ON THE PITFALLS OF HETEROSCEDASTIC UNCERTAINTY ESTIMATION WITH PROBABILISTIC NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Capturing aleatoric uncertainty is a critical part of many machine learning systems. In deep learning, a common approach to this end is to train a neural network to estimate the parameters of a heteroscedastic Gaussian distribution by maximizing the logarithm of the likelihood function under the observed data. In this work, we examine this approach and identify potential hazards associated with the use of log-likelihood in conjunction with gradient-based optimizers. First, we present a synthetic example illustrating how this approach can lead to very poor but stable parameter estimates. Second, we identify the culprit to be the log-likelihood loss, along with certain conditions that exacerbate the issue. Third, we present an alternative formulation in which each data point's contribution to the loss is weighted by the β-exponentiated variance estimate. We show that using an appropriate β largely mitigates the issue in our illustrative example. Fourth, we evaluate this approach on a range of standard benchmarks from the literature and show that it achieves considerable improvements and performs more robustly with respect to hyperparameters, both in predictive RMSE and log-likelihood criteria.

1 INTRODUCTION

Endowing models with the ability to capture *uncertainty* is of crucial importance in machine learning. Uncertainty can be usefully categorized into two main types: *epistemic* uncertainty and *aleatoric* uncertainty (Kiureghian & Ditlevsen, 2009). Epistemic uncertainty accounts for subjective uncertainty in the model, one that is reducible given sufficient data. By contrast, aleatoric uncertainty captures the stochasticity inherent in the observations and can itself be subdivided into *homoscedastic* and *heteroscedastic* uncertainty. Homoscedastic uncertainty is noise that stays constant across different inputs, whereas heteroscedastic uncertainty is one that varies depending on the inputs to the model.

There are well-established benefits for modeling each type of uncertainty. For instance, capturing epistemic uncertainty enables effective budgeted data collection in active learning (Gal et al., 2017), allows for efficient exploration in reinforcement learning (Osband et al., 2016), and is indispensable in cost-sensitive decision making (Amodei et al., 2016). On the other hand, quantifying aleatoric uncertainty enables learning dynamics models of stochastic processes (e.g. for model-based or offline reinforcement learning) (Chua et al., 2018; Yu et al., 2020), improves performance in semantic segmentation, depth regression and object detection (Kendall & Gal, 2017; Harakeh & Waslander, 2021), and allows for risk-sensitive decision making (Dabney et al., 2018; Vlastelica et al., 2021).

Many modern applications of machine learning rely on neural networks and deep learning tools to achieve state-of-the-art performance. However, most neural network models are not readily equipped with a capacity for uncertainty estimation. To address this shortcoming, a multitude of approaches have been proposed in the past few years. Nevertheless, the majority of work in this domain has focused on epistemic uncertainty quantification (Blundell et al., 2015; Gal & Ghahramani, 2016) or uncertainty estimation for classification (Hendrycks & Gimpel, 2017; Guo et al., 2017; Ovadia et al., 2019; Mukhoti et al., 2021).

In this work, we examine a common approach for quantifying aleatoric uncertainty in neural network regression. By assuming that the regression targets follow a particular distribution, we can use a neural network to predict the parameters of that distribution, typically the input-dependent mean and variance when assuming a heteroscedastic Gaussian distribution. Then, the parameters of the network can be learned using maximum likelihood estimation (MLE), i.e. by minimizing the negative

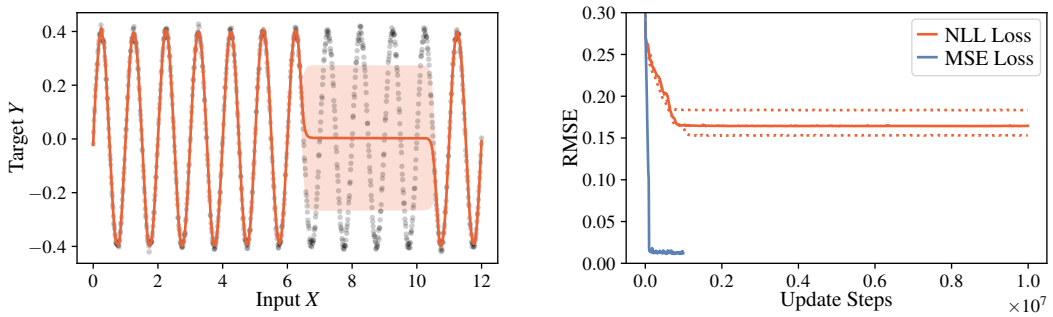

Figure 1: Training a probabilistic neural network to fit a simple sinusoidal fails. Left: learned predictions (orange line) after 10^7 updates, with the shaded region showing the predicted standard deviation. The target function is given by $y(x) = 0.4\sin(2\pi x) + \xi$, where ξ is Gaussian noise with a standard deviation of 0.01. Right: root mean squared error (RMSE) over training. Dotted lines show runs with different random seeds. For comparison, we also plot the training curve when using the mean squared error as the training objective — achieving a perfect mean fit in 10^5 updates.

log-likelihood (NLL) criterion using stochastic gradient descent. This simple procedure, which is the de-facto standard (Nix & Weigend, 1994; Lakshminarayanan et al., 2017; Kendall & Gal, 2017; Chua et al., 2018; Kloss et al., 2021), is known to be subject to overconfident variance estimates. Whereas strategies have been proposed to alleviate this specific issue (Detlefsen et al., 2019; Hu et al., 2020; Stirn & Knowles, 2020), we argue that an equally important issue is that this procedure can additionally lead to subpar mean fits. In this work, we analyse and propose a simple modification to mitigate this issue.

Summary of contributions We demonstrate a pitfall of optimizing the NLL loss for neural network regression, one that hinders the training of accurate mean predictors (see Fig. 1 for an illustrative example). Our analysis identifies the *high dependence of the gradients on the predictive variance* as the primary culprit. More precisely, we hypothesize that the issue arises due to the NLL loss scaling down the gradient of poorly-predicted data points relative to the well-predicted ones, leading to effectively undersampling the poorly-predicted data points.

We then introduce an *alternative loss formulation* that counteracts this by weighting the contribution of each data point to the overall loss by its β-exponentiated variance estimate, *where β controls the extent of dependency of gradients on predictive variance*. This formulation subsumes the standard NLL loss for $\beta = 0$ and allows to lessen the dependency of gradients on the variance estimates for $0 < \beta \leq 1$. Interestingly, using $\beta = 1$ completely removes such dependency for training the mean estimator, yielding the standard mean squared error (MSE) loss – but with the additional capacity of uncertainty estimation. Finally, we empirically show that our modified loss formulation largely mitigates the issue of poor fits, achieving considerable improvements on a set of standard benchmarks while exhibiting more robustness to network architecture and learning rate configurations.

2 Preliminaries

Let X, Y be two random variables describing the input and target, following the joint distribution $P(X, Y)$. We assume that Y is conditionally independent given X and that it follows some probability distribution $P(Y \mid X)$. In the following, we use the common assumption that Y is normally distributed given X; i.e. $P(Y \mid X) = \mathcal{N}(\mu(X), \sigma^2(X))$, where $\mu\colon \mathbb{R}^M \mapsto \mathbb{R}$ and $\sigma^2\colon \mathbb{R}^M \mapsto \mathbb{R}^+$ are respectively the true input-dependent mean and variance functions.[1] Equivalently, we can write $Y = \mu(X) + \epsilon(X)$, with $\epsilon(X) \sim \mathcal{N}(0, \sigma^2(X))$; i.e. Y is generated from X by $\mu(X)$ plus a zero-mean Gaussian noise with variance $\sigma^2(X)$. This input-dependent variance quantifies the heteroscedastic uncertainty, or input-dependent aleatoric uncertainty.

