Reliable Uncertainty Estimates in Deep Neural Networks using Noise Contrastive Priors

Danijar Hafner, Dustin Tran, Alex Irpan, Timothy Lillicrap, James Davidson

Google AI DeepMind

1. Overview

Bayesian neural networks (BNNs) provide uncertainty estimates by modeling a belief distribution over weights.

Only trained on training data → uncertainties may be arbitrary for out-of-distribution (OOD) data as the weight belief can generalize in unforeseen ways.

Contribution. A simple, scalable strategy to train an uncertainty-aware model towards high uncertainty on OOD data.

2. Noise Contrastive Priors

Ideally, we can select the BNN prior to assign high uncertainty to OOD data.

Expressing priors in weight space is difficult.

Instead, define the prior in data space:

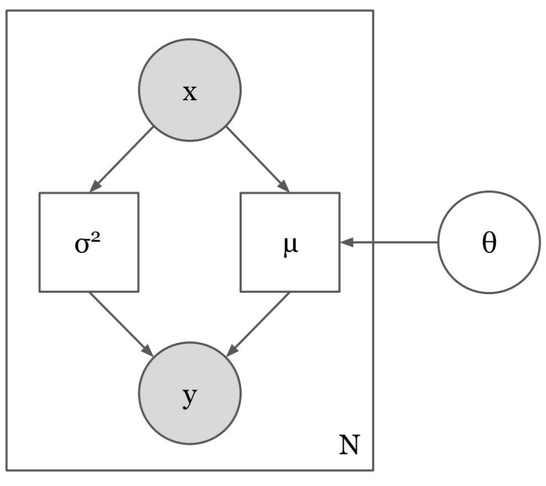
1. Inputs: add noise to mini-batch (will be OOD near data set boundary).

$$\tilde{x} = x + \epsilon \qquad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

2. Set prior on variables such that predictive variance is high on these OOD inputs.

$$\text{e.g.} \quad \mathrm{KL}\left(\mathcal{N}(0, \sigma_z^2) \,||\, p(z|\tilde{x})\right)$$

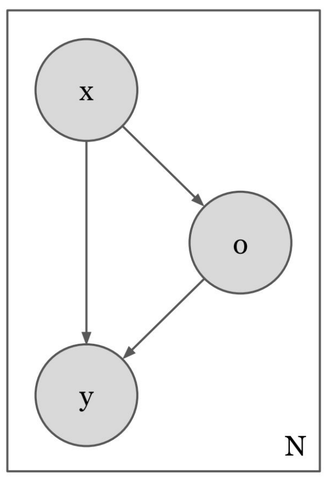
3. Uncertainty-Aware Models



Bayesian Neural Network with NCP on μ

$$\theta \sim p(\theta)$$
$$y \sim \mathcal{N}(\mu(x, \theta), \sigma^2(x))$$

$$\mathcal{L} = -\mathbb{E}_{q_\phi(\theta)}[\log p(y \mid x, \theta)] + \mathrm{KL}(q_\phi(\theta) \,\|\, p(\theta)) \boxed{+ \mathrm{KL}\left(\mathcal{N}(0, \sigma_\mu^2) \,\|\, p(\mu(\tilde{x}, \theta \sim q))\right)}$$

