Multitask Learning for Neural Network Regularization

Julia Nakhleh, Robert Nowak
University of Wisconsin-Madison
jnakhleh@wisc.edu

ETI Annual Workshop, February 20-21, 2024

20

Introduction and Problem Overview

Multi-task learning [1] is a paradigm in which a machine learning model is trained to perform multiple tasks simultaneously. Multi-task learning has been observed to improve model generalization performance on predictive problems in many domains, particularly when performed with neural networks [2]. Typically, the tasks are assumed to be related but distinct, and the intuitive motivation for the multi-task approach is that the similarities and differences between the tasks may allow the model to learn a better representation of the data than standard single-task training. However, rigorous theoretical understanding of the benefits of multi-task learning with neural networks is still largely lacking.

Results, Methods and Next Steps

Our recent work focuses on a particular instance of multi-task learning: single-layer, finite-width, multi-output ReLU neural networks (see Figure 1), each of which is trained to fit a different set of labels for the same dataset. We view the first set of labels as the “true” labels (obtained from some real-world dataset), and the remaining sets of labels as artificial auxiliary labels which can be numerically generated according to specification. We prove mathematically that, when the data is univariate and the auxiliary labels meet mild geometric assumptions, training such a network to interpolate the data points with minimal vector-valued variation (VV) norm [3] will result in the first network output—which fits the “true” dataset—learning the connect-the-dots interpolant of that dataset. The VV norm of the network, given by $\sum_{k=1}^{K} \|w_k\|_2 \|v_k\|_2$ where w_k and v_k are the input and output weight vectors of neuron k, is the same quantity which is minimized by standard weight decay regularization which is commonly used in neural network training. However, minimizing the VV norm of a multi-output network can lead to significantly different solutions than minimizing the same norm for a single-output network. In the scenario we consider, the weight decay problem for a single-output neural network may have multiple solutions, but the VV norm minimization problem when auxiliary outputs are added to the network will have a unique solution.

