Scale Mixtures of Neural Network Gaussian Processes

Hyungi Lee[1] Eunggu Yun[1] Hongseok Yang[1,2,3] Juho Lee[1,4]

[1]Kim Jaechul Graduate School of AI, KAIST, South Korea

[2]School of Computing, KAIST, South Korea

[3]Discrete Mathematics Group, Institute for Basic Science (IBS), South Korea

[4]AITRICS, South Korea

ICLR 2022

Neural Network Gaussian Processes

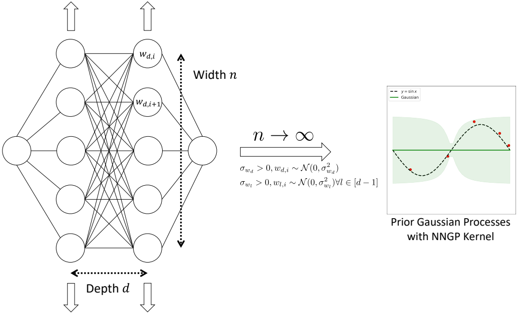

Figure: At Initialization.

- When n goes to infinity, the output of a neural network at initialization converges to a Gaussian Process with NNGP kernel [Neal, 1996, Lee et al., 2018].

Related Works

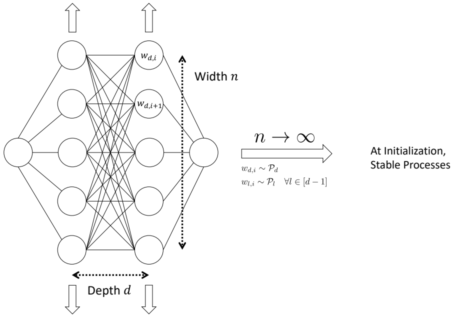

Figure: At Initialization.

- Under an alternative prior specification, the output of a neural network converges to a stable process [Favaro et al., 2020, Bracale et al., 2021].

Limitations

- Convergence results after gradient descent training for only the readout layer or all layers when using Gaussian initialization [Lee et al., 2019].
- Hard to sample and inference for a stable process.
- Limited neural network structures.

Scale Mixture of NNGPs

- Putting a prior distribution on the scale of the readout-layer parameters lets the initial distribution be the following scale mixture of gaussian distribution:

$$\sigma_{w_d}^2 \sim \mathcal{H}, \quad w_{d,i} | \sigma_{w_d}^2 \sim \mathcal{N}(0, \sigma_{w_d}^2)$$
$$\sigma_{w_l} > 0, \quad w_{l,i} \sim \mathcal{N}(0, \sigma_{w_l}^2) \quad \forall l \in [d-1].$$

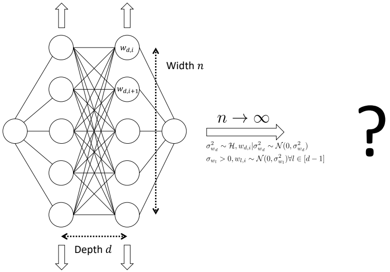

Figure: Our approach.

Scale Mixture of NNGPs

- Our method only changes the constant scale of the readout-layer parameters into random variable.
- Simple, yet flexible.
- Allows efficient inference algorithms, with comparable cost to those for NNGPs.

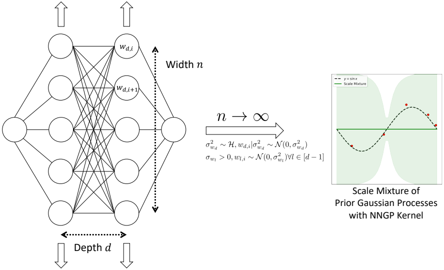

Figure: Our approach.

Heavy tail features of the output distribution

- If we use inverse gamma distribution as prior on the scale, we get Student's t process which has a heavy tail.

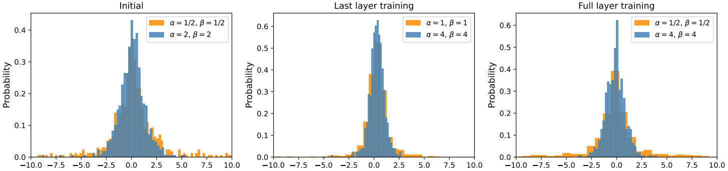

Figure: Impact of the prior hyperparameters to the heaviness of the tail of the output distribution for initial, last layer training and full layer training.

Experimental Results

Table: NLL values on UCI dataset. (m, d) denotes number of data points and features, respectively. We take results from Adlam et al. [2020] except our model.

Dataset	(m, d)	PBP-MV	Dropout	Ensembles	RBF	NNGP	Ours
Boston Housing	(506, 13)	2.54 ± 0.08	$\mathbf{2.40} \pm 0.04$	2.41 ± 0.25	2.63 ± 0.09	2.65 ± 0.13	2.72 ± 0.05
Concrete Strength	(1030, 8)	3.04 ± 0.03	$\mathbf{2.93} \pm 0.02$	3.06 ± 0.18	3.52 ± 0.11	3.19 ± 0.05	3.13 ± 0.04
Energy Efficiency	(768, 8)	1.01 ± 0.01	1.21 ± 0.01	1.38 ± 0.22	0.78 ± 0.06	1.01 ± 0.04	$\mathbf{0.67} \pm 0.04$
Kin8nm	(8192, 8)	$\mathbf{-1.28} \pm 0.01$	-1.14 ± 0.01	-1.20 ± 0.02	-1.11 ± 0.01	-1.15 ± 0.01	-1.18 ± 0.01
Naval Propulsion	(11934, 16)	-4.85 ± 0.06	-4.45 ± 0.00	-5.63 ± 0.05	$\mathbf{-10.07} \pm 0.01$	-10.01 ± 0.01	-8.04 ± 0.04
Power Plant	(9568, 4)	2.78 ± 0.01	2.80 ± 0.01	2.79 ± 0.04	2.94 ± 0.01	2.77 ± 0.02	$\mathbf{2.66} \pm 0.01$
Wine Quality Red	(1588, 11)	0.97 ± 0.01	0.93 ± 0.01	0.94 ± 0.12	-0.78 ± 0.07	$\mathbf{-0.98} \pm 0.06$	-0.77 ± 0.07
Yacht Hydrodynamics	(308, 6)	1.64 ± 0.02	1.25 ± 0.01	1.18 ± 0.21	0.49 ± 0.06	1.07 ± 0.27	$\mathbf{0.17} \pm 0.25$

- Our model shows robust results on the classification tasks.

Summary of our results

- With a simple extension of NNGPs by introducing a scale prior on the last layer weight parameters, we get a broad class of stochastic processes, especially heavy-tailed ones such as Student's t processes.

References

Ben Adlam, Jaehoon Lee, Lechao Xiao, Jeffrey Pennington, and Jasper Snoek. Exploring the uncertainty properties of neural networks' implicit priors in the infinite-width limit. *arXiv preprint arXiv:2010.07355*, 2020.

Daniele Bracale, Stefano Favaro, Sandra Fortini, and Stefano Peluchetti. Infinite-channel deep stable convolutional neural networks. *arXiv preprint arXiv:2102.03739*, 2021.

Stefano Favaro, Sandra Fortini, and Peluchetti Stefano. Stable behaviour of infinitely wide deep neural networks. In *23rd International Conference on Artificial Intelligence and Statistics (AISTATS 2020)*. (seleziona...), 2020.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. In *International Conference on Learning Representations*, 2018.

Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32, 2019.

Radford M Neal. Priors for infinite networks. In *Bayesian Learning for Neural Networks*, pages 29–53. Springer, 1996.