# The Provable Emergence of Neural Collapse

*Marco Mondelli*

## Abstract

Deep neural networks at convergence consistently represent the training data in the last layer via a highly symmetric geometric structure referred to as neural collapse. This empirical evidence has spurred a line of theoretical research aimed at proving the emergence of neural collapse, and the talk will present some progress in this sense.

I will start by focusing on the unconstrained features model where, motivated by the network's perfect expressivity, one treats the last layers' feature vectors as a free variable and explicitly optimizes them together with the last layers' weight matrix. In contrast with existing work restricted to models that are linear or have only two layers, I will show that neural collapse occurs in all the layers of a deep unconstrained features model for binary classification. However, rather surprisingly, as soon as one goes beyond two layers or two classes, the neural collapse solution stops being optimal. Here, the main culprit is a low-rank bias of multi-layer regularization schemes: this bias leads to optimal solutions of even lower rank than the neural collapse.

A limitation of the unconstrained features model is that it is data-agnostic, which puts into question its ability to capture gradient-based training. To address the issue, I will discuss two recent works that prove generic guarantees on neural collapse for networks that end with at least two linear layers and have small training error: neural collapse occurs if either (i) the linear layers are balanced, or (ii) the network operates in the mean-field regime and has small gradient norm. This in turn allows to prove the emergence of neural collapse in the end-to-end gradient descent training of deep neural networks, either in a linearized regime (i.e., where tools from the neural tangent kernel can be used) or in a mean-field regime (i.e., where feature learning takes place).