The Two Phases of Gradient Descent in Deep Learning

Thanks to great experimental work by several research groups studying the behavior of Stochastic Gradient Descent (SGD), we are collectively gaining a much clearer understanding as to what happens in the neighborhood of training convergence.

The story begins with the best paper award winner for ICLR 2017, “Rethinking Generalization”. This paper I first discussed several months ago in a blog post “Rethinking Generalization in Deep Learning”. One interesting observation in that paper is the role of SGD. The observation is extremely radical, where the authors write:

\[
\text{Indeed, in neural networks, we almost always choose our model as the output of running stochastic gradient descent. Appealing to linear models, we analyze how SGD acts as an implicit regularizer. For linear models, SGD always converges to a solution with small norm. Hence, the algorithm itself is implicitly regularizing the solution.}
\]

This is a very odd notion that SGD is being labeled as an ‘implicit regularization’. Coincidentally, another paper: An Empirical Analysis of Deep Network Loss Surfaces by Daniel Jiwoong Im, Michael Tao, Kristin Branson, discusses the structure of loss surfaces of different SGD algorithms and discovers that they all are different:
These experimental measurements seem to back the claim that similar to regularization, the SGD algorithm you select will influence where a network converges to. In short, you different resting placing with different SGD algorithms. This is different to how we conventionally think about SGD. That is, different SGDs just give you differing convergence rates due to different strategies, but we do expect that they all end up at the same results! (BTW, I mention that this fantastic paper was rejected in ICLR 2017. Writing academic papers in the deep learning space is unreasonable competitive.)

Leslie Smith and Nicholay Topin, recently submitted a workshop paper to the ICLR 2017 workshop: “Exploring Loss Function Topology with Cyclic Learning Rate” where they discover some peculiar convergence behavior:
Here, as you monotonically increase and decrease the learning rate, there is a transition near at the convergence regime that a large enough learning rate perturbs the system right off is basin into a space of much higher loss. Then the SGD again quickly converges (also note the faster convergence rate). What exactly is happening here?

A recent paper on Arxiv “Opening the Black Box of Deep Neural Networks via Information” by Ravid Shwartz-Ziv and Naftali Tishby has a elegant interpretation of what goes on in SGD. They describe SGD as having two distinct phases, a drift phase and a diffusion phase. SGD begins in the first phase, basically exploring the multidimensional space of solutions. When it begins converging, it arrives at the diffusion phase where it is extremely chaotic and the convergence rate slows to a crawl. This graph best illustrates this behavior:
That is, the behavior makes a phase transition from high mean with low variance to one with a low mean but high variance. This provides further explanation to Smith et. al's observations, that in the region near convergence, it is highly chaotic. This of course does not fully explain why a high learning rate will knock the system into a place of high loss.

Tomaso Poggio and Qianli Liao have however their own experiments and have a theory: “Theory II: Landscape of the Empirical Risk in Deep Learning”. Where they describe in detail the behavior in that region:
It turns the basin of global minima is flat but has very bumpy. Not only that, but there are many of these basins. They invoke some esoteric math theorem and come up with this conclusion:

We can then invoke Bezout theorem to conclude that there are a very large number of zero-error minima, and that the zero-error minima are highly degenerate, whereas the local non-zero minima, if they exist, may not be degenerate. In the case of classification, zero error implies the existence of a margin, that is a flat region in all dimensions around zero error.

Absolutely fascinating paper, worthy of several reads. There is however one pragmatic take away from this paper “Averaging two models within a basin tend to give a error that is the average of the two models (or less). Averaging two models between basins tend to give an error that is higher than both models”.

In summary, there are a lot of research groups that make a good effort at trying to better understand the behavior of Deep Learning systems. It is through this fundamental research work that we all
collectively gain better ways to improve our own work. Conferences unfortunately have the bias towards valuing novel architectures (the crazier the better) over good experimental data. Unfortunately, that favors the practice of alchemy rather than the pursuit of the science of chemistry.

Note: None of the papers mentioned above (other than ‘rethinking generalization’) has successfully made it past a review board.