[1] For notational convenience, we focus on univariate regression but point out that the work extends to the multivariate case as well.

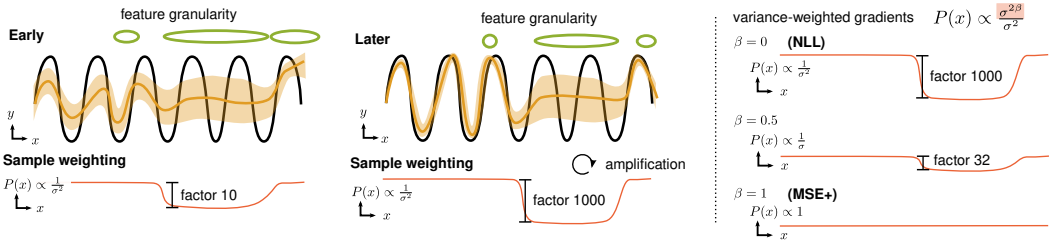

Figure 2: Illustration of the pitfall when training with NLL (negative log-likelihood) and our solution. An initial inhomogeneous feature space granularity (explained in the text) results early on in different fitting quality. The implicit weighting of the squared error in NLL can be seen as a biased data-sampling with $p(x) \sim \frac{1}{\sigma^2(x)}$ (see Eq. 6). Badly fit parts are increasingly ignored during training. On the right, the effect of our solution (Eq. 9) on the relative importance of data points is shown.

To learn estimates $\hat{\mu}(X), \hat{\sigma}^2(X)$ of the true mean and variance functions, it is common to use a neural network f_θ parameterized by θ. Here, $\hat{\mu}(X)$ and $\hat{\sigma}^2(X)$ can be outputs of the final layer (Nix & Weigend, 1994) or use two completely separate networks (Detlefsen et al., 2019).The variance output is hereby constrained to the positive region using a suitable activation function, e.g. softplus. The optimal parameters θ^*_{NLL} can then be found using maximum likelihood estimation (MLE) by minimizing the negative log-likelihood (NLL) criterion $\mathcal{L}_{\text{NLL}}$ under the distribution $P(X, Y)$:

$$\theta^*_{\text{NLL}} = \arg\min_\theta \mathcal{L}_{\text{NLL}}(\theta; \mathcal{D}) = \arg\min_\theta \mathop{\mathbb{E}}_{X,Y}\left[\frac{1}{2}\log\hat{\sigma}^2(X) + \frac{(Y-\hat{\mu}(X))^2}{2\hat{\sigma}^2(X)} + \text{const}\right]. \tag{1}$$

In contrast, standard regression minimizes the mean squared error (MSE) $\mathcal{L}_{\text{MSE}}$:

$$\theta^*_{\text{MSE}} = \arg\min_\theta \mathcal{L}_{\text{MSE}}(\theta; \mathcal{D}) = \arg\min_\theta \mathop{\mathbb{E}}_{X,Y}\left[\frac{(Y-\hat{\mu}(X))^2}{2}\right]. \tag{2}$$

In practice, Eq. 1 and Eq. 2 are optimized using stochastic gradient descent with batches of samples drawn from $P(X, Y)$. The gradients of $\mathcal{L}_{\text{NLL}}$ with respect to $\hat{\mu}(X)$, $\hat{\sigma}^2(X)$ are given by

$$\nabla_{\hat{\mu}}\mathcal{L}_{\text{NLL}}(\theta) = \mathop{\mathbb{E}}_{X,Y}\left[\frac{\hat{\mu}(X)-Y}{\hat{\sigma}^2(X)}\right], \quad \nabla_{\hat{\sigma}^2}\mathcal{L}_{\text{NLL}}(\theta) = \mathop{\mathbb{E}}_{X,Y}\left[\frac{\hat{\sigma}^2(X)-(Y-\hat{\mu}(X))^2}{2\big(\hat{\sigma}^2(X)\big)^2}\right]. \tag{3, 4}$$

3 Analysis

We now return to the example of trying to fit a sinusoidal function from Sec. 1. Recall from Fig. 1 that using the Gaussian NLL as the objective resulted in a sub-optimal fit. In contrast, using MSE as the objective, the model converged without problems in reasonable time. We now analyze the reasons behind this surprising result.

From Eq. 3, we see that the true mean $\mu(X)$ is a minimizer of the NLL loss. It thus becomes clear that a) the solution found in Fig. 1 is not the optimal one, and b) the NLL objective should, in principle, drive $\hat{\mu}(X)$ to the optimal solution $\mu(X)$. So why is the model not converging? We identify two main culprits for the non-convergence of the Gaussian NLL objective:

1. Initial flatness of the feature space can create an undercomplex but locally stable mean fit. This fit results from local symmetries and requires a form of symmetry breaking to escape.
2. The NLL loss scales the gradient of badly-predicted points down relative to well-predicted points, effectively undersampling those points. This effect worsens as training progresses.

These culprits and their effect on training are illustrated in Fig. 2 (left). If the network cannot fit a certain region yet, because its feature space (spanned by the last hidden layer) is too coarse, it results in a high effective data variance. This leads to down-weighting the data from these regions, fueling a vicious circle of self-amplifying the increasingly imbalanced weighting. In the following, we analyse these effects and their reasons in more detail.

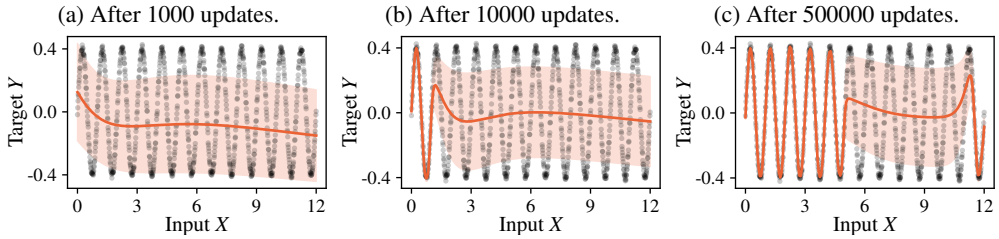

Figure 3: Model fit at different stages of training using the NLL loss function shown in red with $\pm\sigma$ uncertainty band. Black dots mark training data. Fitting the function begins left and is visibly slow.

3.1 Symmetry and Feature Non-Linearity

It is instructive to see how the model evolves over training. Figure 3 shows snapshots of the model at different stages of training. It can be seen that the network first learns essentially the best linear fit while adapting the variance to match the spread of points around the predicted mean. After the linear fit has been established, the situation is locally stable. That is, due to the symmetries of errors below and above the mean fit, there is no incentive to change the situation. Escaping this situation thus requires some form of symmetry breaking. One form of symmetry breaking comes from the stochasticity of mini-batch sampling in SGD, or natural asymmetries contained in the dataset, e.g. outlier points. In addition, we hypothesize that an important form of symmetry breaking is how non-linear the feature space of the network is locally, in order to create the necessary non-linear fit.

