Probing Uncertainty Estimates of Neural Processes

Aditya Grover[1], **Dustin Tran**[2], **Rui Shu**[1], **Ben Poole**[2], **Kevin Murphy**[2]
[1]Stanford University, USA, [2]Google Brain, USA
{adityag,ruishu}@cs.stanford.edu, {trandustin,pooleb,kpmurphy}@google.com

Abstract

A *neural process* defines a family of exchangeable stochastic processes parameterized by deep neural networks. Many variants have been recently proposed and applied to tasks involving few-shot regression. While the specific design choices imposed can significantly affect the quality of uncertainty estimates, the tools for analyzing the inductive biases of probabilistic regression models are lacking, as standard metrics such as log-likelihoods directly focus on direct downstream performance on a specific held-out set. In this work, we analyze the uncertainty estimates obtained via neural processes by proposing a series of metrics that probe the model along various interpretable axis. Such a fine-grained analysis can be useful for model criticism and selection with respect to new tasks and datasets.

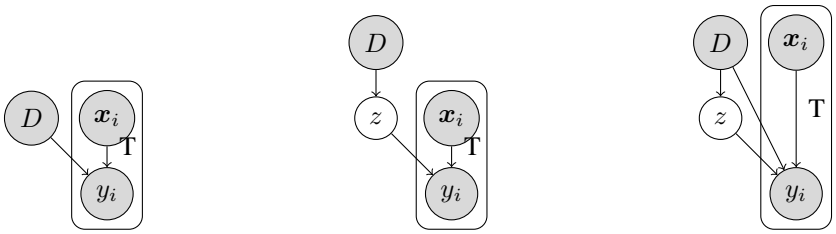
1 Introduction

Recent advancements in probabilistic inference coupled with deep learning has led to a large variety of expressive, uncertainty-aware probabilistic models for regression, such as Bayesian neural networks [Neal, 2012] and deep ensembles [Lakshminarayanan et al., 2017]. A *neural process* contributes to this line of work by proposing a family of exchangeable stochastic processes parameterized via deep neural networks [Garnelo et al., 2018a]. Many variants [Garnelo et al., 2018b, Kim et al., 2019] have been proposed that differ primarily in either the modeling assumptions that define the generative process (e.g., specifying latent variables, inducing points), the inference procedure (e.g., variational posterior family), or the design of neural network parameterizations (e.g., attention, convolutions). These choices can have significant affect on the downstream performance across tasks [Le et al., 2018]. A systematic analysis of the effect of these choices is useful for model criticism and selection on new datasets and tasks. Such an analysis requires careful ablations as well as interpretable statistics that offer insights into the inductive biases of these models.

The current tools available for such an uncertainty-aware analysis of regression models fall into two extremes. On one hand, we have task-specific metrics (e.g., held-out log-likelihoods and mean squared error for regression, cumulative and simple regret for bandit problems) for directly assessing downstream performance of a model. However, the usefulness of these metrics for model criticism and selection on new datasets and tasks can be limited as the metrics lack interpretability, the end tasks could involve regression only as a sub-routine within a larger pipeline (e.g., Bayesian optimization), and the ground truth is available only for a finite set of data points. On the other extreme, one can obtain qualitative insights via visualizations of predictive distributions on low-dimensional *known* data distributions, such as Gaussian mixture models. However, often the absence of a statistical flavor to such an analysis is prone to overfitting on selected examples and in general, presents challenges to draw confident conclusions.

In this work, we seek to bridge the gap in different approaches for analyzing predictive distributions of neural processes via several interpretable statistics such as interpolation and extrapolation. These statistics rely on the knowledge of the true data generative process and hence, in in practice, they are evaluated on synthetic data distributions. Hence, these statistics naturally augment qualitative evaluations with a fine-grained statistical analysis of neural processes that preserves interpretability.

4th workshop on Bayesian Deep Learning (NeurIPS 2019), Vancouver, Canada.



(a) CNP (b) NP (c) ANP

Figure 1: Graphical Models for Neural Processes and Variants. T denotes the observed set of points during training and new test points at test time.

2 Proposed Evaluation Methodology

Consider the setting of noisy, non-linear regression over a bounded domain $\mathcal{X}$:

$$y = f(\boldsymbol{x}) + \epsilon(\boldsymbol{x}) \tag{1}$$

where $\boldsymbol{x} \sim \text{Uniform}(\mathcal{X})$ and $\epsilon(\boldsymbol{x}) \sim \mathcal{N}(0, \sigma(\boldsymbol{x})^2)$ is Gaussian noise. Here, $\sigma(\boldsymbol{x})$ denotes the data (aka aleatoric) uncertainty. A regression model is trained to fit a finite dataset, say D, that comprises of $(\boldsymbol{x}, y)$ points sampled via the above generative process. We assume the model provides two estimates via the predictive distribution $p(y|\boldsymbol{x}, D)$: a predictive mean function $m(\boldsymbol{x}, D)$ and an uncertainty estimate $s(\boldsymbol{x}, D)$.

