Hi everyone.
I got a bad cold prior to the start of the conference and have since lost my voice completely.
So please forgive me for typing out my comments for this talk.

In this talk I’d like to start with a wide view of the philosophy of science, and then get progressively more technical and present some new results as the talk goes on.
Why are we here?

While some progress has been made recently towards a foundational understanding of deep learning, most theory work has been disjointed, and a coherent picture has yet to emerge. Indeed, the current state of deep learning theory is like the fable “The Blind Men and the Elephant”

So to start, I’d like to point out this passage in this workshop’s abstract. The prompt suggests that while there has been some progress in developing a theoretical understanding of deep learning, “most theory work has been disjointed”.
This raises a more general question of what, in general, is the mechanism to align competing theories? In this talk, I’ll argue for a prominent role of experiment in shaping the way we develop and evaluate theories. Essentially, I’m advocating a mode of knowledge creation that more closely resembles the natural sciences. But what sort of science is machine learning? While many of us have the title “research scientist” or “applied scientist” and many of us run “experiments”, to me it always seems unnatural to say “I am a scientist”. So perhaps to start, we can first ask “what is the nature of science?”, and then take a look at how that squares against the modern theory and practice of machine learning.
Recently, I began reading Thomas Kuhn’s “The Structure of Scientific Revolutions.” This is the work in the philosophy of science that introduced the concepts of “paradigm shift” and “normal science.” Being familiar with these terms in the more colloquial sense, I was thinking about the work we do in deep learning and how repetitive and familiar some of it has become. The angle I was looking for was a narrative was that perhaps through Kuhn’s framework we might reasonably describe deep learning as constituting a paradigm shift in 2012 that has since become “normal science”.
Upon reading the book, however it seems that the picture is a bit more complicated. Kuhn describes three main situations. There is the “pre-paradigmatic”, before a compelling theory has taken hold. At this time, there is not consensus about precisely what sorts of entities the world contains or implicitly, which it does not. Here, a field is in an immature stage and much research consists of trying random things. *This might sound familiar.*
On the other hand upon acquisition of a paradigm, a field lurches towards normal science. Here a rigid set of assumptions govern research. Normal science is not a bad thing. It’s the standard state of affairs for most mature sciences. And is often when some of the deepest technical work is done. However, when it exhausts the useful things to do, normal science tends towards “puzzle solving”. The rigid formalism presents many opportunities for fleshing out ever more minute details, but these puzzles are not always undertaken for their actual utility. Instead, often the chief purpose of puzzle solving is to demonstrate the prowess of the solver for solving puzzles. Throughout this time, tightly linked theoretical refinement and experimental science are undertaken, with the goal of bringing theory and nature into tighter alignment.
The final state of affairs is paradigm shift.
Over time normal science encounters more and more problems that it cannot solve.
As an example here, perhaps we might consider the failure of theories based on VC-dimension or Rademacher complexity
to explain the ability of deep neural networks to generalize so well to unseen data.
Paradigm shifts are the “tradition-shattering complement to the tradition-bound activities of normal science”
So where does machine learning fit into this picture?
At this point I’ll paint a little bit of a caricature and hope that you’ll grant me some license.

