**Reading Notes: Henry W. Lin, Max Tegmark, and David Rolnick (2017). Why Does Deep and Cheap Learning Work So Well?, J Stat Phys (2017) 168:1223–1247**

**Uwe Ehret**

**Summary**

In this paper, the authors explain why artificial neural networks (ANNs) generally work well for many applications from physics, and how this is not only due to the mathematical properties of ANNs, but also due to the nature of problems in physics. While doing so, they also link problem settings from physics to Machine Learning (ML) settings.

The authors start by stating typical tasks in ML and physics, i.e. unsupervised learning (estimating patterns in or the pdf of some data), classification (estimating the pdf of a discrete label given some data), prediction (estimating the pdf of a label given time-ordered data), and generation (approximating the pdf of some data given a label). With these tasks in mind, the authors analyze ANNs in terms of their expressibility (which classes of functions can they express) and efficiency (how many resources such as neurons are required to do so). In particular, the authors raise the question how relatively small ANNs can approximate functions from much higher-dimensional problem spaces, i.e. how ANNs can be so "cheap".

To provide the grounds, the authors first establish useful links between key terms in physics, ML and information theory, which helps re-interpreting many established practices in ML as applications of insights from physics about the general nature of physical systems. E.g. a system's Hamiltonian H in physics, indicating a system's energy state given known constraints is equivalent to the conditional entropy of the system in information-theoretic terms, or the notion of 'surprisal' in ML.

Next, the authors explain why small and shallow ANNs work well for many problems, even if they only contain a very small set of operations (addition, multiplication, and some nonlinear transformation): They have been shown to be universal function approximators, but without a guarantee that arbitrarily close approximation can be achieved by feasible-sized ANNs. However, the authors state that many data-relations in both physics and ML are of polynomial form, and that such polynomials can be approximated with arbitrary precision by small, shallow ANNs. Further, these polynomials are often very simple – sparse, symmetric and low-order – which facilitates expression by small ANNs. Reasons for this simplicity are among others i) hierarchical system organization, concentrating energy (variability) on a few key levels, ii) the simplifying/averaging effect of interactions in systems of many particles (the Central Limit Theorem at work), iii) system invariance under certain transformations such as translation or rotation, and iv) locality, i.e. that in physical systems the degree to which one subsystem affects another is a fast-declining function of their distance. Also, the authors remark that often a low-order approximation of a higher-order polynomial sufficiently captures the essential workings of the system under consideration.

In the next section, the authors focus on the question why deep (many-layered) ANNs are often preferable over shallow ones. The main reason is that most of the physical world is hierarchically and compositionally structured, spatially and causally. This implies that complex patterns often arise from distinct sequences of simpler steps, which can be modelled as a (causal) Markov chain, which can easily be realized on the layers of a deep ANN. Decomposing of a multilevel generative process into a sequential hierarchy of simpler steps in an ANN aligns therefore with the real-world system, and dramatically reduces the number of required ANN parameters (compared to modeling the entire systems as a single entity by a single ANN). As most ML applications are concerned with some aspect of the natural world, this also holds for most of them. On a side note, the authors mention that the time evolution of physical system can also be modeled as a Markov chain.

The authors then go on by introducing the concept of a sufficient statistic (a data transformation fully preserving the information content about some target variable of interest) and a minimal sufficient statistic (one which is a sufficient statistic for all other sufficient statistics), and that the information processing inequality, when the '=' holds describes exactly this. If now a hierarchical generative process (e.g. to produce an image of a cat through a sequence of steps), implemented as a compositional Markov chain, is reversed (i.e. classifying an image as cat or dog), then the authors show that it can be "optimally reversed one step at a time; there are functions that optimally undo each of the steps, distilling out all information about the level above that was not destroyed by the Markov process". This essentially means that the structure of the inference problem should reflect the (compositional) structure of the generative process, i.e. the ANN should also take a compositional (deep) form. The authors conclude this section by stating that often, a strictly sufficient statistic cannot be identified, but an (computationally efficient) approximation will do, and that renormalization does just that. Renormalization is a set of procedures that extract key features from noisy data sets, a precondition for supervised learning.

The rest of the paper discusses the question whether deep ANNs can be more efficiently (lesser neurons and/or synapses) represented by shallow ANNs. This is ongoing research, but non-flattening theorems (stating cases where mapping from deep to shallow ANNs comes with efficiency losses) exist e.g. for the class of linear ANNs and ANNs representing polynomials.

**How this might be useful**

* ANNs are provably effective and efficient tools for modeling a large class of systems of the physical world, and are therefore a good general basis for modeling in the earth sciences.
* The world is a hierarchical assemblage of subsystems – we should make use of this organization by hierarchical assemblages of ANN structures (not one big ANN for the entire system), which increases model realism, and greatly reduces model size.

**Open questions**

* Do the author's insights also hold for systems where strongly nonlinear feedback and emergent pattern-formation occurs, if one wishes to explicitly model these effects, and not just the macroscopic effect? Can these also be modeled by relatively small ANNs?
* Systems with long memory, and where the feature importance is not strictly and quickly declining with time distance (e.g. language modeling) do not lend themselves well to modeling as Markov chains
* For classification (Fig. 3, Corollary 2), if the exact compositional form of the generative process is not known, how do we determine how much irrelevant information should be discarded in every step of the reversed process (every layer of a deep neural network)?