Bull. Aust. Math. Soc. (First published online 2026), page 1 of 2*
doi:10.1017/S0004972726102044
*Provisional—final page numbers to be inserted when paper edition is published

LEARNING ABILITY OF DEEP RELU NETWORKS: PAIRWISE TASKS AND GRADIENT DESCENT METHODS

JUNYU ZHOU

(Received 7 August 2026 UTC)

2020 *Mathematics subject classification*: primary 68T07; secondary 41A46, 62G05, 68Q32.

Keywords and phrases: deep ReLU neural networks, pairwise learning, metric and similarity learning, generalisation bounds, gradient descent and stochastic gradient descent, neural tangent kernel.

Deep neural networks have become a cornerstone of modern machine learning, demonstrating remarkable performance across a wide range of applications. Despite this empirical success, the theoretical understanding of deep neural networks, particularly in the context of generalisation, remains relatively underdeveloped. The rectified linear unit (ReLU) is one of the most widely used activation functions in deep learning, valued for its simplicity and effectiveness in training deep networks. In this thesis, we focus on advancing the theoretical analysis of deep ReLU networks from two perspectives: pairwise learning tasks and gradient descent methods.

For pairwise learning tasks, we are concerned with the generalisation performance of nonparametric estimation. Most of the existing work requires the hypothesis space to be convex or a VC class, and the loss to be convex. However, these restrictive assumptions limit the applicability of the results in studying many popular methods, especially kernel methods and neural networks. We significantly relax these restrictive assumptions and establish a sharp oracle inequality of the empirical minimiser with a general hypothesis space for the Lipschitz continuous pairwise losses. As an application, we apply our general results to study pairwise least squares regression and derive an excess generalisation bound that matches the minimax lower bound for pointwise least squares regression up to a logarithmic term. The key novelty here is to construct a structured deep ReLU neural network as an approximation of the true predictor and design the target hypothesis space consisting of the structured networks with controllable complexity. This successful application demonstrates that the general results obtained indeed help us to explore the generalisation performance on a variety of problems that cannot be handled by existing approaches.

Thesis submitted to the University of Sydney in July 2025; degree approved on 22 August 2025; supervisors Ding-Xuan Zhou and Yiming Ying.

© The Author(s), 2026. Published by Cambridge University Press on behalf of Australian Mathematical Publishing Association Inc.

https://doi.org/10.1017/S0004972726102044 Published online by Cambridge University Press

We study the generalisation performance of metric and similarity learning by leveraging the *specific structure* of the true metric (the target function). Specifically, by deriving the explicit form of the true metric for metric and similarity learning with the hinge loss, we construct a structured deep ReLU neural network as an approximation of the true metric, whose approximation ability relies on the network complexity. Here, the network complexity corresponds to the depth, the number of nonzero weights and the computation units of the network. Considering the hypothesis space which consists of the structured deep ReLU networks, we develop the excess generalisation error bounds for a metric and similarity learning problem by estimating the *approximation error* and the *estimation error* carefully. An optimal excess risk rate is derived by choosing the proper capacity of the constructed hypothesis space. To the best of our knowledge, this is the *first ever known* generalisation analysis providing the excess generalisation error for metric and similarity learning. In addition, we investigate the properties of the true metric of metric and similarity learning with general losses.

For gradient descent methods, we focus on studying the statistical generalisation performance of gradient descent (GD) and stochastic gradient descent (SGD) for overparameterised neural networks within the neural tangent kernel regime. Most of the existing work on regression problems is limited to shallow network architectures, leaving a notable gap in the theory for deep neural networks. We address this gap by presenting a comprehensive generalisation analysis for deep ReLU networks trained using GD and SGD. Specifically, we establish the *first known* minimax-optimal rates of excess population risk for both GD and SGD with deep ReLU networks, under the assumption that the network width scales *polynomially* with respect to the network depths and training sample size. Our results demonstrate that with sufficient widths, gradient descent methods for deep ReLU networks can achieve the optimal rates of generalisation on a par with kernel methods.

Some of these research results have appeared in [1].

Reference

[1] J. Zhou, S. Huang, H. Feng, P. Wang and D.-X. Zhou, 'Fine-grained analysis of nonparametric estimation for pairwise learning', *IEEE Trans. Neural Netw. Learn. Syst.* **37**(9) (2026), 4219–4230.

JUNYU ZHOU, School of Mathematics and Statistics,
The University of Sydney, Sydney, New South Wales 2006, Australia
e-mail: 1842081887zjy@gmail.com

https://doi.org/10.1017/S0004972726102044 Published online by Cambridge University Press