Gradient descent aligns the layers of deep linear networks

Matus Telgarsky, University of Illinois Urbana-Champaign

This talk establishes risk convergence and asymptotic weight matrix alignment --- a form of implicit regularization --- of gradient flow and gradient descent when applied to deep linear networks on linearly separable data. In more detail, for gradient flow applied to strictly decreasing loss functions (with similar results for gradient descent with particular decreasing step sizes): (i) the risk converges to 0; (ii) the normalized i-th weight matrix asymptotically equals its rank-1 approximation; (iii) these rank-1 matrices are aligned across layers. In the case of the logistic loss (binary cross entropy), more can be said: the linear function induced by the network --- the product of its weight matrices --- converges to the same direction as the maximum margin solution. This last property was identified in prior work, but only under assumptions on gradient descent which here are implied by the alignment phenomenon.