Let us consider the non-linearity of the feature space. This quantity is not easy to capture. To approximate it, we compute how much the Jacobian J_f of the features $f(x)$ with respect to the input varies in an L2-ball with radius r around a point x, denoted as the Jacobian vriance:[2]

$$V(x) = \frac{1}{|\mathcal{B}|} \sum_{x' \in \mathcal{S}} \left(J_f(x') - \frac{1}{|\mathcal{S}|} \sum_{x'' \in \mathcal{B}} J_f(x'') \right)^2, \quad \mathcal{B} = \{x' \colon \|x - x'\|_2 \leq r\}. \tag{5}$$

Figure 4 visualizes the Jacobian variance over the input space as a function of the training progress. Although initially relatively flat, it gets more fine-granular in parts of the input space – the parts which are later well fit. The region that has a low Jacobian variance gets stuck in this configuration, see Fig. 1. This provides evidence that the non-linearity of the feature space is important for the success or failure of learning. However, it does not answer the question of why gradient descent on the NLL loss would not break out of this situation.

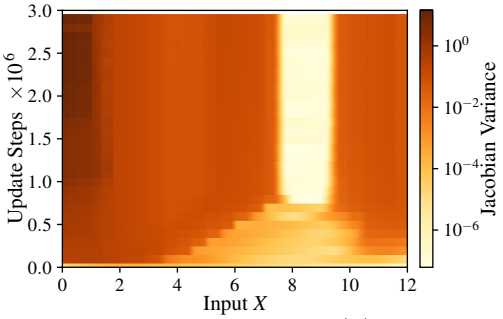

Figure 4: Jacobian variance $V(x)$ (see Eq. 5) over training time.

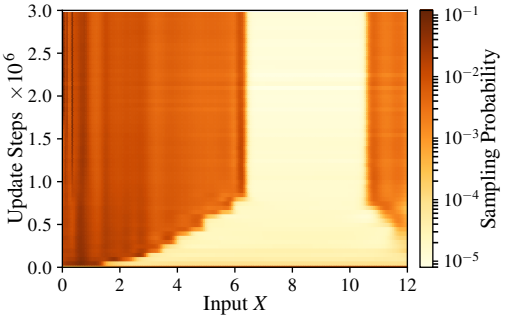

Figure 5: Probability of sampling a data point at x over training time.

3.2 Weighting by Variance Effectively Undersamples

The answer lies in an imbalanced weighting of data points across the input space. Recall that the gradient $\nabla_\mu \mathcal{L}_{\text{NLL}}$ of the NLL with respect to the mean scales the error $\hat{\mu}(X) - Y$ by $\frac{1}{\hat{\sigma}^2(X)}$

[2]This is a form of second-order derivative computed numerically, which also gives non-zero results for networks with ReLU activation (in contrast to, for example, the Hessian).

(Eq. 3). As symmetry is broken and the true function starts to be fit locally, the variance quickly shrinks in these areas to match the reduced MSE. If the variance is well-calibrated, the gradient becomes $\frac{\hat{\mu}(X)-Y}{\hat{\sigma}^2(X)} \approx \frac{\hat{\mu}(X)-Y}{(\hat{\mu}(X)-Y)^2} = \frac{1}{\hat{\mu}(X)-Y}$. That is, data points with already low error will get their contribution in the batch gradient scaled up relatively to high error data points – a "rich get richer" self-amplifying scenario. Thus, the NLL exhibits the opposite of the MSE loss's well-known behavior to focus on high-error samples. If the true variance σ^2 on the well-fit regions is small compared to the error on the badly-fit regions, or if the ratio of well-fit to badly-fit points is high, this can *completely prevent learning progress* on the badly-fit regions.

Another way to view this is to interpret the different weighting of points as *changing the training distribution $P(X, Y)$ to a modified distribution $\tilde{P}(X, Y)$* in which points with high error have a lower probability of getting sampled. This can be shown by defining $\tilde{P}(X, Y) = Z^{-1}\frac{P(X,Y)}{\sigma^2(X)}$, where $Z = \int \frac{P(x,Y)}{\sigma^2(x)}\mathrm{d}x$ is a normalizing constant, and recognizing that the gradient of the NLL equals the gradient of the MSE loss in Eq. 2 under the modified data distribution $\tilde{P}(X, Y)$:

$$\nabla_\mu \mathcal{L}_{\mathrm{NLL}}(\theta) = \mathbb{E}_{(X,Y)\sim\tilde{P}(X,Y)}[\mu(X) - Y] = \nabla_\mu \mathbb{E}_{(X,Y)\sim\tilde{P}(X,Y)}\left[\frac{(Y-\mu(X))^2}{2}\right]. \tag{6}$$

In Fig. 5, we plot $\tilde{P}(X, Y)$ over training time for our sinusoidal example. It can be seen that the virtual probability of sampling a point from the high-error region drops over time, until it is highly unlikely to sample points from this region (10^{-5} as opposed to 10^{-3} for uniform sampling) . As a side note, optimization with SGD typically assumes a fixed training distribution; training under shifting distributions may thus make optimization with SGD more difficult. It also was recently shown that such a shifting training setup can hurt generalization (Igl et al., 2021). These two points may further contribute to the sub-optimal performance of probabilistic neural networks experienced in practice.

Sometimes "weighting-by-variance" is seen as a feature of the Gaussian NLL (Kendall & Gal, 2017), namely that it introduces a self-regularizing property by allowing the network to "ignore" outlier points with high error. This can be desirable if the predicted variance corresponds to the inherent unpredictability (noise), but it is undesirable if it causes premature convergence and ignorance of hard-to-fit regions, as shown above. In our solution, explained below, we allow to control the amount of this self-regularization.

4 Method

To mitigate these problems with NLL training, we develop a solution in the following. We start with a method that uses standard squared error losses to estimate the moments of the distribution to be predicted, which we term "Moment Matching". It fixes the problem with premature convergence when using $\mathcal{L}_{\mathrm{NLL}}$and yields consistent training results, but leads to underestimated variances and will act as one of the baselines in the empirical evaluation.

We then consider the distribution of residual errors, which are drastically different between NLL training and MSE training. This suggests a problem-specific design choice that leads us to our proposed $\beta-$NLL introduced in Sec. 4.3. It allows choosing a meaningful loss-interpolation between NLL and MSE while keeping calibrated uncertainty estimates.

