# PhD Proposal: Two-Scale Dynamics of Neural Networks

Raphaël Berthier

February 9, 2026

*This project targets an excellent theory-oriented student. It could start with an internship in the spring of 2026, followed by the PhD starting in the fall of 2026. If interested, feel free to contact me at* `raphael.berthier@inria.fr`.

**Context and Objectives.** Shallow neural networks, specifically single-hidden-layer models, are functions of the form

$$f_{a,u,b}(x) = \sum_{i=1}^{m} a_i \sigma(\langle u_i, x \rangle + b_i),$$

where $x \in \mathbb{R}^d$ and $\sigma : \mathbb{R} \to \mathbb{R}$ is a non-linear activation function, such as ReLU $\sigma(z) = \max(z, 0)$. In statistical learning, the parameters $a, u, b$ are typically trained using (stochastic) gradient descent on a data-fitting loss term.

Existing approaches to study these non-convex dynamics focus on simplified regimes. The Neural Tangent Kernel (NTK) approximation [13] linearizes the dynamics around initialization, but it fails to capture feature selection or explain the practical success of neural networks [9]. Mean-field theory for neural networks [15, 8, 17, 16] applies only in the limit of infinitely many neurons. This proposal explores a novel framework, the *two-scale regime*, to investigate non-linear dynamics with (i) a moderate number of neurons and (ii) feature selection mechanisms.

**Two-Scale Regime.** In this regime, gradient steps for $u$ and $b$ are taken to be infinitesimally smaller than those for $a$, introducing fast-slow dynamics. On the fast timescale, only $a$ evolves significantly while $u$ and $b$ remain fixed, leading to linear regression dynamics that are well-understood. On the slower timescale, $u$ and $b$ adjust while $a$ remains optimally adapted to $u$ and $b$. This separation simplifies the analysis of $u$ and $b$'s slower dynamics by leveraging the fast equilibrium of $a$.

Fast-slow systems have a long history in mathematics and physics (e.g., [3, Chapter 2]). Two-scale algorithms have been used in stochastic approximation and optimization [6, 7, 11, 18, 12], but have not been much applied to neural network analysis. Singular perturbation theory [3], particularly Tikhonov's theorem, allows the decoupling of fast and slow dynamics. In the neural network setting, this decoupling provides a powerful framework to analyze interactions between the two layers.

**Research Directions.**

1. **One-dimensional case:** In dimension $d = 1$, with ReLU activation $\sigma(z) = \max(z, 0)$, shallow neural networks represent piecewise affine functions. The fast dynamics of $a$ yield the best piecewise affine approximation of the target function for a fixed partition, while the slow dynamics of $u$ and $b$ refine the partition itself. This setup connects to classical numerical analysis problems on free-knot spline approximation. This connection between free-knot approximation and neural networks in the two-scale regime remains largely unexplored and will be a focus of this research. First steps were taken in [14].

2. **High-dimensional case ($d \gg 1$):** In the case of "single-index models" $f_*(x) = \varphi_*(\langle u_*, x \rangle)$, the two-scale regime separates the alignment of $u_i$ with $u_*$ (slow dynamics) from

the approximation of the non-linearity $\varphi_*$ by $a_i$ (fast dynamics). This incremental learning phenomenon was demonstrated in [4]. Future work will extend these results to "multi-index models", $f_*(x) = \varphi_*(Px)$, where $P$ projects onto a low-dimensional subspace. The two-scale regime could simplify this setting and offer insights into broader neural network behaviors [2, 1, 10, 5].

3. **Numerical and theoretical extensions:** This research will also address practical extensions to finite learning rates and stochastic gradient descent. While the two-scale regime assumes an infinite timescale separation, quantifying its behavior under finite conditions will help bridge the gap between theoretical insights and real-world applications.

**Summary.** This thesis aims to formalize the two-scale regime for neural networks and demonstrate its utility in analyzing complex dynamics with a moderate number of neurons. By combining tools from numerical analysis, singular perturbation theory, and modern deep learning, this work seeks to contribute new theoretical insights and practical guidelines for neural network training.

# References

[1] Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks. In *Conference on Learning Theory*, pages 4782–4887. PMLR, 2022.

[2] Emmanuel Abbe, Enric Boix-Adsera, Matthew S Brennan, Guy Bresler, and Dheeraj Nagaraj. The staircase property: How hierarchical structure can guide deep learning. *Advances in Neural Information Processing Systems*, 34:26989–27002, 2021.

[3] Nils Berglund and Barbara Gentz. *Noise-induced phenomena in slow-fast dynamical systems: a sample-paths approach*. Springer Science & Business Media, 2006.

[4] Raphaël Berthier, Andrea Montanari, and Kangjie Zhou. Learning time-scales in two-layers neural networks. *Foundations of Computational Mathematics*, pages 1–84, 2024.

[5] Alberto Bietti, Joan Bruna, and Loucas Pillaud-Vivien. On learning gaussian multi-index models with gradient flow. *arXiv preprint arXiv:2310.19793*, 2023.

[6] Vivek S Borkar. Stochastic approximation with two time scales. *Systems & Control Letters*, 29(5):291–294, 1997.

[7] Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

[8] Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

[9] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32, 2019.

[10] Amit Daniely and Eran Malach. Learning parities with neural networks. *Advances in Neural Information Processing Systems*, 33:20356–20365, 2020.

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[12] Mingyi Hong, Hoi-To Wai, Zhaoran Wang, and Zhuoran Yang. A two-timescale stochastic algorithm framework for bilevel optimization: Complexity analysis and application to actor-critic. *SIAM Journal on Optimization*, 33(1):147–180, 2023.

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

[14] Pierre Marion and Raphaël Berthier. Leveraging the two-timescale regime to demonstrate convergence of neural networks. *Advances in Neural Information Processing Systems*, 36, 2024.

[15] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

[16] Grant Rotskoff and Eric Vanden-Eijnden. Trainability and accuracy of neural networks: An interacting particle system approach. *arXiv preprint arXiv:1805.00915*, 2018.

[17] Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: a central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020.

[18] Csaba Szepesvári. *Algorithms for reinforcement learning*. Springer, 2010.