Subject: Optimization and generalization of deep neural networks

Dissertation supervisor: George Deligiannidis

Pre-requisite knowledge(listed as essential, recommended, useful)

Essential: Part A probability, SB3a Applied Probability. Strongly recommended: SC10 Algorithmic Foundations of learning

Description of proposal:

Deep neural networks, that is artificial neural networks with many hidden layers, have achieved state-of-the-art performance in various machine learning tasks. However, this remarkable success remains to a large extent mysterious as traditional mathematical approaches fail to capture their performance.

To put things more concretely, let $\mathcal{X} \subset \mathbb{R}^d$, $\mathcal{Y} = \{\pm 1\}$ and consider a classification problem where one is interested in learning a distribution $\mathcal{P}$ on $\mathcal{X} \times \mathcal{Y}$ given a training sample $(\boldsymbol{x_i}, y_i)_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$. In the simplest possible terms then, a neural network can be represented by a non-linear, function $f(\cdot;\cdot) : \mathcal{X} \times \mathcal{W} \mapsto \mathcal{Y}$, where $\mathcal{W}$ is the typically very high-dimensional space where the *weights* live. The weights are essentially the parameters that we learn from the data typically by applying some optimization algorithm, like Stochastic Gradient Descent (SGD), to minimize the following objective

$$\boldsymbol{W}^* := \text{argmin}_{\boldsymbol{w}\in\mathcal{W}} L_{\mathsf{emp}}[f(\cdot;\boldsymbol{w})] := \text{argmin}_{\boldsymbol{w}\in\mathcal{W}} \sum_{i=1} L\left(y_i, f(\boldsymbol{x_i};\boldsymbol{w})\right),$$

where L is some loss function.

The hope then is the model $f^* := f(\cdot;\boldsymbol{W}^*)$ will *generalize well*, in the sense that it will perform well as a classifier on new, unseen examples. In other words we would like to control the *generalization gap*

$$\left|\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{P}} L\Big(y, f^*(\boldsymbol{x})\Big) - L_{\mathsf{emp}}[f^*]\right|.$$

Classical learning theory provides upper bounds for the generalization gap, that hold with high probably, and have the form $\mathcal{C}_{\mathcal{W}} n^{-1/2}$, where the constant $\mathcal{C}_{\mathcal{W}}$ measures the complexity of class of functions captured by our model and tends to grow with the expressiveness, or *capacity*, of the model class, e.g. Vapnik-Chervonenkis (VC) dimension or Rademacher complexity. This becomes especially important in deep neural networks, where the number of parameters tends to be much larger than the size of the data set, which in principle means that we can over-fit, or even memorize, the data. This in turn means that the fitted model should not generalize well at all.

This however is not what happens in practice. Deep neural networks are indeed over-parameterized enough to achieve zero training loss, but contrary to what theory suggests they also generalize very well. This is a crucial issue in deep learning and is being intensely studied over the past few years. Several ideas have appeared that attempt to explain the performance of deep neural networks. Some of them focus on a careful analysis of SGD on the non-convex objective and argue that SGD tends to converge on those local minima that generalize well. A related approach suggests that SGD implicitly solves a regularized problem thus improving generalization. Other approaches attempt to obtain generalization bounds directly on classifiers that interpolate, and thus perfectly fit any data set, without appealing to standard learning theory bounds.

Possible avenues of investigation:
The first aim of the project, common to all students, will be to study the fundamentals of learning theory, give a concise review of some recent developments in deep neural networks clearly explaining why classical theory fails to explain their performance.

After this each student will explore and give an exposition a recent paper. Some overlap is to be expected. Some possibilities are given below:

Recent papers

[1] Mikhail Belkin, Daniel J Hsu, and Partha Mitra. "Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate". In: *Advances in Neural Information Processing Systems*. 2018, pp. 2300–2311.

[2] Alon Brutzkus et al. "Sgd learns over-parameterized networks that provably generalize on linearly separable data". In: *arXiv preprint arXiv:1710.10174* (2017).

[3] Yuanzhi Li and Yingyu Liang. "Learning overparameterized neural networks via stochastic gradient descent on structured data". In: *Advances in Neural Information Processing Systems*. 2018, pp. 8157–8166.

[4] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. "Learning and generalization in overparameterized neural networks, going beyond two layers". In: *arXiv preprint arXiv:1811.04918* (2018).

[5] Simon S Du et al. "Gradient descent provably optimizes over-parameterized neural networks". In: *arXiv preprint arXiv:1810.02054* (2018).

[6] Alon Brutzkus and Amir Globerson. "Why do Larger Models Generalize Better? A Theoretical Perspective via the XOR Problem". In: *International Conference on Machine Learning*. 2019, pp. 822–830.

Useful reading and further references

[7] Chiyuan Zhang et al. "Understanding deep learning requires rethinking generalization". In: *arXiv preprint arXiv:1611.03530* (2016).

[8] Chiyuan Zhang et al. "Identity crisis: Memorization and generalization under extreme overparameterization". In: *arXiv preprint arXiv:1902.04698* (2019).

[9] Suriya Gunasekar et al. "Implicit bias of gradient descent on linear convolutional networks". In: *Advances in Neural Information Processing Systems*. 2018, pp. 9461–9471.

[10] Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

[11] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.