4.1 Moment Matching

A natural solution to the problem of estimating distributions would be to estimate the sufficient statistics of the distribution. In the case of the Gaussian distribution, these would be the first two moments μ, σ^2 fully describing the distribution. So why not defining losses based on the moment estimators? Starting from the conditional mean of the predictions y: $\mathbb{E}[Y \mid X]$, we can define the squared deviation as a loss and see that the MSE is an upper bound:

$$\left(\mathbb{E}[Y \mid X] - \hat{\mu}(X)\right)^2 = \mathbb{E}[(Y - \hat{\mu}(X)) \mid X]^2 \leq \mathbb{E}\left[(Y - \hat{\mu}(X))^2 \mid X\right] := \mathcal{L}^{\hat{\mu}}_{MM}\,. \tag{7}$$

Notice that $\mathcal{L}^{\mu}_{MM}$ is the standard MSE loss. Thus, fitting the mean $\hat{\mu}(x)$ can follow standard (non-distributional) regression procedures. Interestingly, due to the moment-matching viewpoint, we

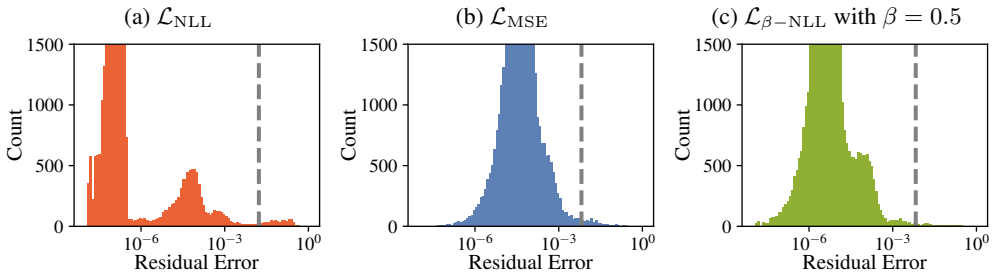

Figure 6: Distribution over residual prediction errors depending on the choice of loss function for the ObjectSlide dataset (see Sec. 5.3). Dashed line shows the RMSE. (a) NLL loss: residuals are multimodal. There is a long tail of difficult data points that are ignored, whereas easy regions are fit to high accuracy. (b) Using MSE loss results in a log-normal residual distribution. (c) Our loss $\mathcal{L}_{\beta-\text{NLL}}$, introduced below, yields highly accurate fits in easy regions without ignoring hard examples.

can analogously define a loss for the variance (second central moment). The variance is defined as $\mathbb{E}\big[(Y-\mu(X))^2 \mid X\big]$; analogously, we get the following loss: $\mathbb{E}\Big[\big((Y-\hat{\mu}(X))^2-\hat{\sigma}^2(X)\big)^2 \mid X\Big]$. To use the same physical unit as $\mathcal{L}^{\mu}_{MM}$, we reformulate it in terms of the standard deviation as:

$$\mathcal{L}^{\sigma}_{MM} := \mathbb{E}\left[\left(\sqrt{(Y-\hat{\mu}(X))^2}-\hat{\sigma}(X)\right)^2 \mid X\right]. \tag{8}$$

Thus, the moment matching loss is simply given by the sum of both losses: $\mathcal{L}_{MM} = \mathcal{L}^{\mu}_{MM} + \mathcal{L}^{\sigma}_{MM}$.

4.2 Residual Distribution and Allocation of Function Approximator Capacity

The moment matching loss gives the same weighting to all data points. Since we are dealing with function approximators, the question is whether their capacity should be used equally. As discussed in Sec. 3.2, the NLL loss gives high weight to data points with low (predicted) variance and low weight to those with high variance. This is appropriate in case these variances are caused by true aleatoric uncertainty in the data. However, due to the use of function approximators, there is also the case where data points cannot be well predicted, resulting in high predicted variance, although the ground truth is corrupted by little noise. The handling of these *difficult points* is different for the different loss functions. Figure 6 shows the distribution of the residuals for a dynamics prediction dataset containing *easy* and *hard* to model areas. We observe a drastic difference between the NLL loss and the MSE/moment matching loss.

How important are the data points with high uncertainty? Are these outliers or stem from truly noisy regions? In this case, we would be willing to allocate less of the function approximator's capacity to them. Or are these simply difficult samples? In our dynamics prediction example, we analysed the data points with high uncertainty and found that they were actually *the most important ones* to get correct (because they captured non-trivial interactions in the physical world). Following this insight, this suggests that there is no one-loss-fits-all, but rather that the modeler should have the choice to select which behavior is desired for the application at hand.

4.3 Variance-weighting the gradients of the NLL

In this section, we introduce our proposed solution to the problems occurring when training with $\mathcal{L}_{\text{NLL}}$. The most important problem to solve is the premature convergence of NLL training to highly suboptimal mean fits. We identified in Sec. 3 the main reason being the relative down-weighting of badly fit data points with its self-amplifying character. Effectively, NLL weights the mean squared error per data point with $\frac{1}{\sigma^2}$, which can be interpreted as sampling data points with $P(x) \propto \frac{1}{\sigma^2}$ (Sec. 3.2). Consequently, we propose modifying this distribution by introducing a parameter β allowing to *interpolate* between NLL's and a completely uniform data point importance. The resulting sampling distribution is given by $P(x) \propto \frac{\sigma^{2\beta}}{\sigma^2}$ and illustrated in Fig. 2 (right).

How could this weighting be achieved? We simply introduce the variance-weighting term $\sigma^{2\beta}$ to the $\mathcal{L}_{\text{NLL}}$ loss such that it acts as a factor on the gradient. We denote the resulting loss as $\beta-$NLL given

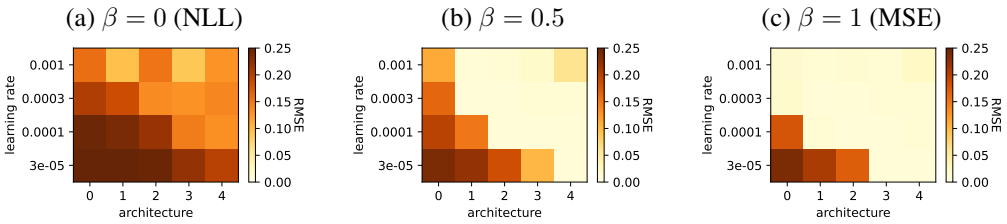

Figure 7: Convergence properties analyzed on the sinusoidal toy regression problem. The root mean squared error (RMSE) is displayed as a color code, depending on model-architecture (see Sec. C.1) and learning rate after 200 000 epochs (mean over 3 independent trials). The original NLL ($\beta = 0$) does not obtain good RMSE fits for most hyperparameter settings. Figure S1 shows the NLL scores.

by:

$$\mathcal{L}_{\beta-\text{NLL}} := \lfloor \hat{\sigma}^{2\beta}(X) \rfloor \mathcal{L}_{\text{NLL}} = \mathop{\mathbb{E}}_{X,Y} \lfloor \hat{\sigma}^{2\beta}(X) \rfloor \left[\frac{1}{2} \log \hat{\sigma}^2(X) + \frac{(Y - \hat{\mu}(X))^2}{2\hat{\sigma}^2(X)} + \text{const} \right]. \tag{9}$$

where $\lfloor \cdot \rfloor$ denotes the *stop gradient* operation. The stop gradient makes the variance weighting term to be treated as a prefactor or learning rate. In this way, the gradients of $\mathcal{L}_{\beta-\text{NLL}}$ are:

$$\nabla_{\hat{\mu}} \mathcal{L}_{\beta-\text{NLL}}(\theta) = \mathop{\mathbb{E}}_{X,Y} \left[\frac{\hat{\mu}(X) - Y}{\hat{\sigma}^{2-2\beta}(X)} \right], \quad \nabla_{\hat{\sigma}^2} \mathcal{L}_{\beta-\text{NLL}}(\theta) = \mathop{\mathbb{E}}_{X,Y} \left[\frac{\hat{\sigma}^2(X) - (Y - \hat{\mu}(X))^2}{2\hat{\sigma}^{4-2\beta}(X)} \right]. \tag{10, 11}$$

Naturally, for $\beta = 0$, we recover the original NLL loss. For $\beta = 1$ the gradient w.r.t. μ in Eq. 10 is equivalent to the one of MSE. However, for the variance, the gradient in Eq. 11 is a new quantity with $2\sigma^2$ in the denominator. This is in contrast to the moment matching loss (Eq. 8). For values $0 < \beta < 1$, we get different loss interpolations. Particularly interesting is the case of $\beta = 0.5$, where the data points are weighted with $\frac{1}{\sigma}$ (inverse standard deviation instead of inverse variance). In Sec. 5, we show that $\beta = 0.5$ generally achieves the best trade-off between accuracy and log-likelihood.

