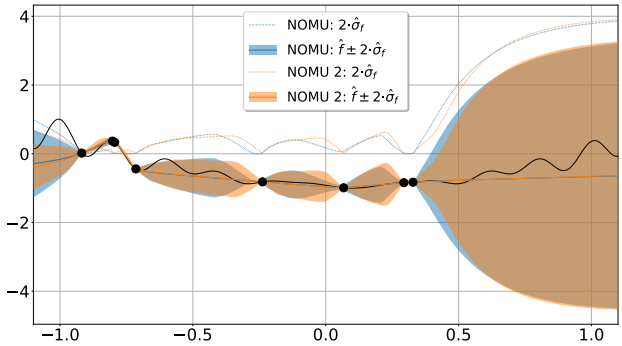


Figure 27. NOMU's UBs (c=2) for $D^{\text{art}} = 16$ (blue) and $D^{\text{art}} = 1024$ (orange).

Varying the architecture and L2-regularization In Appendix D.4.1 we visualize how certain changes to the architecture and different L2-regularization parameters λ for different parts of the network influence NOMU's model uncertainty estimate $\hat{\sigma}_f$. In particular, we show in Appendix D.4.1 how the choice of the architecture and the L2-regularization determine the degree to which NOMU fulfills desiderata D4 (Metric Learning).

Finally, in Appendix H.2, we empirically show that NOMU is robust with respect to its hyperparameters within a certain range.

H.2. Quantitative Sensitivity Analysis

We now present an extensive quantitative sensitivity analysis of NOMU's loss hyperparameters: π_{sqr}, π_{exp}, c_{exp}, and D^{art} in the generative test-bed setting (see Section 4.1.2 for details on this setting). Additionally, we also consider in this analysis the hyperparameters corresponding to the readout map, i.e.., ℓ_{min} and ℓ_{max}. We decided to perform the quantitative sensitivity analysis in the generative test-bed setting, since it offers a particularly rich variety of different test functions and thus exposes each hyperparameter selection to hundreds of different test functions.

Setting We use the same default hyperparameters as in Section 4.1.2. This includes NOMU's loss hyperparameters: π_{sqr}, π_{exp}, c_{exp}, and D^{art}, the hyperparameters of the readout map: ℓ_{min} and ℓ_{max} as well as all other hyperparameters. For the following sensitivity analysis, we then vary at each time only a *single* hyperparameter, i.e., one of π_{sqr}, π_{exp}, c_{exp}, D^{art}, ℓ_{min}, and ℓ_{max}, and set all other hyperparameters to their default values (i.e., we perform a *ceteris paribus* analysis as in Appendix H.1).

In Table 15, we present for each considered hyperparameter a grid of five different values which we use to test its sensitivity. The **NOMU** column in Table 15 corresponds to the NOMU's default hyperparameters used in the generative test-bed setting (Section 4.1.2). The columns **NOMU1** to **NOMU4** in Table 15 then correspond to deviations from these original hyperparameters.

Results In Table 16, Table 17, and Table 18 we present for each of the ceteris paribus runs average NLL values for input dimensions 1D, 2D, and 5D, respectively. Each cell in those tables represents a single hyperparameter selection where we use NOMU's default hyperparameters except for the hyperparameter of the corresponding row which we choose according to the cell's column, e.g., to obtain the result for the cell (π_{exp}, **NOMU3**), we use the default NOMU hyperparameters from Section 4.1.2 *except* for π_{exp}

Table 15. Grid selection for each hyperparameter (HP). The column NOMU corresponds to NOMU's default hyperparameters used in the generative test-bed setting. NOMU1 to NOMU4 correspond to deviations from these default hyperparameters. For D^{art}, d denotes the input dimension.

HP	NOMU1	NOMU2	NOMU	NOMU3	NOMU4
π_{SQR}	0.01	0.02	0.1	0.5	1
π_{EXP}	0.001	0.002	0.01	0.05	0.1
c_{EXP}	10	20	30	45	90
D^{ART}	$\frac{100\cdot d}{4}$	$\frac{100\cdot d}{2}$	$100\cdot d$	$2\cdot(100\cdot d)$	$4\cdot(100\cdot d)$
ℓ_{MIN}	0.01	0.05	0.1	0.15	0.20
ℓ_{MAX}	0.50	0.75	1	2.0	4.0

which we set according to column **NOMU3** in Table 15 to $\pi_{\text{exp}} := 0.05$. Overall, we can make the following three main observations:

1. The majority of all cells in Table 16, Table 17, and Table 18 are marked in grey. This shows that, their corresponding hyperparameters lead to average NLL results which are statistically on par with the results obtained via NOMU's default hyperparameters in the **NOMU** columns. This highlights NOMU's robustness with respect to all considered hyperparameters within the chosen grids. Furthermore, it confirms our claim from the main paper that using generic hyperparameters for NOMU often works well without much hyperparameter-tuning.

