Noise Contrastive Priors for Functional Uncertainty

Danijar Hafner[12], Dustin Tran[1], Alex Irpan[1], Timothy Lillicrap[3], James Davidson[4]

[1] Google Brain [2] University of Toronto [3] DeepMind [4] Third Wave Automation

1 We introduce NCP

1. Bayesian neural networks (BNNs) provide uncertainty estimates by modeling a belief distribution over weights. However, it remains open how to specify the prior.
2. An independent Normal prior imposes weak constraints on the function posterior, allowing it to generalize in unforeseen ways outside of the data distribution.
3. We propose Noise Contrastive Priors (NCP) to obtain robust uncertainty estimates by training the model to output high uncertainty outside of the training distribution.
4. For this, we define an input prior, which adds noise to the current mini-batch, and an output prior, which is a wide distribution given these inputs.
5. Our contribution is a simple and scalable method to train any uncertainty-aware model towards high uncertainty on out of distribution (OOD) inputs.

2 Noise Contrastive Priors

Ideally, we select the BNN prior to assign high uncertainty to OOD data. But expressing this as distribution over weights is difficult.

We use a *data prior* instead to express the idea directly on inputs:

$$p_{\text{prior}}(x, y) = p_{\text{prior}}(x)\, p_{\text{prior}}(y \mid x)$$

This means we can choose an input prior and an output prior.

1. Inputs: copy the mini-batch and augment it with noise (gives contrastive inputs that are OOD but near the data set boundary):

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

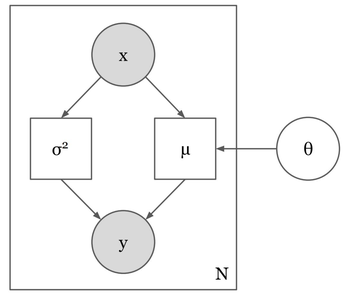
2. Set prior on variables of the model (e.g. the model's prediction) such that predictive variance is high on the OOD inputs:

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

We center the output prior around the targets for the mini-batch to encourage the network OOD to generalize yet be uncertain.

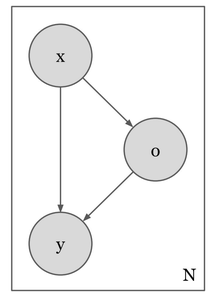
The resulting loss function minimizes the cross entropy both on training data and imagined "pseudo-data" from the prior.

3 Uncertainty-Aware Models



Bayesian Neural Network with NCP on μ (BNN+NCP)

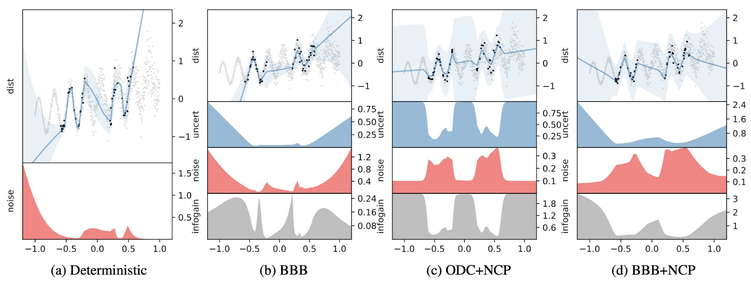
$$\theta \sim q_\phi(\theta)$$
$$y \sim \text{Normal}(\mu(x, \theta), \sigma^2(x))$$
$$\mathcal{L} = -\,\mathbb{E}_{p_{\text{train}}(x,y)}\left[\mathbb{E}_{q_\phi(\theta)}[\ln p(y \mid x, \theta)]\right] + D_{\text{KL}}[\text{Normal}(\mu_\mu, \sigma_\mu^2) \parallel q(\mu(\tilde{x}))]$$



Out of Distribution Classifier with NCP on o (ODC+NCP)

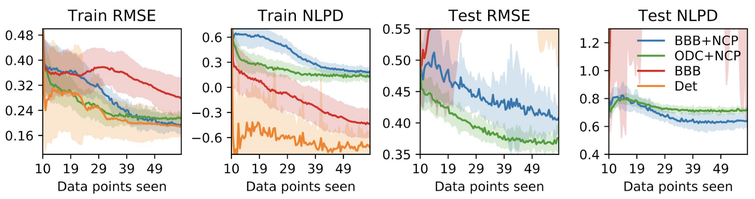
$$o \sim \text{Bernoulli}(\pi(x))$$
$$y \sim \begin{cases} \mathcal{N}(\mu(x), \sigma^2(x)) & \text{if } o = 0 \\ \mathcal{N}(0, \sigma_o^2) & \text{if } o = 1 \end{cases}$$
$$\mathcal{L} = -\log \mathcal{N}(y \mid \mu(x), \sigma^2(x)) - \log \text{Bernoulli}(0 \mid \pi(x)) - \log \text{Bernoulli}(1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

4 Motivating Regression Task

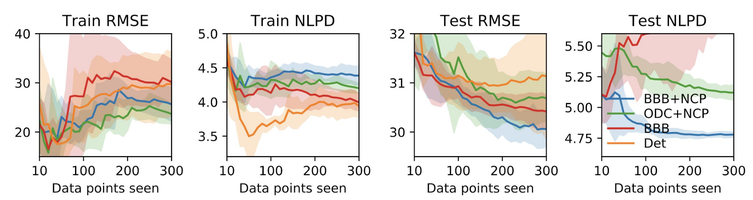


(a) Deterministic (b) BBB (c) ODC+NCP (d) BBB+NCP

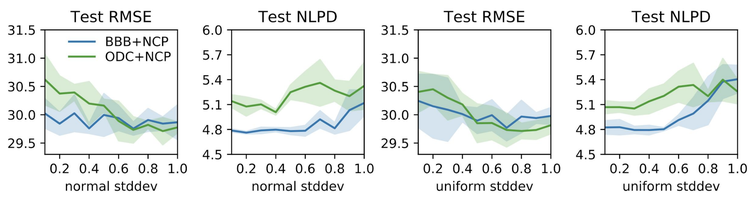
Active learning on 1D regression task where targets can be acquired only within two bands. The deterministic network and Bayes by Backprop (BBB) generalize in unforeseen ways outside of the distribution of visible data. NCP trains these OOD predictions towards a prior, leading to successful active learning.



5 Active Learning for Flights Delays



Actively learning flight delays. Models trained with NCP achieve substantially lower negative log predictive density (NLPD) and BBB+NCP achieves the lowest root mean squared error (RSME).



Robustness to different input noise patterns. Curves show final test performance (lower is better). NCP is robust to the type of input noise and improves over the baselines in all settings.

6 Large-Scale Regression

To evaluate the scalability of NCP, we train on the full 700k data points of the flights dataset in a passive learning setting, where we outperform all previously published work.

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
BBB	**4.38**	**24.59**
BBB+NCP	**4.38**	**24.71**
ODC+NCP	**4.38**	**24.68**

In conclusion, we find NCP effective for low-dim regression tasks. In the future, it should be investigated how this applies to images, e.g. using data augmentation as the input noise.

Contact: mail@danijar.com Twitter: @danijarh

Project website: danijar.com/ncp