# Shaping the Neural Linear Model

Max Guo

Data to Actionable Knowledge Lab
Summer Program for Undergraduates in Data Science (SPUDS)
Harvard University

Summer 2022

# Research Overview

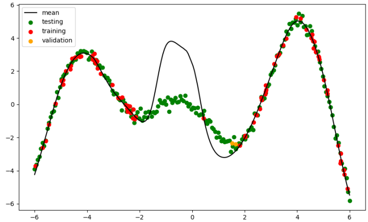
**Problems:**

- Deep neural networks do not provide predictive uncertainties.
- Probabilistic inference via Gaussian Processes (GPs) is computationally expensive and requires complicated tuning to perform well on some types of data.
- Combining Bayesian inference with neural networks via a model known as the *Neural Linear Model* (NLM) may suffer from overfitting to the data and tuning issues.

**My Research:**

- Develop and analyze variations of the NLM that can combine the strengths of the traditional NLM and fixed-kernel GPs.

# Lack of Predictive Uncertainty in Neural Networks



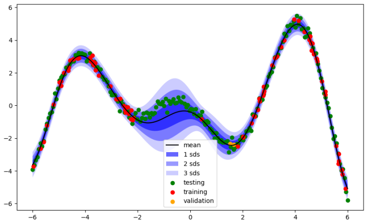




Figure: Feedforward Neural Network vs. Gaussian Process with RBF Kernel

# Neural Linear Model (NLM)

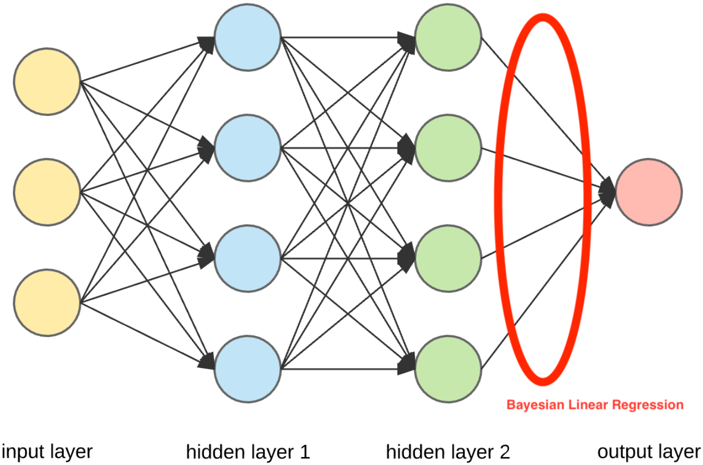



Figure: Neural Linear Model, Visual

# Predictive Uncertainty in NLMs

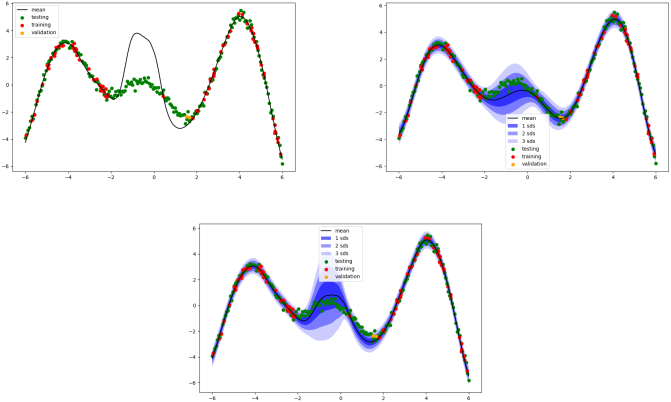



Figure: Top: Feedforward Neural Network (Left) GP (Right)
Bottom: NLM

# Problems with NLMs

- NLMs also suffer from problems:
  - NLMs require much hyperparameter tuning
  - NLMs may not express desirable uncertainties in areas of data sparsity.

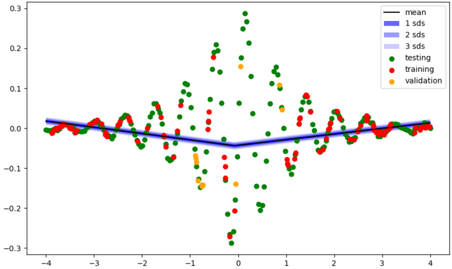



Figure: Trained NLM does not fit data at all.

# Problems with GPs

- Besides computational issues, fixed-kernel GPs do not adapt well to varying data attributes across the input space (e.g. smoothness).