2. Since NOMU with the default hyperparameters already outperforms *all* other considered benchmark methods in this setting, i.e., each **NOMU** column represents the winning method among benchmarks (see Table 2), we see that all grey marked deviations of hyperparameters lead to results that outperform all other considered benchmark methods too. Moreover, all except for one (5D (ℓ_{min}, **NOMU1**)) NOMU models corresponding to cells which are not marked in grey, i.e., with hyperparameters that lead to statistically worse results than NOMU's default hyperparameters, are as good or better than the best benchmark methods from Table 2.

3. By varying NOMU's hyperparameters, we can even obtain better results (i.e., with a smaller average NLL) than the ones reported in Table 2 of the main paper, e.g. in 1D with $\ell_{\text{min}} := 0.01$ the average NLL $= -1.83 < -1.65$. While these improvements are not statistically significant, these results still suggest that systematic hyperparameter-tuning could improve the performance of NOMU even further.

Table 16. Sensitivity analysis for 1D generative test-bed setting. We present for each hyperparameter (HP) and its five corresponding grid-points the average NLL (without const. $\ln(2\pi)/2$) and a 95% CI over 200 BNN samples. Results which are statistically on par with NOMU's default HPs, i.e., the column **NOMU**, are marked in grey. Note that, the best benchmark method for this experiment is GP with NLL $= -1.08 \pm 0.22$ (see Table 2).

HP	NOMU1	NOMU2	NOMU	NOMU3	NOMU4
π_{SQR}	-1.59±0.11	-1.63±0.10	-1.65±0.10	-1.55±0.13	-1.54±0.14
π_{EXP}	-1.64±0.09	-1.67±0.09	-1.65±0.10	-1.62±0.10	-1.60±0.10
c_{EXP}	-1.77±0.12	-1.73±0.10	-1.65±0.10	-1.49±0.12	-1.11±0.13
D^{ART}	-1.65±0.09	-1.62±0.11	-1.65±0.10	-1.63±0.12	-1.65±0.10
ℓ_{MIN}	-1.83±0.09	-1.72±0.10	-1.65±0.10	-1.58±0.10	-1.47±0.12
ℓ_{MAX}	-1.49±0.12	-1.60±0.10	-1.65±0.10	-1.63±0.16	-1.66±0.14

Table 17. Sensitivity analysis for 2D generative test-bed setting. We present for each hyperparameter (HP) and its five corresponding grid-points the average NLL (without const. $\ln(2\pi)/2$) and a 95% CI over 200 BNN samples. Results which are statistically on par with NOMU's default HPs, i.e., the column **NOMU**, are marked in grey. Note that, the best benchmark method for this experiment is DE with NLL $= -0.77 \pm 0.07$ (see Table 2).

HP	NOMU1	NOMU2	NOMU	NOMU3	NOMU4
π_{SQR}	-1.18±0.04	-1.18±0.04	-1.16±0.05	-1.15±0.04	-1.15±0.04
π_{EXP}	-1.15±0.04	-1.15±0.05	-1.16±0.05	-1.18±0.04	-1.18±0.04
c_{EXP}	-1.17±0.04	-1.19±0.04	-1.16±0.05	-1.11±0.05	-1.00±0.05
D^{ART}	-1.16±0.05	-1.16±0.05	-1.16±0.05	-1.16±0.04	-1.15±0.05
ℓ_{MIN}	-1.07±0.05	-1.17±0.04	-1.16±0.05	-1.14±0.05	-1.12±0.05
ℓ_{MAX}	-1.13±0.05	-1.15±0.05	-1.16±0.05	-1.16±0.04	-1.16±0.04

Table 18. Sensitivity analysis for 5D generative test-bed setting. We present for each hyperparameter (HP) and its five corresponding grid-points the average NLL (without const. $\ln(2\pi)/2$) and a 95% CI over 200 BNN samples. Results which are statistically on par with NOMU's default HPs, i.e., the column **NOMU**, are marked in grey. Note that, the best benchmark method for this experiment is GP with NLL $= -0.33 \pm 0.02$ (see Table 2).