As a remark, the new loss $\mathcal{L}_{\beta-\text{NLL}}$ is not meant for performance evaluation – it is designed to result in meaningful gradients. We found that its absolute value can be counter-intuitive due to the prefactor. The model performance during training should be monitored with the original (negative) log-likelihood and optionally the RMSE for the mean fit.

5 Experiments

We investigate the proposed changes in the loss function for training neural networks to predict a Gaussian distribution on toy examples and real-world datasets, the UCI dataset, and challenging dynamics model datasets.

As motivated above, the standard NLL loss often leads to suboptimal fits, for instance, where the mean fit is far from its desired quality. To analyze this quantitatively, we report the performance for many hyperparameter settings, such as learning rates and network architectures and consider both the quality in predictions with respect to the root mean squared error (RMSE) and the negative log likelihood (NLL).

5.1 Synthetic Datasets

Sinusoidal without heteroscedastic noise We first consider the dataset already used as an illustrative example in Fig. 1 – a simple sine curve with small additive noise: $y = \sin(x) + \xi$, with ξ being Gaussian noise with standard deviation $\sigma = 0.01$. One would expect that a network with sufficient capacity can easily learn to fit this function.

We find that for the standard NLL ($\beta = 0$) the network does not converge to a reasonable mean fit. There is a trend that larger networks and learning rates show better results, but when comparing this to β−NLL with $\beta = 0.5$ we see that the networks are indeed able to fit the function without problems. As expected, the same holds for the mean squared error loss (MSE) which is recovered for $\beta = 1$. The quality of the fit w.r.t. NLL is shown in Fig. S1 for completeness.

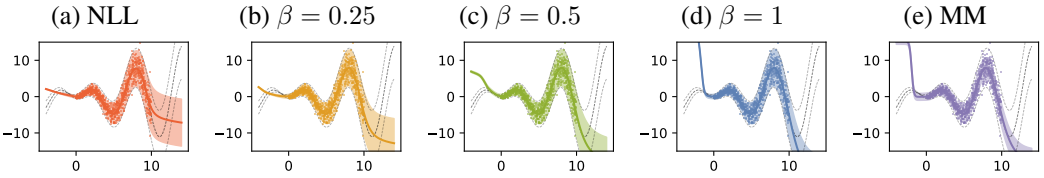

Figure 8: Fits for the heteroscedastic sine example from Detlefsen et al. (2019). Dotted lines show the ground truth mean and $\pm 2\sigma$, respectively. Inside of the training regime, all $\beta-$NLL variants (a-d) yield well-calibrated uncertainty estimates. Outside the training regime, the variance estimates are smaller for larger β. Moment matching (e) is significantly underestimating the variance everywhere.

Table 1: Results for UCI Regression Datasets. We report predictive log-likelihood (higher is better) and standard deviation, together with dataset size, input and output dimensions.

Loss	β	boston (506, 13, 1)	concrete (1030, 8, 1)	energy (768, 8, 2)	kin8m (8192, 8, 1)	naval (11934, 16, 2)	power plant (9568, 4, 1)	protein (45730, 9, 1)	wine-red (1599, 11, 1)	yacht (308, 6, 1)
$\mathcal{L}_{\beta-\text{NLL}}$	0	-2.86 $\pm$ 0.48	**-3.25 $\pm$ 0.31**	**-1.71 $\pm$ 0.95**	**1.14 $\pm$ 0.04**	6.72 $\pm$ 0.25	**-2.81 $\pm$ 0.06**	**-2.80 $\pm$ 0.05**	**-1.03 $\pm$ 0.24**	-2.86 $\pm$ 5.05
$\mathcal{L}_{\beta-\text{NLL}}$	0.25	**-2.75 $\pm$ 0.41**	**-3.31 $\pm$ 0.49**	**-1.75 $\pm$ 0.96**	**1.14 $\pm$ 0.03**	**6.90 $\pm$ 0.19**	**-2.80 $\pm$ 0.05**	**-2.79 $\pm$ 0.04**	**-0.98 $\pm$ 0.12**	-1.97 $\pm$ 1.11
$\mathcal{L}_{\beta-\text{NLL}}$	0.5	**-2.64 $\pm$ 0.36**	**-3.29 $\pm$ 0.35**	**-1.77 $\pm$ 1.26**	**1.14 $\pm$ 0.04**	**6.96 $\pm$ 0.24**	**-2.81 $\pm$ 0.05**	**-2.78 $\pm$ 0.02**	**-0.99 $\pm$ 0.15**	-2.47 $\pm$ 1.64
$\mathcal{L}_{\beta-\text{NLL}}$	0.75	**-2.72 $\pm$ 0.41**	**-3.27 $\pm$ 0.33**	**-1.82 $\pm$ 0.96**	**1.14 $\pm$ 0.04**	**7.06 $\pm$ 0.22**	**-2.81 $\pm$ 0.05**	**-2.79 $\pm$ 0.02**	**-1.02 $\pm$ 0.22**	**-1.87 $\pm$ 0.54**
$\mathcal{L}_{\beta-\text{NLL}}$	1.0	-2.85 $\pm$ 0.85	**-3.23 $\pm$ 0.32**	-1.91 $\pm$ 0.51	**1.13 $\pm$ 0.04**	6.95 $\pm$ 0.27	**-2.81 $\pm$ 0.05**	**-2.80 $\pm$ 0.03**	**-0.97 $\pm$ 0.09**	-2.27 $\pm$ 1.04
$\mathcal{L}_{\text{MM}}$	—	-3.42 $\pm$ 0.98	-3.49 $\pm$ 0.37	-2.74 $\pm$ 1.74	1.00 $\pm$ 0.06	6.76 $\pm$ 0.28	-2.92 $\pm$ 0.09	-2.98 $\pm$ 0.06	-1.22 $\pm$ 0.42	-11.2 $\pm$ 30.2

Sinusoidal with heteroscedastic noise A commonly used illustrative example introduced in Detlefsen et al. (2019) is a sine curve with increasing amplitude and noise: $y = x\sin(x) + x\xi_1 + \xi_2$, with ξ being Gaussian noise with standard deviation $\sigma = 0.3$. The functional form is much easier, so fitting the mean is achieved with all losses. Here we sanity-check that the newly introduced loss is still delivering good uncertainty estimates. Figure 8 displays the predictions of the best models (w.r.t. NLL validation loss).

5.2 UCI Regression Datasets

As a standard real-world benchmark in predictive uncertainty estimation, we consider the UCI dataset (Hernández-Lobato & Adams, 2015). Table 1 reports the log-likelihoods of the test data comparing the different loss variants. We report RMSE in Table S1. The results are encouraging: $\beta-$NLL achieves the same predictive log-likelihood as the NLL loss or better, while clearly improving the predictive accuracy on most datasets.

