Convergence in learning

Joakim Andén and Kathlén Kohn

Many machine learning models rely on optimization procedures that fit various parameters to a given training dataset. While the convergence behavior of classical methods (such as shrinkage models, decision trees, and support vector machines) is relatively well understood, less is known for deeper models, such as neural networks. A similar problem occurs for models based on low-rank matrix factorization. Both methods involve non-convex optimization and convergence to global minima is not guaranteed. Nonetheless, neural networks and low-rank matrix factorization achieve impressive results in a wide variety of tasks. Understanding the convergence behavior of these and characterizing the resulting models is therefore of great importance. Recent work has started illuminating these questions for specific classes of neural networks [1, 2, 3] and matrix factorization methods [4, 5]. Other work has focused on characterizing the properties of the networks once converged [6, 7, 8, 9]. We plan to extend these results by combining ideas from algebraic geometry and statistical signal processing, while anchoring the theoretical analysis in concrete problems in computer vision and cryogenic electron microscopy (cryo-EM). To this end, we propose a series of projects, each illuminating different aspects of the above problem. We envision that one of these projects, possibly in combination with one or two side projects, would form the core research subject of a doctoral thesis.

Scattering transforms. These transforms were introduced as convolutional networks with fixed weights that guarantee certain invariance and stability properties with respect to translation and deformation [6]. By explicitly encoding these symmetries, networks consisting of scattering transforms followed by a linear layer have achieved significant success in various classification and regression tasks [10, 11, 12, 13]. Since only a single linear layer is optimized, the problem becomes convex and a global optimum is easily found. In addition, the fixed structure of the scattering transform and the relative simplicity of the linear layer simplifies analysis and increases interpretability of the model. Understanding these networks therefore amounts to characterizing the space of linear combinations of scattering transform coefficients [7]. Determining the properties of this space would specify the limitations of these networks and suggest extensions that would increase their expressivity. This project would involve collaborations with the group of S. Mallat (École normale supérieure) and M. Eickenberg (Flatiron Institute) with applications to audio and image classification.

Attractors of autoencoders. An important aspect of autoencoders is their ability to memorize the training data, which has been recently explored from a dynamical systems perspective [8, 9]. Empirical results suggest that training examples form attractors of these autoencoders, but the theoretical reasons behind that mechanism are still not clear. Algebraic techniques can be applied in the setting of ReLU autoencoders with Euclidean loss, as the underlying geometric problem is to find a closest point on a semi-algebraic set. We conjecture that all training examples are attractors in a (global) minimum of the Euclidean loss of a sufficiently deep ReLU autoencoder. This project would investigate that conjecture as well as further conditions under which attractors are formed. Possible collaborators are G. Montúfar (UCLA & Max Planck Institute MiS Leipzig) or C. Uhler (MIT).

Convergence of linear networks. As linear networks are the easiest type of neural networks, many of their properties are well understood, including the structure of their critical points when using the Euclidean loss function [1]. However, it remains an open problem to show that generic initializations of the network converge under gradient flow to a global minimum [2, 3]. Another project is therefore to investigate this conjecture and possibly expand it to other types of networks or loss functions. The project would be conducted in collaboration with A. Eftekhari (Umeå University) or H. Rauhut (RWTH Aachen University).

Convergence in low-rank matrix factorization models. A similar behavior is observed in low-rank matrix models. Recent work has shown that global convergence is possible under certain settings of linear matrix measurement using Gaussian matrices [4, 5]. However, this convergence behavior is observed for much wider classes of measurement operators. In particular, replacing the Gaussian sensing matrices with certain integral operators representing tomographic projection, global convergence is observed in a wide range of configurations. This particular setup has applications to the heterogeneity problem in cryo-EM, where a low-rank factorization model can be used to characterize the structural variability of three-dimensional density maps representing the imaged molecule [14, 15]. This project would investigate this behavior and extend previous convergence results to wider settings of measurement operators. It would be conducted in collaboration with A. Singer (Princeton University).

[1] Matthew Trager, Kathlén Kohn, and Joan Bruna. Pure and spurious critical points: a geometric study of linear networks. In *International Conference on Learning Representations*, 2020.

[2] Bubacarr Bah, Holger Rauhut, Ulrich Terstiege, and Michael Westdickenberg. Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers. *arXiv:1910.05505*, 2019.

[3] Armin Eftekhari. Training linear neural networks: Non-local convergence and complexity results. *arXiv:2002.09852*, 2020.

[4] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proc. STOC*, pages 665–674, 2013.

[5] Dohyung Park, Anastasios Kyrillidis, Constantine Carmanis, and Sujay Sanghavi. Non-square matrix sensing without spurious local minima via the Burer–Monteiro approach. In *Proc. AISTATS*, pages 65–74, 2017.

[6] Stéphane Mallat. Group invariant scattering. *Comm. Pure Appl. Math.*, 65(10):1331–1398, 2012.

[7] Stéphane Mallat. Understanding deep convolutional networks. *Phil. Trans. R. Soc. A*, 374(2065):20150203, 2016.

[8] Adityanarayanan Radhakrishnan, Karren Yang, Mikhail Belkin, and Caroline Uhler. Memorization in overparameterized autoencoders. *arXiv:1810.10333*, 2018.

[9] Adityanarayanan Radhakrishnan, Mikhail Belkin, and Caroline Uhler. Overparameterized neural networks can implement associative memory. *arXiv:1909.12362*, 2019.

[10] Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1872–1886, 2013.

[11] Joakim Andén and Stéphane Mallat. Deep scattering spectrum. *IEEE Trans. Signal Process.*, 62:4114–4128, 2014.

[12] Joakim Andén, Vincent Lostanlen, and Stéphane Mallat. Joint time–frequency scattering. *IEEE Trans. Signal Process.*, 67(14):3704–3718, July 2019.

[13] Michael Eickenberg, Georgios Exarchakis, Matthew Hirn, Stéphane Mallat, and Louis Thiry. Solid harmonic wavelet scattering for predictions of molecule properties. *J. Chem. Phys.*, 148(24):241732, 2018.

[14] Joakim Andén and Amit Singer. Structural variability from noisy tomographic projections. *SIAM J. Imaging Sci.*, 11(2):1441–1492, 2018.

[15] Amit Moscovich, Amit Halevi, Joakim Andén, and Amit Singer. Cryo-EM reconstruction of continuous heterogeneity by Laplacian spectral volumes. *Inverse Prob.*, 36(2):024003, 2020.