HP	NOMU1	NOMU2	NOMU	NOMU3	NOMU4
π_{SQR}	-0.37±0.03	-0.37±0.02	-0.37±0.02	-0.36±0.02	-0.35±0.02
π_{EXP}	-0.33±0.02	-0.34±0.02	-0.37±0.02	-0.40±0.02	-0.40±0.02
c_{EXP}	-0.39±0.02	-0.38±0.02	-0.37±0.02	-0.37±0.02	-0.34±0.05
D^{ART}	-0.37±0.02	-0.37±0.02	-0.37±0.02	-0.37±0.02	-0.37±0.02
ℓ_{MIN}	-0.21±0.03	-0.33±0.02	-0.37±0.02	-0.38±0.02	-0.39±0.02
ℓ_{MAX}	-0.39±0.02	-0.37±0.02	-0.37±0.02	-0.35±0.02	-0.33±0.02

I. Details on our Notation

In Section 2, we are using a slightly overloaded notation, where we use the same symbol f for different mathematical objects. Sometimes, we use f for a function-valued random variable $F : (\Omega, \Sigma, \mathbb{P}) \to Y^X$. and sometimes we use f for the specific unknown ground truth function $f_{\text{true}} := F(\omega)$ (i.e., $\forall x \in X : f_{\text{true}}(x) = (F(\omega))(x)$). While we used this slightly overloaded notation for the sake of readability in the main paper, we will now introduce our Bayesian uncertainty framework in its full mathematical detail:

In practice, there exists an unknown ground truth function f_{true}. In the classical Bayesian paradigm, one assumes that everything unknown (i.e., here f_{true}) was sampled from a prior distribution. Specifically, $f_{\text{true}} := F(\omega)$ is a realization of a random variable $F : (\Omega, \Sigma, \mathbb{P}) \to Y^X$ distributed according to a prior belief, i.e., a prior distribution. Using this notation, one can mathematically describe in a rigorous way the full Bayesian data generating process as follows:

Let $F : (\Omega, \Sigma, \mathbb{P}) \to Y^X$ be a function-valued random variable distributed according to a prior belief, i.e., a prior

distribution. Moreover, let $(\mathcal{X}_i, \mathcal{E}_i)_{i\in\{1,\dots,n^{\text{train}}\}}$ denote the random variable representing the input data points and the corresponding data noise, i.e., a family off i.i.d random variables independent of F, where each random variable $(\mathcal{X}_i, \mathcal{E}_i) : (\Omega, \Sigma, \mathbb{P}) \to X \times \mathbb{R}$ fulfills $\forall x \in X : (\mathcal{E}_i | \mathcal{X}_i = x) \sim \mathcal{N}(0, \sigma_{\text{n}}(x))$. Finally, let $\mathcal{Y}_i$ be the random variable associated to the targets, which we define as $\mathcal{Y}_i : (\Omega, \Sigma, \mathbb{P}) \to Y, \omega \mapsto \mathcal{Y}_i(\omega) := F(\omega)(\mathcal{X}_i(\omega)) + \mathcal{E}_i(\omega)$.

With this notation in place, the objects used in the main paper in Section 2: $x_i = \mathcal{X}_i(\omega)$, $y_i = \mathcal{Y}_i(\omega)$ and $\varepsilon_i = \mathcal{E}_i(\omega)$ are the values one actually observe in practice as training data, i.e., the realizations of the data generating process.

Therefore, $\sigma_f(x)$ from Equation (1) should be interpreted for all $x \in X$ as

$$\sigma_f(x) = \sqrt{\mathbb{V}\left[F(\cdot)(x) | \forall i \in \{1, \dots, n^{\text{train}}\} : (\mathcal{X}_i, \mathcal{Y}_i) = (x_i, y_i)\right]}, \tag{94}$$

which, for all $x \in X$, is defined mathematically even more rigorously via the conditional variance as follows:

$$\sqrt{\mathbb{V}\left[F(\cdot)(x) \middle| (\mathcal{X}_i, \mathcal{Y}_i)_{i\in\{1,\dots,n^{\text{train}}\}}\right](\omega)}. \tag{95}$$

Throughout the paper it should always be clear from the context if x_i refers to the random variable $\mathcal{X}_i$ or its realization $\mathcal{X}_i(\omega)$ where the same holds true for y_i, ε_i and f.

Importantly, note that in the setting which we consider in this paper, one only observes data coming from a *single* function $f = F(\omega)$ and one does not have access to more functions f_i. I.e., we neither have access to other samples of the random variable F nor do we consider multiple i.i.d random variables F_i in contrast to the setting considered for neural processes as described in Appendix F.