For qualitative analysis, it is standard to pre-select, known data distributions for evaluation that allow for unbounded access (i.e., we can query f and σ for all $\boldsymbol{x} \in \mathcal{X}$). For example, a standard practice in prior works is to consider sine curves, Gaussian mixture models, etc. Note that the knowledge of the true data distribution is only used for evaluation and not during learning or test-time inference (which would make the task trivial). Next, we discuss our proposed metrics.

2.1 Self-certainty

The *self-certainty* (SC) statistic measures the discrepancy in the uncertainty metrics at the observed (labelled) datapoints relative to the ground-truth data uncertainty. Let the set of observed $\boldsymbol{x}$ be denoted as $D_{\boldsymbol{x}} = \{\boldsymbol{x} | (\boldsymbol{x}, y) \in D\}$. Then, the self-certainty statistic is defined as:

$$SC = \sum_{\boldsymbol{x} \in \mathcal{X}} \|s(\boldsymbol{x}, D) - \sigma(\boldsymbol{x})\|^2. \tag{2}$$

When there is no data noise (i.e., $\sigma(\boldsymbol{x}) = 0$ for all $\boldsymbol{x}$), this property alludes to the standard requirement of an *interpolator* which any flexible enough model can satisfy. For an ideal probabilistic regression model, we expect the self-certainty to collapse as we collect additional data.

2.2 Inclusion@k

The *inclusion* statistic is parameterized by a positive real $k > 0$ and measures the expected coverage of the target mean function f given upper and lower bounds on the predictive mean. Formally:

$$I(k) = \mathbb{E}_{\boldsymbol{x} \sim \text{Uniform}(\mathcal{X})}[\mathbb{1}(|f(\boldsymbol{x}) - m(\boldsymbol{x}, D)| < ks(\boldsymbol{x}, D))] \tag{3}$$

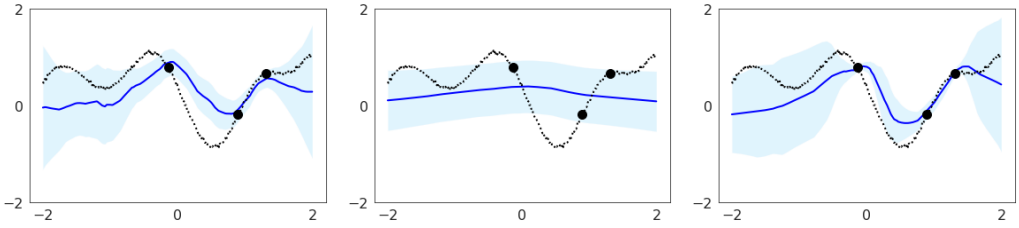
We evaluate the expectation via Monte-Carlo in practice. The above statistic can be used a proxy for calibration in regression based-scenarios.

2.3 Uncertainty-increase@δ

To probe uncertainty at unobserved points, we propose to evaluate their uncertainties relative to the nearest neighbor points. For every test point $\boldsymbol{x}$, we check if its uncertainty estimate $s(\boldsymbol{x}, D)$ exceeds that of its nearest neighbor in D, denoted as $NN(\boldsymbol{x}, D)$. For low-dimensional distributions, the ℓ_1 or ℓ_2 distance could be used directly in the space defined over $\boldsymbol{x}$ whereas for high-dimensional

Table 1: Statistical Evaluation of Neural Process Variants averaged over 1000 trials.

Model	NLL	MSE	Self-certainty (SC)
CNP	**0.626** $\pm$ 0.042	**0.232** $\pm$ 0.01	0.057 $\pm$ 0.003
NP	0.858 $\pm$ 0.037	0.361 $\pm$ 0.014	0.273 $\pm$ 0.009
ANP	0.834 $\pm$ 0.083	0.247 $\pm$ 0.011	**0.043** $\pm$ 0.004



(a) CNP (b) NP (c) ANP

Figure 2: Qualitative vizualizations of uncertainty for a test function and three context points (bold black). Predictive distribution in blue and target function in black.

structured distributions, we could consider alternate feature spaces. With no data noise, we expect higher uncertainty at unobserved points than at observed points i.e., $NN(\boldsymbol{x}, D)$ vs. $NN(\boldsymbol{x}, \mathcal{X} - D)$.