5.3 Dynamics models

As a major application of uncertainty estimation lies in model-based reinforcement learning (RL), we test the different loss functions on two dynamics predictions tasks of varying difficulty, ObjectSlide and Fetch-PickAndPlace. In both tasks, the goal is to predict how an object will move from the current state and the agent's action. Whereas ObjectSlide (Seitzer et al., 2021) is a simple 1D-environment, Fetch-PickAndPlace (Plappert et al., 2018) is a complex 3D robotic-manipulation environment. The models are trained on state-action trajectories collected by running RL agents. See Sec. B for more details.

For both datasets and all loss functions, we first perform a grid search over different hyperparameter configurations (see C.2). The results also serve for a sensitivity analysis presented in Fig. 9: how much does one have to tune hyperparameters for each loss function to work well in practice? We then take the best performing configurations and evaluate them on a hold-out test set (Table 2). The sensitivity analysis reveals that both the NLL- and the MM-loss are vulnerable to hyperparameter settings, whereas the $\beta-$NLL loss achieves good results over a wide range of settings. When focusing on the best found model configurations for each loss in Table 2, one can see that the NLL loss results in poor predictive performance and also exhibits quite a high variance across random seeds; the MM-loss results in poor log-likelihood fits. Our method performs well on both accuracy and log-likelihood fit, for a range of β-values, where $\beta = 0.5$ generally achieves the best trade-off between both.

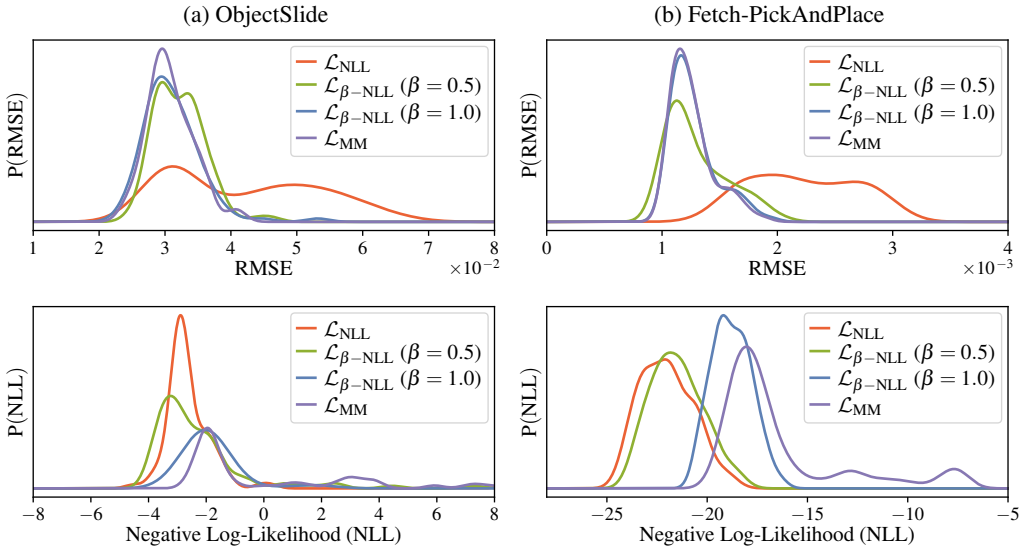

Figure 9: Sensitivity analysis of loss functions to hyperparameter settings. We plot the distribution over RMSE (top row) and NLL (bottom row) as a function of hyperparameters, on validation sets of the ObjectSlide (a) and Fetch-PickAndPlace (b) datasets. The values stem from a grid search over different model configurations (see C.2 for a description of the settings). It can be seen that the NLL loss is sensitive when evaluating RMSE, whereas the MM loss is sensitive when evaluating NLL (in fact, many runs for MM ended with exploding NLLs). In contrast, the β-NLL loss does not exhibit the same sensitivity and achieves relatively stable results regardless of the concrete settings.

Table 2: Test results for dynamics models, using best configurations found in grid search. Reported standard deviation is over 5 random seeds. Although the NLL loss with $\beta = 0$ achieves the top-performing log-likelihood on some runs, its average over random seeds suffers due to high variance.

		1D-Slide		**Fetch-PickAndPlace**	
Loss	β	RMSE ↓	LL ↑	RMSE ↓	LL ↑
$\mathcal{L}_{\beta-\text{NLL}}$	0	0.0192 ± 0.006	7.97 ± 3.62	0.00163 ± 0.00008	18.72 ± 7.32
$\mathcal{L}_{\beta-\text{NLL}}$	0.25	0.0107 ± 0.004	9.03 ± 0.47	0.00102 ± 0.00004	24.43 ± 1.64
$\mathcal{L}_{\beta-\text{NLL}}$	0.5	$\mathbf{0.0064 \pm 0.002}$	$\mathbf{9.28 \pm 0.75}$	$\mathbf{0.00096 \pm 0.00002}$	$\mathbf{24.68 \pm 0.08}$
$\mathcal{L}_{\beta-\text{NLL}}$	0.75	0.0087 ± 0.003	6.61 ± 1.83	0.00098 ± 0.00001	22.77 ± 0.17
$\mathcal{L}_{\beta-\text{NLL}}$	1.0	0.0074 ± 0.001	6.58 ± 0.29	0.00102 ± 0.00001	21.32 ± 0.07
$\mathcal{L}_{\text{MM}}$		0.0078 ± 0.001	diverges	0.00104 ± 0.00003	19.33 ± 1.31
$\mathcal{L}_{\text{MSE}}$		0.0068 ± 0.001	—	0.00103 ± 0.00000	—

6 Discussion

We highlight a deep problem frequently occurring when optimizing probabilistic neural networks using the common NLL loss: training gets stuck in suboptimal function fits. Due to the predictive variance-based weighting of mean prediction errors in the NLL loss, these bad fits often stay unnoticed, unless the MSE is evaluated. With our analysis on feature space non-linearity and induced data-set weighting (interpreted as a biased dataset sampling). we reveal the underlying reason: initially badly fit regions (due to low feature non-linearity) get increasingly less weight and result in premature convergence. We propose a simple solution by introducing a family of loss functions called $\beta-$NLL. Effectively, the gradient of the original NLL loss is scaled by the β-exponentiated per-sample variance. This allows for a meaningful interpolation between NLL loss and MSE loss while providing well-behaved uncertainty estimates (even for the MSE case). The hyperparameter β gives practitioners the choice to control the self-regularization strength of NLL: how unimportant should high-noise regions or difficult-to-predict data points be in the fitting process. In most cases $\beta = 0.5$ will be a good starting point. We hope that our simple solution will improve the usability and performance of predictive uncertainty deep networks and increase their applicability.

REPRODUCIBILITY STATEMENT

All settings are described in detail in Sec. B and Sec. C. We make full code and data available under https://sites.google.com/view/pitfalls-uncertainty. Both will be made publicly available after the review phase.

REFERENCES

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *ArXiv*, abs/1606.06565, 2016.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In Francis Bach and David Blei (eds.), *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1613–1622, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/blundell15.html.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, J. Schneider, John Schulman, Jie Tang, and W. Zaremba. OpenAI Gym. *ArXiv*, abs/1606.01540, 2016.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/3de568f8597b94bda53149c7d7f5958c-Paper.pdf.

Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1096–1105. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/dabney18a.html.

Nicki S. Detlefsen, Martin Jørgensen, and Søren Hauberg. Reliable training and estimation of variance networks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/07211688a0869d995947a8fb11b215d6-Paper.pdf.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/gal16.html.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian active learning with image data. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1183–1192. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/gal17a.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Ali Harakeh and Steven L. Waslander. Estimating and evaluating regression predictive uncertainty in deep object detectors. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YLewtnvKgR7.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Hkg4TI9xl.

José Miguel Hernández-Lobato and Ryan P. Adams. Probabilistic backpropagation for scalable learning of Bayesian neural networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pp. 1861–1869. JMLR.org, 2015.

Shi Hu, Nicola Pezzotti, and Max Welling. A new perspective on uncertainty quantification of deep ensembles. *ArXiv*, 2020.

Maximilian Igl, Gregory Farquhar, Jelena Luketina, Wendelin Boehmer, and Shimon Whiteson. Transient non-stationarity and generalisation in deep reinforcement learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=Qun8fv4qSby.

Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/2650d6089a6d640c5e85b2b88265dc2b-Paper.pdf.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural Safety*, 31(2):105–112, 2009. ISSN 0167-4730. doi: https://doi.org/10.1016/j.strusafe.2008.06.020. URL https://www.sciencedirect.com/science/article/pii/S0167473008000556. Risk Acceptance and Risk Communication.

Alina Kloss, Georg Martius, and Jeannette Bohg. How to train your differentiable filter. *Autonomous Robots*, 45:562–578, June 2021. doi: 10.1007/s10514-021-09990-9. URL https://link.springer.com/article/10.10072Fs10514-021-09990-9.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf.

Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip H. S. Torr, and Yarin Gal. Deterministic neural networks with appropriate inductive biases capture epistemic and aleatoric uncertainty. *CoRR*, abs/2102.11582, 2021. URL https://arxiv.org/abs/2102.11582.

David A. Nix and Andreas S. Weigend. 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 vol.1, 1994. doi: 10.1109/ICNN.1994.374138.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped DQN. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/8d8818c8e140c64c743113f563cf750f-Paper.pdf.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/8558cb408c1d76621371888657d2eb1d-Paper.pdf.

Cristina Pinneri, Shambhuraj Sawant, Sebastian Blaes, and Georg Martius. Extracting strong policies for robotics tasks from zero-order trajectory optimizers. In *9th International Conference on Learning Representations (ICLR 2021)*, May 2021.

Matthias Plappert, Marcin Andrychowicz, Alex Ray, Bob McGrew, Bowen Baker, Glenn Powell, J. Schneider, Joshua Tobin, Maciek Chociej, P. Welinder, V. Kumar, and W. Zaremba. Multi-goal reinforcement learning: Challenging robotics environments and request for research. *ArXiv*, abs/1802.09464, 2018.

Maximilian Seitzer, Bernhard Schölkopf, and Georg Martius. Causal influence detection for improving efficiency in reinforcement learning. In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*. Curran Associates, Inc., December 2021. URL https://arxiv.org/abs/2106.03443.

Andrew Stirn and David A. Knowles. Variational variance: Simple and reliable predictive variance parameterization. *ArXiv*, abs/2006.04910, 2020.

Marin Vlastelica, Sebastian Blaes, Cristina Pinneri, and Georg Martius. Risk-averse zero-order trajectory optimization. In *Conference on Robot Learning*, 2021.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. MOPO: Model-based offline policy optimization. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 14129–14142. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/a322852ce0df73e204b7e67cbbef0d0a-Paper.pdf.

APPENDIX

A ADDITIONAL RESULTS

Synthetic Dataset – NLL fit In Fig. S1, we provide the results for the NLL metric on the sinusoidal dataset.

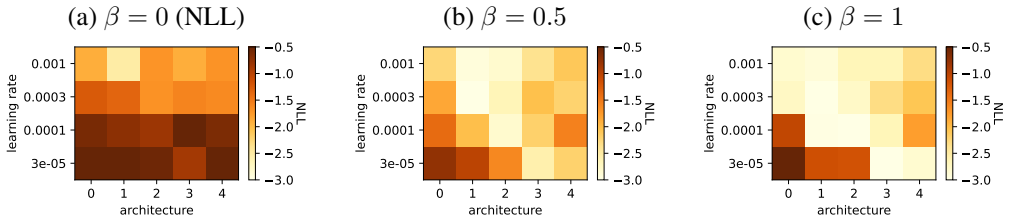

Figure S1: Convergence properties analyzed on the sinusoidal toy regression problem. Same as Fig. 7, but for the negative log-likelihood (NLL). Due to the bad mean fit, the original NLL loss ($\beta = 0$) is also bad for most hyperparameter settings. With $\beta > 0.5$ good fits are obtained for many settings. Also for $\beta = 1$, which corresponds to MSE for fitting the mean, good uncertainty predictions are obtained with our $\beta-$NLL as testified by the low NLL scores.

UCI Dataset In addition to the results presented in the main text in Table 1 (where we report log-likelihoods of the predictions), we include the achieved RMSE values in Table S1.

Table S1: Results for UCI Regression Datasets. RMSE (lower is better) and standard deviation, together with dataset size, input and output dimensions.

Loss	β	boston (506, 13, 1)	concrete (1030, 8, 1)	energy (768, 8, 2)	kin8m (8192, 8, 1)	naval (11934, 16, 2)
$\mathcal{L}_{\beta-\text{NLL}}$	0	3.56 ± 1.04	6.08 ± 0.63	1.93 ± 0.37	0.09 ± 0.00	0.00 ± 0.00
$\mathcal{L}_{\beta-\text{NLL}}$	0.25	3.48 ± 1.13	5.79 ± 0.72	1.71 ± 0.41	0.08 ± 0.00	0.00 ± 0.00
$\mathcal{L}_{\beta-\text{NLL}}$	0.5	$\mathbf{3.42 \pm 1.02}$	5.61 ± 0.63	$\mathbf{0.98 \pm 0.52}$	0.08 ± 0.00	0.00 ± 0.00
$\mathcal{L}_{\beta-\text{NLL}}$	0.75	$\mathbf{3.43 \pm 1.04}$	5.67 ± 0.71	$\mathbf{1.01 \pm 0.57}$	0.08 ± 0.00	0.00 ± 0.00
$\mathcal{L}_{\beta-\text{NLL}}$	1.0	3.50 ± 0.93	$\mathbf{5.55 \pm 0.75}$	1.32 ± 0.64	0.08 ± 0.00	0.00 ± 0.00
$\mathcal{L}_{\text{MM}}$	—	4.02 ± 1.15	6.28 ± 0.79	1.91 ± 0.47	0.08 ± 0.00	0.00 ± 0.00