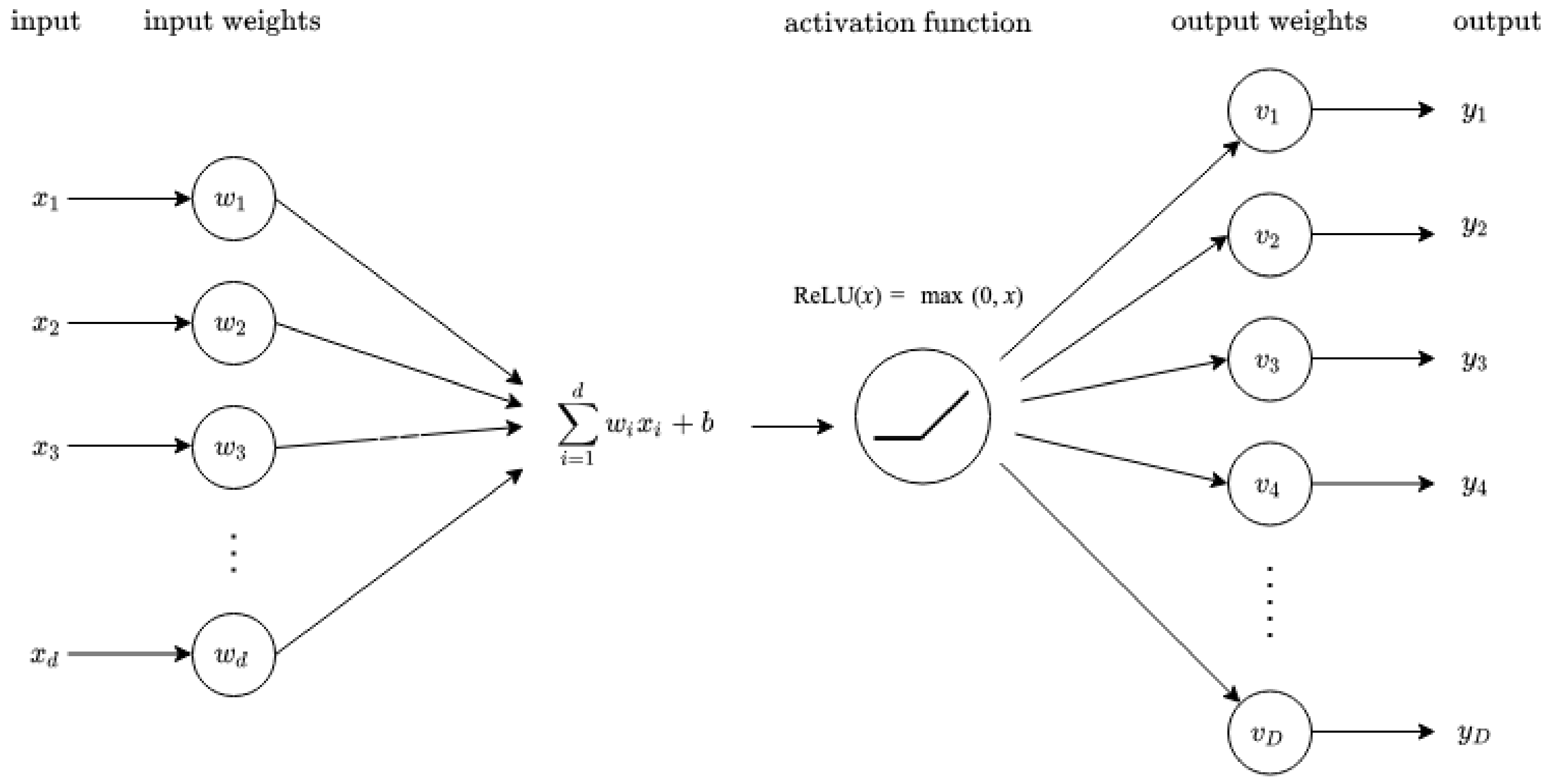

Figure 1: a multi-output ReLU neuron.

Results, Methods and Next Steps (cont.)

We also perform numerical experiments which show that for multivariate data, the aforementioned training procedure with Gaussian auxiliary labels leads the first output to learn functions which are on average smoother (in that they have a smaller average gradient norm) than those obtained using traditional weight decay regularization on the first output alone. See Figure 2 for an example. We observe that the smoothness of the learned functions depends on the variance of the Gaussian auxiliary labels as well as the number of auxiliary outputs, with more auxiliary outputs resulting in smoother functions. We also find that these smooth solutions tend to have smaller average and maximum distance between test and training data points. The mathematical properties of functions learned using random auxiliary tasks on multivariate data is an ongoing research area that we are exploring.

Relevance to Program Objectives

Neural networks exhibit state-of-the-art performance on a wide variety of pattern-recognition tasks, and can be trained as effective anomaly detectors across multiple problem domains. However, their demonstrated success is largely empirical, and many of their mathematical properties are not yet well-understood. Improved understanding of the benefits of multi-task learning for ReLU neural networks is relevant for developing a mathematically rigorous theory of deep learning, and it may help guide machine learning practitioners in their usage of this approach when applying multi-task learning to real-world problems.