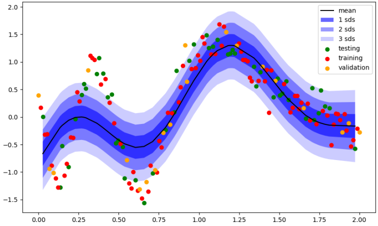



Figure: Fixed-kernel GP does not capture all of the function.

# Research Goal

**Goal**: Create a variation on the NLM to capture strengths of both NLMs and fixed-kernel GPs.

**Goal, Rephrased**: Because NLMs are GPs with data-adaptive kernels, can we create a spectra of models that interpolates between NLMs and GPs?


Figure: Slider between NLM and Fixed-kernel GP

# Proposed Models

I consider three variations on NLMs that can interpolate between a traditional NLM and a fixed-kernel GP:

1. NLMSubset
2. NLMDecoupled
3. NLMRegularize

# NLMSubset and NLMDecoupled

- Train a subset of the bases normally
- Train the other subset to match a fixed-kernel GP on mini-batches of data in terms of the Gram matrix.
- (For NLMSubset) Combine the losses for the loss function:

$$L(\theta) = L_1(\theta_1) + \lambda \cdot L_2(\theta_2)$$

- Number of bases is the sliding parameter.

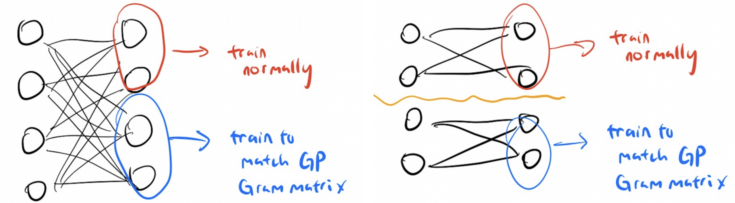



Figure: NLMSubset (left), NLMDecoupled (right)

# NLMRegularize

- Modify the training objective so all of the bases jointly attempt to fit the data and match the GP Gram matrix:

$$L(\theta) = L_1(\theta) + \lambda \cdot L_2(\theta)$$

- Regularizing constant $\lambda$ is the sliding parameter.

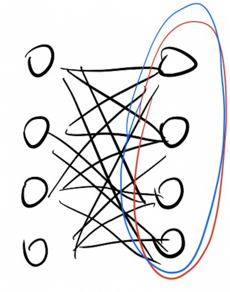


Figure: NLMRegularize

# Preliminary Results

On synthetic 1 dimensional datasets, these methods often give us desirable solutions to the problems that traditional NLMs and GPs present:

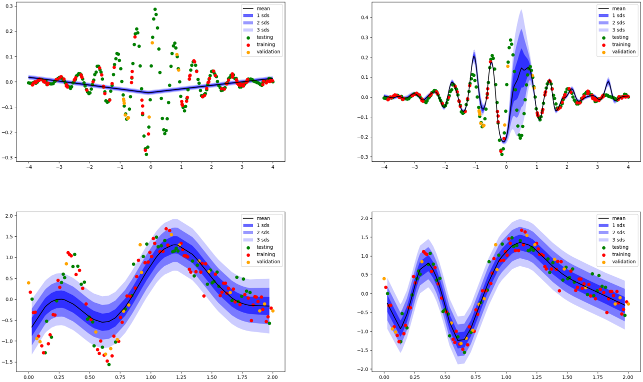
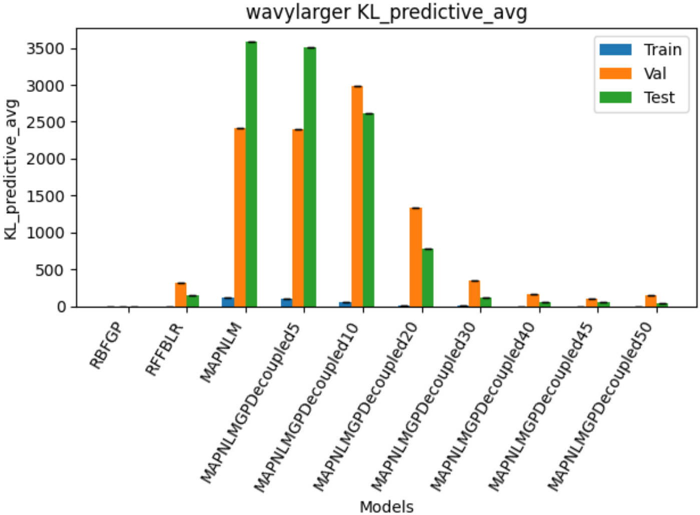



Figure: Wavy (top) and nonstationary (bottom) datasets, where NLMs and GPs fail, respectively (left). NLMRegularize (right) succeeds.