Loss	β	power plant (9568, 4, 1)	protein (45730, 9, 1)	wine-red (1599, 11, 1)	yacht (308, 6, 1)
$\mathcal{L}_{\beta-\text{NLL}}$	0	4.06 ± 0.18	4.49 ± 0.09	0.64 ± 0.04	$\mathbf{1.22 \pm 0.46}$
$\mathcal{L}_{\beta-\text{NLL}}$	0.25	4.04 ± 0.17	4.35 ± 0.05	0.64 ± 0.04	1.73 ± 0.97
$\mathcal{L}_{\beta-\text{NLL}}$	0.5	4.04 ± 0.17	4.31 ± 0.02	0.64 ± 0.04	2.35 ± 1.40
$\mathcal{L}_{\beta-\text{NLL}}$	0.75	4.04 ± 0.15	$\mathbf{4.28 \pm 0.02}$	0.64 ± 0.03	1.97 ± 1.01
$\mathcal{L}_{\beta-\text{NLL}}$	1.0	4.06 ± 0.17	4.31 ± 0.04	0.64 ± 0.03	2.08 ± 1.11
$\mathcal{L}_{\text{MM}}$	—	4.07 ± 0.15	4.32 ± 0.06	0.65 ± 0.04	3.02 ± 1.34

B DATASETS AND TRAINING SETTINGS

Sinusoidal without heteroscedastic noise This dataset is created by taking 1 000 uniformly spaced points on the interval $[0, 12]$ as inputs x and applying the function $y(x) = 0.4 \sin(2\pi x) + \xi$ to them to create the targets y, where ξ is Gaussian noise with a standard deviation of 0.01.

Sinusoidal heteroscedastic noise We use the synthetic data as introduced in Detlefsen et al. (2019). From the functional form $y = x \sin(x) + x\xi_1 + \xi_2$, with Gaussian noise with standard deviation $\sigma = 0.3$ for ξ_1 and ξ_2, we sample 500 points uniformly spaced in the interval $[0, 10]$. The model is a MLP with one hidden layer with 50 units and $\tanh$ activation function (as used in Detlefsen et al. (2019) and Stirn & Knowles (2020)).

UCI Datasets We use the UCI datasets suite commonly used to benchmark uncertainty estimation, stemming from the UCI Machine Learning Repository[3] In particular, we use the training-test protocol from (Hernández-Lobato & Adams, 2015; Gal & Ghahramani, 2016), and their data splits [4].

Inputs and targets are whitened on the training set. Metrics are reported in the original scale of the data. Each dataset is divided into 20 randomly sampled train-test splits (80%-20%). For each split, we further divide the training set into 80% training data and 20% validation data and search for an optimal learning rate from the set $\{10^{-4}, 3 \cdot 10^{-4}, 7 \cdot 10^{-4}, 10^{-3}, 3 \cdot 10^{-3}, 7 \cdot 10^{-3}\}$ by monitoring log-likelihood on the validation set. We train for a maximum of 20 000 updates, besides on the larger "kin8m", "power plant", "protein" and "naval" datasets, where we train for a maximum of 100 000 updates We perform early stopping with a patience of 50 epochs, retrain the model with the best found learning rate on the full training set, and then evaluate on the test split. The reported performance and standard deviations are taken as averages over all test splits. Note that the performance we report is not comparable with other publications, as performance is known to differ strongly over different data splits. Some other works also perform early-stopping on the test set, which distorts the results.

As model, we use a single-layer relu hidden network with 50 neurons, aside from "protein", where we use 100 neurons. The batch size is 256.

ObjectSlide This environment consists of an agent whose task it is to slide an object to a target location (Seitzer et al., 2021). The continuous state space consists of 4 dimensions: agent and object position and velocity, and the continuous action space is a one-dimensional movement command. The forward prediction task consists of predicting the change in object position in the next state from current state and action. The dataset we use consists of 180 000 transitions collected using a random policy, which we split into training, validation and testing sets with 60 000 transitions each. Inputs and targets are whitened on the training set. Metrics are reported in the original scale of the data. We train for a maximum of 5 000 epochs with a batch size of 256, and evaluate the model with the best validation log-likelihood on the test set afterwards.

Fetch-PickAndPlace We use the Fetch-PickAndPlace environment (Plappert et al., 2018) from OpenAI Gym (Brockman et al., 2016) as a challenging real-world scenario. The task of the agent is to use position-controlled 7 DoF robotic arm to lift an object to a target location in space. The state space is 25-dimensional and the action space is 4-dimensional. As in ObjectSlide, the prediction task is to predict the 3D-dimensional change in object position from the current state and action. We use 840 000 transitions collected using the APEX method (Pinneri et al., 2021) as our dataset, which we split into 70% training, 15% validation and 15% testing data. Inputs and targets are whitened on the training set. Metrics are reported in the original scale of the data. We train for a maximum of 500 epochs with a batch size of 256, and evaluate the model with the best validation log-likelihood on the test set afterwards.

C Hyperparameter Settings and Implementation Details

For all experiments, we used the Adam optimizer (Kingma & Ba, 2015) with standard settings $\beta_1 = 0.9, \beta_2 = 0.999$. We parametrize the Gaussian distribution using two linear layers on top of shared features produced by a MLP. The variance $\hat{\sigma}^2(x)$ is constrained to the positive region using the $\mathrm{softplus}(x) = \log(1 + \exp(x))$ activation function We additionally add a small constant of 10^{-8} to prevent the variance from collapsing to zero, and clamp the maximum variance to 1 000.

C.1 Sinusoidal Regression Problem

The sinusoidal fit in Fig. 1 results from a network with two hidden layers and 128 neurons each, $\tanh$ activation, and optimized with learning rate $5 \cdot 10^{-4}$ and a batch size of 100. For the experiment in Sec. 5.1, we scan over learning rates and architectures with different hidden layers and units per layer, as detailed in Table S2.

[3] https://archive.ics.uci.edu

[4] available under https://github.com/yaringal/DropoutUncertaintyExps

Table S2: Architectures used for the sinusoidal regression task. Fully connected feed-forward neural networks with $\tanh$ activation function.

Architecture #	0	1	2	3	4
# Hidden Layers	2	2	2	3	3
# Units per Layer	32	64	128	128	256

C.2 OBJECTSLIDE AND FETCH-PICKANDPLACE

For each tested loss function, we performed a grid search for the ObjectSlide and Fetch-PickAndPlace datasets, We report the parameters in Table S3. For the results in Table 2, we retrained the model configurations with the best validation log-likelihood on the grid search, and evaluated them on the hold-out test set.

The best found configuration on ObjectSlide coincided to 3 hidden layers with 128 neurons, relu activation and learning rate 0.001. The best found configuration on Fetch-PickAndPlace coincided to 4 hidden layers with 128 neurons, relu activation and learning rates 0.0003 for $\mathcal{L}_{\mathrm{NLL}}$, $\mathcal{L}_{\beta-\mathrm{NLL}}$ with $\beta = 0.25$ and $\beta = 0.5$, and learning rate 0.001 for $\mathcal{L}_{\mathrm{MSE}}$, $\mathcal{L}_{\mathrm{MM}}$, $\mathcal{L}_{\beta-\mathrm{NLL}}$ with $\beta = 0.75$ and $\beta = 1.0$.

Table S3: Hyperparameter settings for grid search on ObjectSlide and Fetch-PickAndPlace datasets. We run 96 configurations per loss function.

Hyperparameter	Set of Values
Learning Rate	$\{3 \cdot 10^{-5}, 10^{-4}, 3 \cdot 10^{-4}, \cdot 10^{-3}\}$
# Hidden Layers	$\{2, 3, 4\}$
# Units per Layer	$\{128, 256, 386, 512\}$
Activation	$\{\tanh, \mathrm{relu}\}$