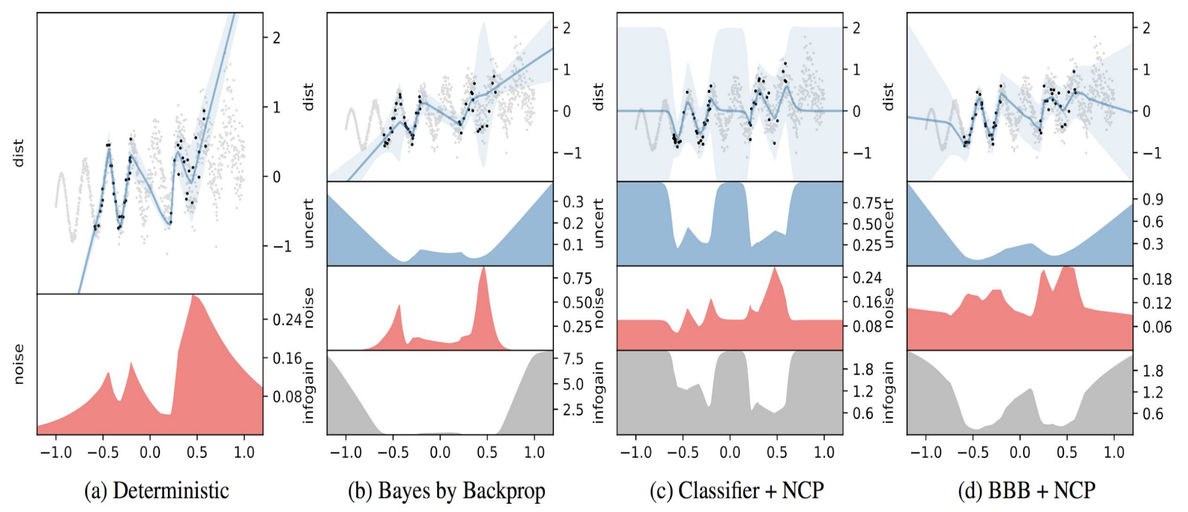


OOD Classifier Model with NCP on o

$$o \sim \mathrm{Bernoulli}(\pi(x))$$
$$y \sim \begin{cases} \mathcal{N}(\mu(x), \sigma^2(x)) & \text{if } o = 0 \\ \mathcal{N}(0, \sigma_o^2) & \text{if } o = 1 \end{cases}$$

$$\mathcal{L} = -\log \mathcal{N}(y \mid \mu(x), \sigma^2(x)) - \log \mathrm{Bernoulli}(0 \mid \pi(x)) \boxed{- \log \mathrm{Bernoulli}(1 \mid \pi(\tilde{x}))}$$

4. Predictive Distributions



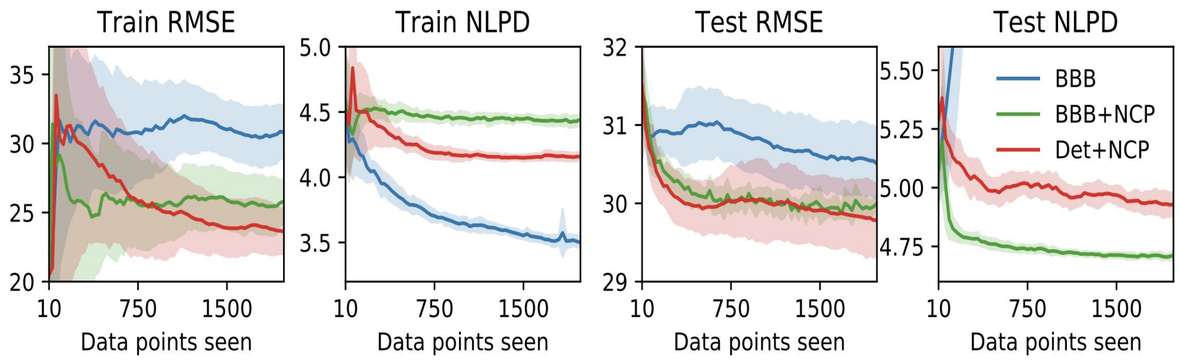
Toy regression active learning where data can only be acquired within two bands.

Deterministic network and BBB overfit. NCP gives high uncertainty estimates for OOD data.

5. Active Learning

Reliable uncertainty estimates enable faster active learning on several tasks (see the paper).

Select data points by expected information gain.



Flight delay regression data set, 700k data points.

Models with NCP improve while the BNN without NCP severely overfits to the training data points.

6. Large-Scale Regression

New state of the art on the full 700k data points of the flights benchmark (passive learning).

Model	NLPD	RMSE
gPoE (Deisenroth & Ng 2015)	8.1	—
SAVIGP (Bonilla et al. 2016)	5.02	—
SVI GP (Hensman et al. 2013)	—	32.60
HGP (Ng & Deisenroth 2014)	—	27.45
MF (Lakshminarayanan et al. 2016)	4.89	26.57
Bayes by Backprop	**4.38**	**24.59**
Flipout+NCP	**4.38**	**24.71**
Deterministic+NCP	**4.38**	**24.68**