# Preliminary Results

On synthetic 1 dimensional datasets, each of these methods gives us a desired interpolation of models:



# Future Directions

Though the three models are relatively successful in carefully constructed, simple scenarios, more research is required before ascertaining conclusions.

Questions that need to be addressed:

- What distinguishes these three types of models from each other?
- Does the interpolation for these models generalize to higher dimensional, real-world datasets?
- How do these methods compare to kernel-learning in GPs?

[ACKS20] Isac Arnekvist, J. Frederico Carvalho, Danica Kragic, and Johannes A. Stork. The effect of Target Normalization and Momentum on Dying ReLU, May 2020. Number: arXiv:2005.06195 arXiv:2005.06195 [cs, stat].

[ACS+19] Vahdat Abdelzad, Krzysztof Czarnecki, Rick Salay, Taylor Denounden, Sachin Vernekar, and Buu Phan. Detecting Out-of-Distribution Inputs in Deep Neural Networks Using an Early-Layer Output, October 2019. Number: arXiv:1910.10307 arXiv:1910.10307 [cs, stat].

[ADH+19] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On Exact Computation with an Infinitely Wide Neural Net. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

# References II


[BOFG] Davis Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle, and John Guttag. What is the State of Neural Network Pruning? page 18.

[CS14] Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, January 2014.

[CSHvdV15] Ismaël Castillo, Johannes Schmidt-Hieber, and Aad van der Vaart. Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5), October 2015.

[DSV20] Laya Das, Abhishek Sivaram, and Venkat Venkatasubramanian. Hidden representations in deep neural networks: Part 2. Regression problems. *Computers & Chemical Engineering*, 139:106895, August 2020.

# References III

[FAL17] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pages 1126–1135. PMLR, 2017.

[GDMB11] Gonzalo Garcia-Donato and Miguel Angel Martinez-Beneito. Inferences in Bayesian variable selection problems with large model spaces. Technical Report arXiv:1101.4368, arXiv, January 2011. arXiv:1101.4368 [stat] type: article.

[HAMS20] Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-Learning in Neural Networks: A Survey. Technical Report arXiv:2004.05439, arXiv, November 2020. arXiv:2004.05439 [cs, stat] type: article.

# References IV


[HC14] P. Richard Hahn and Carlos M. Carvalho. Decoupling shrinkage and selection in Bayesian linear models: a posterior summary perspective. Technical Report arXiv:1408.0464, arXiv, August 2014. arXiv:1408.0464 [stat] type: article.

[HPTT16] Hengyuan Hu, Rui Peng, Yu-Wing Tai, and Chi-Keung Tang. Network Trimming: A Data-Driven Neuron Pruning Approach towards Efficient Deep Architectures. Technical Report arXiv:1607.03250, arXiv, July 2016. arXiv:1607.03250 [cs] type: article.

[JGH18a] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

# References V


[JGH18b] Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[Jol82] Ian T. Jolliffe. A Note on the Use of Principal Components in Regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 31(3):300–303, 1982. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.2307/2348005.

[KAKP22] M. Karagiannopoulos, Dionysios Anyfantis, Sotiris Kotsiantis, and P. Pintelas. Feature selection for regression problems. *Proceedings of HERCMAâ07*, June 2022.

[Kum17] Siddharth Krishna Kumar. On weight initialization in deep neural networks. Technical Report arXiv:1704.08863, arXiv, May 2017. arXiv:1704.08863 [cs] type: article.

# References VI

[LBG11] Fraser Lewis, Adam Butler, and Lucy Gilbert. A unified approach to model selection using the likelihood ratio test. *Methods in Ecology and Evolution*, 2(2):155–162, 2011. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.2041-210X.2010.00063.x.

[LBN+18] Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep Neural Networks as Gaussian Processes. Technical Report arXiv:1711.00165, arXiv, March 2018. arXiv:1711.00165 [cs, stat] type: article.

[LSSK20] Lu Lu, Yeonjong Shin, Yanhui Su, and George Em Karniadakis. Dying ReLU and Initialization: Theory and Numerical Examples. *Communications in Computational Physics*, 28(5):1671–1706, June 2020. arXiv:1903.06733 [cs, math, stat].

