Dominique Ostermayer | 26.10.2023

Escaping the Curse of Dimensionality

- **The role of machine learning**. In many applications of machine learning, such as image analysis, the ML model is used to approximate an unknown function mapping input to output. The usual reason why such a function is unknown is its complexity. In contrast, in many problems in chemistry such a function is known — e.g., the Schrödinger equation for electronic structure problems. The role of machine learning is then to bypass the impossible or resource-intensive evaluation of such a function .
- **Data availability**. Since the availability of storage increases faster than the availability of processing power, most applications of machine learning are limited not by data availability, but by processing power (*i.e.*, algorithmic efficiency, data selection, and pre-processing). However, machine learning in chemistry is often limited by data availability.
- **Curse of dimensionality**. The curse of dimensionality describes the problem of sampling a high-dimensional space. Because of the exponential increase of volume with an increase in dimensions, the number of data points to cover the space with the same density of data points as before also increases exponentially (in the worst case, which is the usual case).
- **Bias-variance dilemma**. In statistical models, flexibility in the form of a high number of parameters introduces variance, whereas limiting flexibility by expert knowledge in the form of assumptions about parameter space and the underlying probability distribution of the data introduces bias. Usually, it is not possible to simultaneously minimise both the bias and the variance errors. For neural networks (NNs), it was long believed that bias decreases and variance increases with the number of hidden layers. However, recent results suggest that overparameterisation of NNs does not necessarily increase variance.

Chem. Rev. **2021**, *121*, 16, 10142.
Angew. Chem. Int. Ed. **2023**, 62, e202219070 .
CHIMIA **2023**, *77*, 22.

Overcoming Data Scarcity

- **Transfer learning**. The idea of transfer learning is to train a model on one data set and transfer the model to another data set without completely retraining the model. In chemistry, a major application of transfer learning is the initialisation of a model on relatively cheap, thus expansive, semiempirical or DFT datasets and then finetune it on very few data points obtained with an expensive method (e.g., coupled cluster) to improve the quality of machine learning prediction to that of an expensive *ab initio* method.
- **Active learning**. Active learning is in the long tradition of adaptive methods in numerics. The idea is to generate the data of the most information content for the model, e.g., by performing experiments or calculations for that part of the space that is not yet well covered.
- **Inductive bias**. This is what is most commonly meant when chemists speak of expert knowledge and chemical intuition. More generally, the need for data is reduced by limiting the number of parameters of the model or by projecting the data to a lower-dimensional space through assumptions about the behaviour of chemical systems. This is usually achieved by descriptors or model choice.
- **Auto-differentiable models/differentiable programming**. The training of NNs relies on auto-differentiation — the process of evaluating partial derivatives automatically by reducing the derivative to elementary arithmetic operations and some elementary functions by the chain rule. Using auto-differentiable physical models has several advantages: (1) itmakes variational problems of the kind that arise from differential equations easier to minimise; (2) it facilitates the creation of hybrid models from a combination of physical models with NNs; (3) it provides easier access to properties that derive from higher-order derivatives.