Abstract:

The tsunami of data and the increasing availability of massive computing power create many opportunities for AI in general, and for data science in particular. However, in many instances, the results delivered by machine learning algorithms are difficult to 'understand' or 'explain'.
Black boxes as they are, we basically do not know when and how they are successful, and when and why not. The whole field is pervaded, if not swamped, by hundreds of 'heuristic tricks', such as tampering with the neuronal activation function (sigmoid, RELU (rectified linear unit),…), experimenting with the number of neurons and hidden layers, selecting optimization methods (numerical ones, such as back-propagation or Gauss-Newton, or randomized ones, such as genetic algorithms), specifying user choices (initial guesses, step sizes, convergence tests, early stopping and/or regularization measures, seeds and cross-over/mutation rates, ...), etc. Therefore, machine learning is still more of an art than a science, and many results published in the literature do not even satisfy the minimal scientific requirement of 'reproducibility'.

Admittedly, this situation is not new. Even in the ('classical') field of system identification or statistical modeling of linear time-invariant finite dimensional dynamical linear systems, the origins of which predate those of deep learning by many years, heuristics prevail. Of course, there is a certain systematic methodology to tackle data-based identification problems: 1. Collect data; 2. Select a pre-specified model class that is parametrized by unknown parameters; 3.
Choose an appropriate approximation criterion; 4. 'Solve' the resulting nonlinear optimization problem by numerical algorithms that output `optimal' parameters; 5. Validate the resulting model on validation/test sets or by assessing the quality of predictions. 6. Re-iterate the whole loop when necessary.

It is in Step 4 that still plenty of heuristics are required. Typically, the optimization problem is nonlinear and provably non-convex. The nonlinearities originate in the assumptions that are made on the data model (e.g. additive misfit on the observations, or unobserved additional inputs), and the fact that products between unknowns occur (e.g. between unknown model parameters and unobserved inputs as in ARMAX models, or between unknown model matrices and states in subspace identification methods). The nonlinearities can be dealt with depending on the identification framework one deploys: by invoking asymptotic orthogonality and numerical linear algebra tools in subspace identification, by using the machinery of instrumental variables in errors-in-variables approaches, or by implementing iterative nonlinear least squares algorithms like in Prediction-Error-Methods. Despite hundreds of person-years of experience, and thousands of published papers, books and software suites, all of these 'solutions' contain plenty of heuristics.

The main message of this talk is that we have to be braver: There is a lot of old and new mathematics out there that we collectively ignore, but that could provide us with a deeper understanding of our identification approaches. Specifically, for the problem of identifying LTI models from observed data, the central observation is that all nonlinearities involved are multivariate polynomial. With least squares as an approximation criterion, we end up with a multivariate polynomial optimization problem. From algebraic geometry, we know that such problem is fundamentally equivalent to a large-scale eigenvalue problem. We will show how the set of first order optimality conditions comprises a multi-parameter eigenvalue problem (MEVP), the solution of which is surprisingly little studied in the literature, requiring 'new'
mathematics: Such a MEVP can be solved by recursively building up a quasi-block-Toeplitz block Macaulay matrix, the null space of which is block multi-shift invariant (a property studied in operator theory). By then applying multi-dimensional (exact) realization theory in that null space (new multi-dimensional system theory), we can find the optimal parameters from the eigenvectors and -values of a specific large-scale matrix.

This ‘solution’ as described, satisfies the very scientific definition of the word, because a set of a priori minimal assumptions on data and model leads to a sequence of mathematically verifiable and reconstructable steps that uniquely characterize the optimal solution to be an eigenvalue problem. Even if heuristics would still be needed to compute the optimal solution because of the mere size of the problem, we now know that we are looking for a minimizing eigenvalue-eigenvector pair of a matrix constructed from the data and chosen model class.

Can we learn something about the challenges posed by deep learning from this very special case of linear system identification? The answer is a resounding ‘yes’, and we will provide some first indications why our mathematical framework as outlined above might also be applicable to neural networks. The take home message is that more - old and new - mathematics is mandatory to understand and explain deep learning. The endeavor is difficult, challenging, time consuming, requires patience and endurance, and the research involved is definitely high-risk high-gain. The journey will take us into the realms of mathematical disciplines such as one- and multidimensional system theory, algebraic geometry, operator theory and numerical linear algebra and optimization. For sure, diving deeper in the mathematical depts of neural networks is largely a-typical in the current massive wave of heuristic deep learning activities. t will require a creative understanding of mathematics, that is more profound than the acquired know-how that most practitioners of AI and deep learning currently possess. Yet, there is no alternative if we want to turn the toolbox of deep learning heuristics into real science.