[LWL+17] Weibo Liu, Zidong Wang, Xiaohui Liu, Nianyin Zeng, Yurong Liu, and Fuad E. Alsaadi. A survey of deep neural network architectures and their applications. *Neurocomputing*, 234:11–26, April 2017.

[LXMZ] Tao Luo, Zhi-Qin John Xu, Zheng Ma, and Yaoyu Zhang. Phase Diagram for Two-layer ReLU Neural Networks at Infinite-width Limit. page 47.

# References VIII

[LZM+17] Nir Levine, Tom Zahavy, Daniel J Mankowitz, Aviv Tamar, and Shie Mannor. Shallow updates for deep reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

[MAT+20] Hartmut Maennel, Ibrahim Alabdulmohsin, Ilya Tolstikhin, Robert J. N. Baldock, Olivier Bousquet, Sylvain Gelly, and Daniel Keysers. What Do Neural Networks Learn When Trained With Random Labels?, November 2020. Number: arXiv:2006.10455 arXiv:2006.10455 [cs, stat].

[MBW20] Wesley J. Maddox, Gregory Benton, and Andrew Gordon Wilson. Rethinking Parameter Counting in Deep Models: Effective Dimensionality Revisited. Technical Report arXiv:2003.02139, arXiv, May 2020. arXiv:2003.02139 [cs, stat] type: article.

# References IX

[NBMS17] Behnam Neyshabur, Srinadh Bhojanapalli, David Mcallester, and Nati Srebro. Exploring Generalization in Deep Learning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[noa] Introduction: what is overdispersion?

[NRK21] Thao Nguyen, Maithra Raghu, and Simon Kornblith. Do Wide and Deep Networks Learn the Same Things? Uncovering How Neural Network Representations Vary with Width and Depth. Technical Report arXiv:2010.15327, arXiv, April 2021. arXiv:2010.15327 [cs] type: article.

[OC16] Chris Olah and Shan Carter. Attention and Augmented Recurrent Neural Networks. *Distill*, 1(9):e1, September 2016.

[OR19] Sebastian W. Ober and Carl Edward Rasmussen. Benchmarking the Neural Linear Model for Regression. Technical Report arXiv:1912.08416, arXiv, December 2019. arXiv:1912.08416 [cs, stat] type: article.

[SDNSL20] Jascha Sohl-Dickstein, Roman Novak, Samuel S. Schoenholz, and Jaehoon Lee. On the infinite width limit of neural networks with a standard parameterization. Technical Report arXiv:2001.07301, arXiv, April 2020. arXiv:2001.07301 [cs, stat] type: article.

[SRS+15a] Jasper Snoek, Oren Rippel, Kevin Swersky, Ryan Kiros, Nadathur Satish, Narayanan Sundaram, Md Mostofa Ali Patwary, Prabhat, and Ryan P. Adams. Scalable Bayesian Optimization Using Deep Neural Networks. Technical Report arXiv:1502.05700, arXiv, July 2015. arXiv:1502.05700 [stat] type: article.

# References XI


[SRS+15b] Jasper Snoek, Oren Rippel, Kevin Swersky, Ryan Kiros, Nadathur Satish, Narayanan Sundaram, Mostofa Patwary, Mr Prabhat, and Ryan Adams. Scalable bayesian optimization using deep neural networks. In *International conference on machine learning*, pages 2171–2180. PMLR, 2015.

[Ste69] G. W. Stewart. On the Continuity of the Generalized Inverse. *SIAM Journal on Applied Mathematics*, 17(1):33–45, 1969. Publisher: Society for Industrial and Applied Mathematics.

[TLY+21] Sujay Thakur, Cooper Lorsung, Yaniv Yacoby, Finale Doshi-Velez, and Weiwei Pan. Uncertainty-Aware (UNA) Bases for Deep Bayesian Regression Using Multi-Headed Auxiliary Networks. Technical Report arXiv:2006.11695, arXiv, December 2021. arXiv:2006.11695 [cs, stat] type: article.

# References XII

[TMK17] Surat Teerapittayanon, Bradley McDanel, and H. T. Kung. BranchyNet: Fast Inference via Early Exiting from Deep Neural Networks, September 2017. Number: arXiv:1709.01686 arXiv:1709.01686 [cs].

[YH20] Greg Yang and Edward J Hu. Feature learning in infinite-width neural networks. *arXiv preprint arXiv:2011.14522*, 2020.

[YH21] Greg Yang and Edward J. Hu. Feature Learning in Infinite-Width Neural Networks. Technical Report arXiv:2011.14522, arXiv, May 2021. arXiv:2011.14522 [cond-mat] type: article.