To obtain aggregate statistics, we bin the test points by their distances to the nearest neighbor in D. For any bin, let δ denote the average distance between the bin points and their respective nearest neighbors in D. The interpolation/extrapolation behavior for all test points $\boldsymbol{x}_\delta$ at bin distance δ can then be summarized via the following uncertainty-increase (UI) statistic:

$$UI(\delta) = \frac{\sum_{\boldsymbol{x} \in \boldsymbol{x}_\delta} \mathbb{1}(s(\boldsymbol{x}, D) > NN(\boldsymbol{x}, D))}{|\boldsymbol{x}_\delta|} \tag{4}$$

2.4 Uncertainty-reduction@k

The *uncertainty-reduction* (UR) statistic measures the expected reduction in uncertainty as new points are added to the dataset. Let D'_k be a dataset consisting of k labelled points sampled independently via the data generating process in Eq. 1. The UR statistic can be formally expressed as:

$$UR(k) = \mathbb{E}_{\boldsymbol{x} \sim \text{Uniform}(\mathcal{X})}[\mathbb{E}_{D, D'_k} \| s(\boldsymbol{x}, D \cup D'_k) - s(\boldsymbol{x}, D) \|] \tag{5}$$

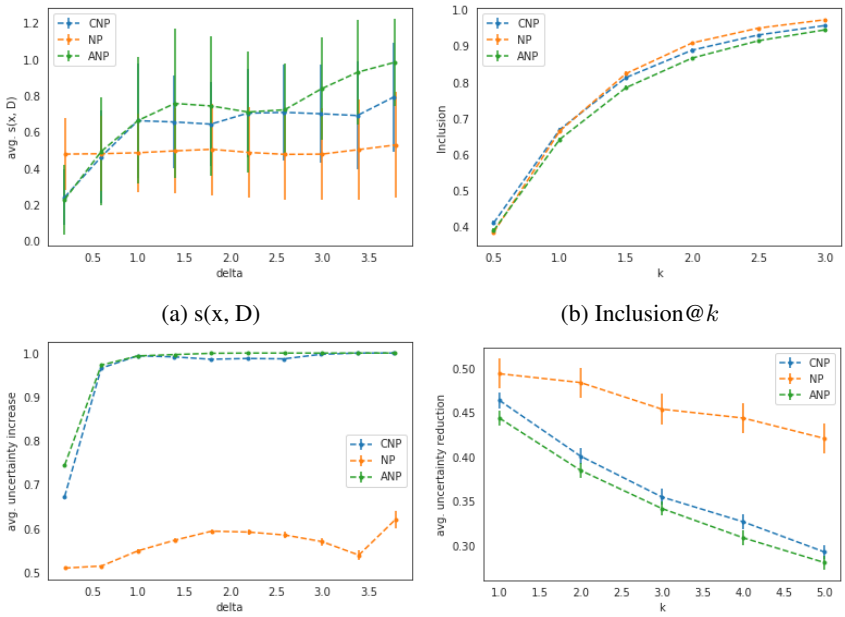
In practice, we evaluate the expectation via Monte-Carlo.

3 Experiments on Neural Processes

We compare three variants of Neural Processes:

1. Conditional Neural Processes (CNP) which map the observed dataset and a test point $\boldsymbol{x}$ directly to its prediction $\boldsymbol{y}$ [Garnelo et al., 2018b]. These models do not include any latent variables.
2. Neural Processes (NP) which modify the CNP model to include global latents [Garnelo et al., 2018a]
3. Attentive Neural Processes (ANP) which include global latents as well as an attentive mapping from the observed dataset to the predictions y [Kim et al., 2019]

The graphical models for the three approaches are shown in Figure 1. We consider the 1D synthetic regression setup followed in all the above three prior works. The distribution over target regression functions is a GP with zero mean and a squared-exponential kernel and randomly sampled kernel hyperparameters. After training, we sample 1000 functions for testing, with number of context points varying from 1 to 10.



(a) s(x, D)

(b) Inclusion@k

(c) Uncertainty-increase@δ

(d) Uncertainty-reduction@k

Figure 3: Quantitative evaluation of interpretable metrics across variants of neural processes.

The qualitative and quantitative results are shown in the Figures 2, 3 and Table 1 above. In Table 1, we see that while CNP outperforms ANP and NP on global metrics such as negative log-likelihoods (NLL) and mean-squared error (MSE), it has a lower self-certainty score than ANP. This is confirmed qualitatively in Figure 2a vs. c.

NPs which are known to underfit [Kim et al., 2019] also show significantly different behavior on all the metrics shown in Figure 3. For NPs, the uncertainty estimate $s(x, D)$ and uncertainty increase is almost independent of δ while the reduction in uncertainty is much slower as a function of k relative to CNPs and ANPs. From Figure 3c, d, we also see that rate of uncertainty increase (as we move away from observed data) and decrease (as we add more data) is superlinear (in δ) for CNPs and ANPs and linear (in k) respectively for all the three models.

References

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413, 2017.

Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J Rezende, SM Eslami, and Yee Whye Teh. Neural processes. *arXiv preprint arXiv:1807.01622*, 2018a.

Marta Garnelo, Dan Rosenbaum, Chris J Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo J Rezende, and SM Eslami. Conditional neural processes. *arXiv preprint arXiv:1807.01613*, 2018b.

Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh. Attentive neural processes. *arXiv preprint arXiv:1901.05761*, 2019.

Tuan Anh Le, Hyunjik Kim, Marta Garnelo, Dan Rosenbaum, Jonathan Schwarz, and Yee Whye Teh. Empirical evaluation of neural process objectives. In *NeurIPS workshop on Bayesian Deep Learning*, 2018.