Theory *is* practically relevant!

or "Simulations are not scalable but theory is scalable"

I just watched "Theoretical Statistics is the Theory of Applied Statistics: How to Think About What We Do" [1] by Prof. Andrew Gelman, from last year's New York R conference. I have to admit he covers quite a lot, some of which I feel I do not have the experience, yet, for appreciating it thoroughly enough. But when Prof. Gelman discusses the value of theory in applied fields (like statistics), it really resonated with my previous research experiences in statistical physics and on the interplay between randomised perfect sampling algorithms and Markov Chain mixing as well as my current perspective on the status quo of deep learning.

So essentially in this post I give more evidence for Prof. Gelman's statements "*simulations are not scalable but theory is scalable*" and "*theory scales*" from different disciplines. Of course you should watch the entire video to get the whole context and learn much more from Prof. Gelman, but I guess one should get the point of the following without doing so.

The theory of finite size scaling in statistical physics: I devoted quite a significant amount of my PhD and post-doc research to finite size scaling, where I applied and checked the theory of finite size scaling for critical phenomena. In a nutshell, the theory of finite size scaling allows us to study the behaviour and infer properties of physical systems in thermodynamic limits (close to phase transitions) through simulating (sequences) of finite model systems. This is required, since our current computational methods are far from being, and probably will never be, able to simulate real physical systems. So that completely agrees with Prof. Gelman's remark.

In a recent publication my coauthors and I explain apparent inconsistencies in the theory of finite size scaling for specific statistical physics models in high dimensions. The paper is called "Geometric explanation of anomalous finite-size scaling in high dimensions" [2].

Central limit theorems: Prof. Gelman mentions central limit theorems in the talk as an example. So here comes another example from my research; We studied limit theorems for extreme value problems and applied those to the study of the runtime of a perfect sampling

algorithm, which has a random runtime. Interesting and surprising outcomes of this line of research are deep and very practical connections between relaxation times of a related family of Markov chains and the standard deviation of the perfect sampling algorithm's runtime. Importantly, the theory of extreme values with its universal limit laws (in particular the Fisher–Tippett–Gnedenko theorem) has been essential for us to come to our conclusions. See "On the Coupling Time of the Heat-Bath Process for the Fortuin–Kasteleyn Random–Cluster Model" [3].

Deep Learning: Here comes a question I have been thinking about for a while now (including discussions with my former PhD advisor Martin Weigel). As of today, I think there is no real answer to my question or the theories I know of are unsatisfactory for the question:

Essentially I am asking for a finite size scaling theory of deep learning. In particular, is there a (universal) theory that can quantify how deep learning models behave on larger problem instances, based on results from sequences of smaller problem instances.

As an example, how do we have to adapt a, say, convolutional neural network architecture and its hyperparameters to sequences of larger (unexplored) problem instances (e.g. increasing the resolution of colour fundus images for the diagnosis of diabetic retinopathy, see "Convolutional Neural Networks for Diabetic Retinopathy" [4]) in order to *guarantee a fixed precision over the whole sequence of problem instances without the need of ad-hoc and manual adjustments to the architecture and hyperparameters for each new problem instance.*

A very early approach of a finite size scaling analysis of neural networks (admittedly for a rather simple "architecture") can be found here [5]. An analogue to this, which just crossed my mind, is the study of Markov chain mixing times, where people typically try to understand the dependence of mixing time scales in Markov chains as problem instances become larger [6].

To cut a long story short, people please appreciate theory in applied settings and do not dismiss theory as practically irrelevant, especially if you think big data!

[1] https://m.youtube.com/watch?time_continue=25&v=cuE9eHSbjNI

[2] https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.118.115701

[3] https://link.springer.com/article/10.1007/s10955-017-1912-x

[4] https://www.sciencedirect.com/science/article/pii/S1877050916311929

[5] https://arxiv.org/abs/cond-mat/9611027

[6] http://pages.uoregon.edu/dlevin/MARKOV/markovmixing.pdf