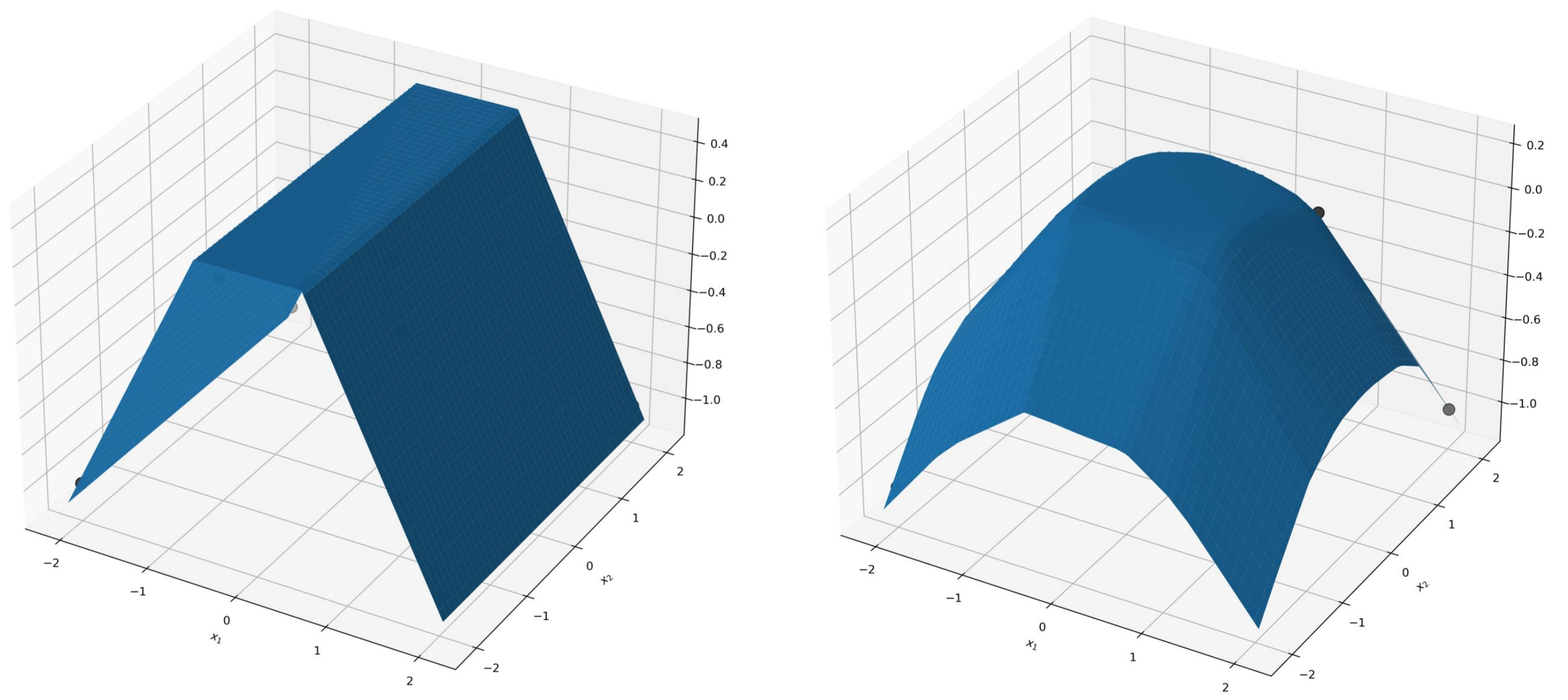

Figure 2: Left: non-smooth solution learned by a single-output ReLU neural network. Right: smooth solution learned by a multi-output ReLU neural network fitting auxiliary Gaussian random labels in addition to the primary dataset (pictured here).

References

[1] R. Caruana, “Multitask Learning,” Mach. Learn., vol. 28, no. 1, pp. 41–75, Jul. 1997, doi: 10.1023/A:1007379606734.

[2] Y. Zhang and Q. Yang, “A Survey on Multi-Task Learning,” IEEE Trans. Knowl. Data Eng., vol. 34, no. 12, pp. 5586–5609, Dec. 2022, doi: 10.1109/TKDE.2021.3070203.

[3] J. Shenouda, R. Parhi, K. Lee, and R. D. Nowak, “Vector-Valued Variation Spaces and Width Bounds for DNNs: Insights on Weight Decay Regularization.” arXiv, May 25, 2023. doi: 10.48550/arXiv.2305.16534.

This material is based upon work supported by the Department of Energy / National Nuclear Security Administration under Award Number(s) DE-NA0003921.