A Survey of Our Results on Provable Neural Training

Anirbit *anirbit.mukherjee@manchester.ac.uk*
Department of Computer Science
The University of Manchester

We give a summary of 8 of our papers on provable neural training, organized into two sections. Section 1 covers our results for depth 2 (i.e. one layer of activation) for unrestricted widths, and Section 2 covers our provable neural training results for arbitrary depths.

1 Provable Learning for Depth 2 Nets at Any Width

A versatile model of iterative stochastic optimizers attempting to minimize an objective function $L(\boldsymbol{W})$ at a constant step-size $h > 0$, is the Langevin Monte Carlo (LMC) algorithm, defined as,

$$\boldsymbol{W}_{(k+1)h} = \boldsymbol{W}_{kh} - h\nabla L(\boldsymbol{W}_{kh}) + \sqrt{2}(\boldsymbol{B}_{(k+1)h} - \boldsymbol{B}_{kh})$$

where, $k = 0, 1, 2 \ldots$ & $(\boldsymbol{B}_t)_{t\geq 0}$ is the Brownian motion. Albeit the proofs of LMC's convergence under various conditions and evidence of it being a good neural trainer, since the key empirical study in (Neelakantan et al., 2015) — and yet, till our following works, this fundamental question had remained unknown, if LMC provably converges in any practical machine learning setting, let alone on any neural nets.

•[1–3] (Kumar et al., 2025; Gopalani & Mukherjee, 2025; Gopalani et al., 2024) In this breakthrough work we uncovered a novel mechanism (via isoperimetric inequalities) for proving the convergence of noisy gradient algorithms on neural nets with one layer of activation — for any data and size.

This is the first true "beyond NTK" proof of neural training.

The key idea here is to realize that a *constant amount of regularization* when incorporated into the empirical risk $L(\boldsymbol{W})$ leads to the corresponding Gibbs' measure ($\sim e^{-\beta L(\boldsymbol{W})}$) satisfying the Poincaré inequality. This constant amount of regularization does not depend on the size of the net.

In particular, this work includes a proof for continuous-time noisy-GD (SDE) converging on nets with any number of SoftPlus gates, which, not being Lipschitz smooth is a regime close to real world (since SoftPlus is a differentiable approximation to ReLU) and particularly inaccessible via other proof techniques.

•[4, 5] In (Karmakar & Mukherjee, 2022) and (Karmakar et al., 2023) classes of neural nets with one ReLU gate and a generalized convolution layer were identified, respectively, for which stochastic "pseudo-gradient" algorithms can be given that provably converge exponentially fast under realizable data conditions. We note that, except for these, there are no other stochastic algorithms known for these settings that are this fast under similar data assumptions.

•[6] (Arora et al., 2018) was my fist paper as a PhD student. I had initiated this study and this was the first paper in neural nets for all the other senior authors, including my PhD adviser. I had invented the preliminary version of all the theorems here. This paper went on to be highly cited and has become a part of the syllabus of certain deep-learning courses around the world.

This paper showed for the first time that there exists algorithms that can find the global minima of ReLU neural net loss functions for arbitrary width and data, in time polynomial in the number of data points. This paper also presented new circuit complexity results for nets, that there exists a continuum of deep neural net functions that need the width to be super-exponential in the presented depth, to be approximated at lower depths. The techniques in this paper opened up new avenues of interaction between neural nets and polyhedral theory and a number of groups around the world have developed these ideas further. *Some of the conjectures we made in (Arora et al., 2018) have only recently begun to be settled, (Froese & Hertrich, 2023)*

A stream of literature has now developed, that tries to get similar fast run-time exact optimization algorithms for high-depth nets. These pursuits have opened up new interfaces with machine learning of the classical field of structure-sensitive algorithm design for NP-Hard optimization questions.

2 Provable Learning for Deep Nets

It has scarcely been made rigorous as to why "adaptive"/current gradient dependent step-schedules are needed for optimal performance in deep-learning. We have made two foundational progresses in this field.

•[7] In (Tucat et al., 2025) we invented $\delta-$GClip, which is the following form of step-length scheduled Gradient Descent, which is the *first provably convergent step-schedule for deep-learning*, $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - h(\boldsymbol{x}_t) \cdot \nabla f(\boldsymbol{x}_t)$, with $h(\boldsymbol{x}_t) \coloneqq \eta \cdot \min\left\{1, \max\left\{\delta, \frac{\gamma}{\|\nabla f(\boldsymbol{x}_t)\|}\right\}\right\}$.

In experiments, we showed that beyond the ambit of the proof, a mini-batched version of this algorithm also matches the best known training on VAEs and standard transformers like ViT and BERT. *Thus, our algorithm $\delta-$GClip is the first provably convergent neural trainer that is also competitive against the best heuristics.*

•[8] Additionally, in (De et al., 2018) we showed the *first* proof of convergence of the ADAM algorithm which is the most widely used deep-learning algorithm.

References

Raman Arora, Amitabh Basu, Poorya Mianjy, and Mukherjee, Anirbit. **Understanding Deep Neural Networks with Rectified Linear Units**. In *International Conference on Learning Representations*, 2018.

Soham De, Mukherjee, Anirbit, and Enayat Ullah. **Convergence guarantees for RMSProp and ADAM in non-convex optimization and their comparison to Nesterov acceleration on autoencoders**. In *ICML Workshop on Modern Trends in Nonconvex Optimization for Machine Learning*, 2018. URL https://doi.org/10.48550/arXiv.1807.06766.

Vincent Froese and Christoph Hertrich. Training neural networks is NP-hard in fixed dimension. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=VAQp2EnZeW.

Pulkit Gopalani and Anirbit Mukherjee. **Global Convergence of SGD On Two Layer Neural Nets**. *Information and Inference: a Journal of the IMA*, January 2025. ISSN 2049-8764. doi: 10.1093/imaiai/iaae035.

Pulkit Gopalani, Samyak Jha, and Anirbit Mukherjee. **Global Convergence of SGD For Logistic Loss on Two Layer Neural Nets**. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/pdf?id=9TqAUYB6tC.

Sayar Karmakar and Anirbit Mukherjee. **Provable training of a ReLU gate with an iterative non-gradient algorithm**. *Neural Networks*, 151:264–275, 2022.

Sayar Karmakar, Anirbit Mukherjee, and Theodore Papamarkou. **Depth-2 neural networks under a data-poisoning attack**. *Neurocomputing*, 532:56–66, 2023. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2023.02.034.

Dibyakanti Kumar, Samyak Jha, and Anirbit Mukherjee. **Langevin Monte-Carlo Provably Learns Depth Two Neural Nets at Any Size and Data**. Technical report, arXiv.org, March 2025.

Arvind Neelakantan, Luke Vilnis, Quoc V Le, Ilya Sutskever, Lukasz Kaiser, Karol Kurach, and James Martens. Adding gradient noise improves learning for very deep networks. *arXiv preprint arXiv:1511.06807*, 2015.

Matteo Tucat, Anirbit Mukherjee, Mingfei Sun, Procheta Sen, and Omar Rivasplata. **Regularized Gradient Clipping Provably Trains Wide and Deep Neural Networks**. *Transactions on Machine Learning Research*, 2025, 2025.