**Blue: could try on my experiment**

**Red: could use in my paper**

**On the Learning Dynamics of Deep Neural Networks**

How does the confidence of a classifier evolve throughout learning?

How does the loss used during training impact its dynamics?

Which properties of the features present in a dataset impact learning, and how?

**Independent mode learning** We show that, similarly to the case of linear networks and under certain initial conditions, learning happens independently between different classes, i.e. classes induce a partition of the network activations, corresponding to orthogonal modes of the data.

**Learning dynamics** We prove that in accordance to experimental findings, the hidden activations and the classification error of the network show a sigmoidal shape with slow learning at the beginning followed by fast saturation of the curve. We also characterize a region in the initialization space where learning is frozen or eventually dies out.

**Hinge loss** We study how using the hinge loss impacts learning and quantitatively compare it to the classic cross-entropy loss. We show that the hinge loss allows one to solve a classification task much faster, by providing strong gradients no matter how close to convergence the neural network is.

**Gradient starvation** Finally, we identify a phenomenon that we call gradient starvation where the most frequent features present in the dataset starve the learning of other very informative but less frequent features. Gradient starvation occurs naturally when training a neural network with gradient descent and might be part of the explanation as to why neural networks generalize so well. They intrinsically implement a variant of Occam’s razor (Ariew, 1976): the simplest explanation is the one they converge to first.