Machine learning
On one hand we have machine learning theory. Here skilled researchers work with a set of assumptions to produce theorems. The mode of knowledge creation resembles pure mathematics. Quality of work is assessed based on the beauty of the ideas and results, but insufficient attention is paid to justifying/experimentally validating assumptions to bring them closer into alignment with nature. In some ways the state of learning theory resembles “normal science”. The field is rigorous and mature, but perhaps a bit too consumed with puzzle solving, where the utility of solving the puzzle isn’t always sufficiently clear.
On the other hand we have the modern practice of deep learning. Here we talk a lot about “experiments”, but these are seldom scientific experiments. Most often, their purpose is to demonstrate utility, not to produce insights. Moreover, applied deep learning exhibits some aspects of the pre-paradigmatic: namely, much of our activity consists of “trying random things”, or, being the “S” in stochastic graduate-student descent.
So what I’d like to argue for is a more wide-spread practice of machine learning science, engaging theory and experiment in a tighter loop. Unfortunately, “science-y” papers are currently a difficult genre of paper to publish. I’d argue that we need to change our objective function to encourage more of this sort of work. In the next part of the talk I’d describe a series of papers (by other researchers) leading up to some recent developments in the theory of deep learning that I think exhibit precisely this tight loop between theory and experiment. And then I’ll describe some new surprising experimental results inspired by this theory that perhaps confirm the usefulness of this theory.
So to start. We have mature paradigms for explaining the inductive bias of machine learning models based on VC dimension and Rademacher complexity. Recently, Chiyuan Zhang and collaborators showed that typical deep learning models, with datasets of typical size encountered in practice are actually capable of fitting arbitrary labelings, suggesting that no straightforward theory based on VC-dimension or Rademacher complexity can fully account for the ability of neural networks to generalize well.

This result also shook the conventional wisdom in the deep learning community that architectures themselves are a source of inductive bias. If for common architecture, there exist settings of the weights for which the models can classify all training points correctly and all test points incorrectly, then something more is required to explain the ability of deep neural networks to generalize.

In a recent work, inspired by these experiments, Daniel Soudry and collaborators looked to see what inductive bias if anything could be accounted for by the choice of the optimization algorithm. To start, they began analyzing linear neural networks on separable data. It makes sense to analyze separable data because for all modern deep neural networks, all common datasets are separable. However, this convergence is slow. Interestingly, they showed that on separable data, with cross-entropy loss, linear networks optimized by gradient descent converge to the max margin solution.
One nice aspect of this theory is that it explains a commonly observed phenomena. Over the course of many epochs of training, the training error goes to 0. Long after the training error hits zero, although the validation loss bottoms out and starts to increase (due to more confident mistakes on unseen examples) the validation error continues to decrease.
What about weighted ERM?

- SGD on linear nets w linearly separable data insensitive to weights
- What if this holds on nonlinear nets, for which all data is separable?
- Weighted ERM is a key component in domain adaptation, causal inference, off-policy RL, etc.
- In new draft with student Jonathan Byrd, we ask, “what if anything, is the effect of the “IW” in IW-ERM?”

So to complete another cycle from theory to experiment to theory to experiment: Seeing these results, I wondered if neural networks converge to something resembling a margin-based solution, then if these results hold, what if anything is the effect of weighted risk minimization? For linear nets, the theory dictates that there should be no effect. This is a serious concern because importance weighting is often undertaken with deep neural networks, as a basic step in domain adaptation algorithms, causal inference, and off-policy evaluation, for example. Although the existing theory only covers linear nets (both 1 and multi-layer), and convolutional linear nets, we used experiment to see what happens in the nonlinear case.
Indeed, although importance weights effect the solution early in training, the models eventually converge to a solution that is weight-insensitive.
Notably, adding L2 regularization does restore the influence of importance weighting. But this seems problematic because regularization appears to be the wrong abstraction for modulating the impact of importance sampling. Why should we fiddle with the regularizer to recover the desired effect of importance weighting?
Moreover dropout regularization, which is often thought of as interchangeable with techniques like l2 regularization, does not have this effect of restoring some impact of importance sampling on the learned solution.
We also looked at class weighting with the CIFAR dataset. Here we trained on DOGS and CATS only, with various weightings of the classes. Presumably severely upweighting dogs might cause the classifier to converge to a solution that classifies more holdout images as DOGS. We also might like to think that the classifier would classify more “off-manifold” images as DOGS. Again, we find that while the impact of importance weighting is significant early in training, we converge within 500 epochs, for importance weights up to 256, to a weight-insensitive solution.
Here we look at convergence across all values of the importance weights.
In short, I’d like to argue that we do a bit more “science” in machine learning. This means moving the objective function for both theory and experiment, valuing theories more for accurately predicting things that we didn’t expect to find in the natural world, and placing a greater emphasis on experiments whose purpose is to confer insight.