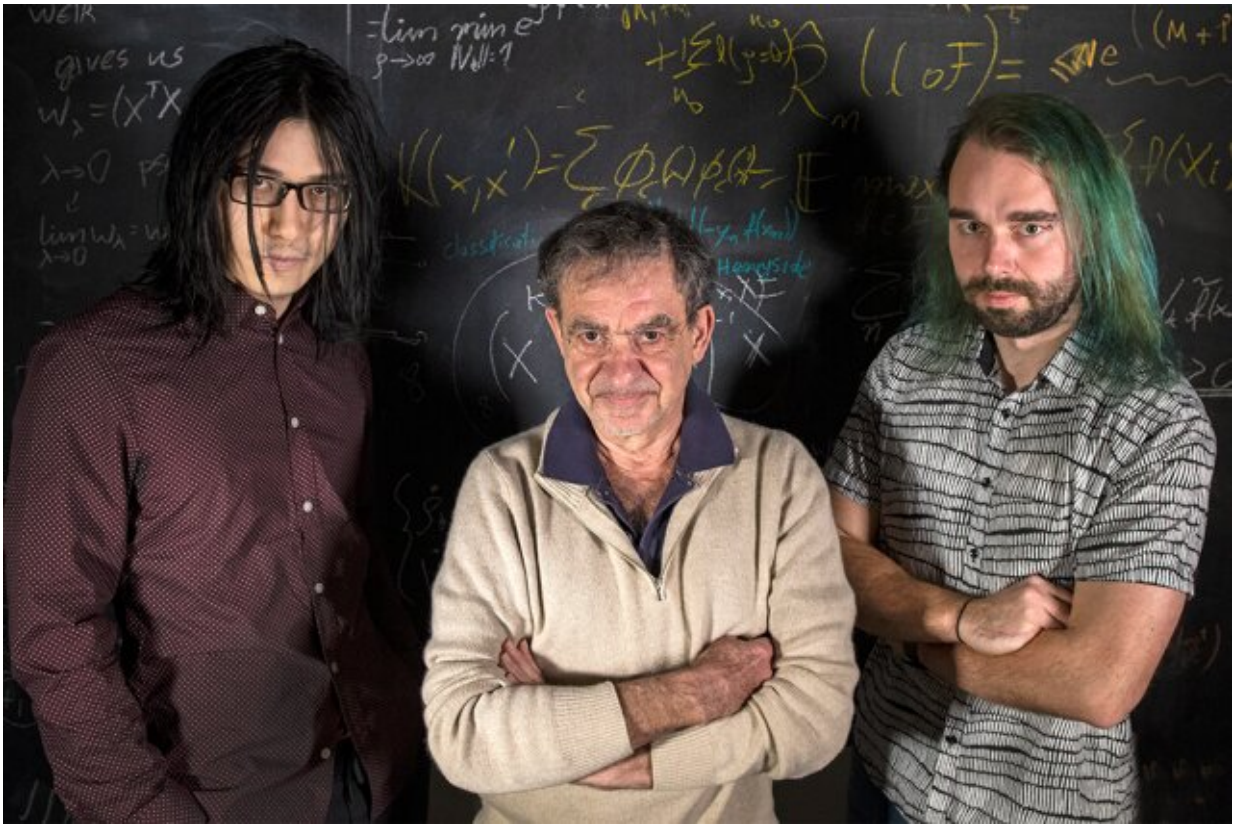
Why deep networks generalize despite going against statistical intuition

March 2 2020, by Kris Brewer



MIT researchers (left to right) Qianli Liao, Tomaso Poggio, and Andrzej Banburski stand with their equations. Credit: Kris Brewer

Introductory statistics courses teach us that, when fitting a model to some data, we should have more data than free parameters to avoid the

danger of overfitting—fitting noisy data too closely, and thereby failing to fit new data. It is surprising, then, that in modern deep learning the practice is to have orders of magnitude more parameters than data. Despite this, deep networks show good predictive performance, and in fact do better the more parameters they have. Why would that be?

It has been known for some time that good performance in machine learning comes from controlling the complexity of networks, which is not just a simple function of the number of free parameters. The complexity of a classifier, such as a neural network, depends on measuring the "size" of the space of functions that this network represents, with multiple technical measures previously suggested: Vapnik–Chervonenkis dimension, covering numbers, or Rademacher complexity, to name a few. Complexity, as measured by these notions, can be controlled during the learning process by imposing a constraint on the norm of the parameters—in short, on how "big" they can get. The surprising fact is that no such explicit constraint seems to be needed in training deep networks. Does deep learning lie outside of the classical learning theory? Do we need to rethink the foundations?

In a new Nature Communications paper, "Complexity Control by Gradient Descent in Deep Networks," a team from the Center for Brains, Minds, and Machines led by Director Tomaso Poggio, the Eugene McDermott Professor in the MIT Department of Brain and Cognitive Sciences, has shed some light on this puzzle by addressing the most practical and successful applications of modern deep learning: classification problems.

"For classification problems, we observe that in fact the parameters of the model do not seem to converge, but rather grow in size indefinitely during gradient descent. However, in classification problems only the normalized parameters matter—i.e., the direction they define, not their size," says co-author and MIT Ph.D. candidate Qianli Liao. "The not-so-

obvious thing we showed is that the commonly used gradient descent on the unnormalized parameters induces the desired complexity control on the normalized ones."

"We have known for some time in the case of regression for shallow linear networks, such as kernel machines, that iterations of gradient descent provide an implicit, vanishing regularization effect," Poggio says. "In fact, in this simple case we probably know that we get the best-behaving maximum-margin, minimum-norm solution. The question we asked ourselves, then, was: Can something similar happen for deep networks?"

The researchers found that it does. As co-author and MIT postdoc Andrzej Banburski explains, "Understanding convergence in deep networks shows that there are clear directions for improving our algorithms. In fact, we have already seen hints that controlling the rate at which these unnormalized parameters diverge allows us to find better performing solutions and find them faster."

What does this mean for machine learning? There is no magic behind deep networks. The same theory behind all linear models is at play here as well. This work suggests ways to improve deep networks, making them more accurate and faster to train.

More information: Tomaso Poggio et al. Complexity control by gradient descent in deep networks, *Nature Communications* (2020). DOI: 10.1038/s41467-020-14663-9