NP FS Connections to AI/ML

Tanmoy Bhattacharya, LANL

Dec 14, 2022

LA-UR-22-32894

AI/ML: the next frontier

The scientific method has progressed through various revolutions:

1. Observations, classification and criticism
2. Repeatable and controlled experiments
3. Positivism and Quantitative Methods
4. Rigorous statistical analysis
5. Simulations
6. AI/ML

Huge opportunity if used well!

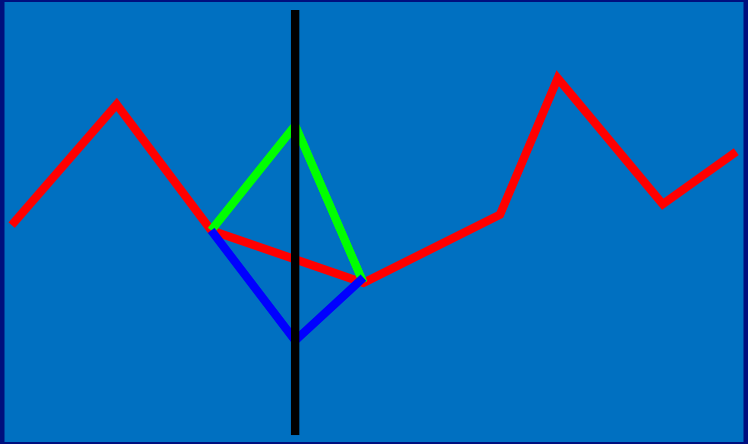
At each stage need to connect to previous levels.

Similarities and differences with other methods

1. Like simulations, AI/ML methods are black boxes: answers can’t be checked in reasonable time by human brains.
2. Unlike simulations, no underlying ‘theory’: it is more like clustering and interpolation. Output is ‘like’ that of similar inputs.
3. Interpolation and extrapolation very similar in high dimensions: almost all data on the boundary, almost all movement in a ‘new’ direction.
4. Validity can be checked statistically. The ‘complexity’ of the interpolating function space driven by data availability.
5. Power of AI/ML increases with increasing data!

Impossibility of interpolation

1. No free lunch theorem: As many functions through every unseen point as any other.
2. Theory provides a prior, i.e., function class of solutions to interpolate, e.g., smoothness.
3. In high dimensions, need exponentially large number of data points to 'cover' each dimension.
4. Assume data a low-dimensional pancake.
5. Splines and ML work by making implicit assumptions and testing on leave out.



Examples of use

1. Surrogate models: Interpolate between known points.
2. Speed up calculations.
3. Investigate theory space: interpolate between theories.
4. Classify events into known types.
5. Compare distributions.
6. Generate distributions.
7. Find natural ‘categories’ of events.
8. Detect anomalies.
9. Suggest experiments!

A zoo of methods

AI/ML techniques finds statistical patterns in the input data and reconstruct the the manifold they inhabit; they can then carry out interesting tasks:

1. Clustering: find clusters of similar data
2. Classification: predict the class to which unseen elements belong
3. Regression: predict dependent variable on new instances
4. Generation of new data indistinguishable from already seen data
5. Natural parameterizaton: discover latent space

Uncertainty quantification

Science needs accuracy bounds.

1. Bayesian methods: fully Bayesian methods often difficult since model space is unclear and priors difficult to specify.
2. Generally speaking uncertainty can be gleaned from:
   1. Neighborhood in input space: high-dimensional space almost always sparse.
   2. Neighborhood in output space: uncertainty independent of output.
   3. Neighborhood in model space: Bayesian priors on opaque model parameters.
   4. Supervised learning of uncertainty as another task: needs more data.

Robust Use

Alternative to uncertainty quantification: use ML to provide guesses for

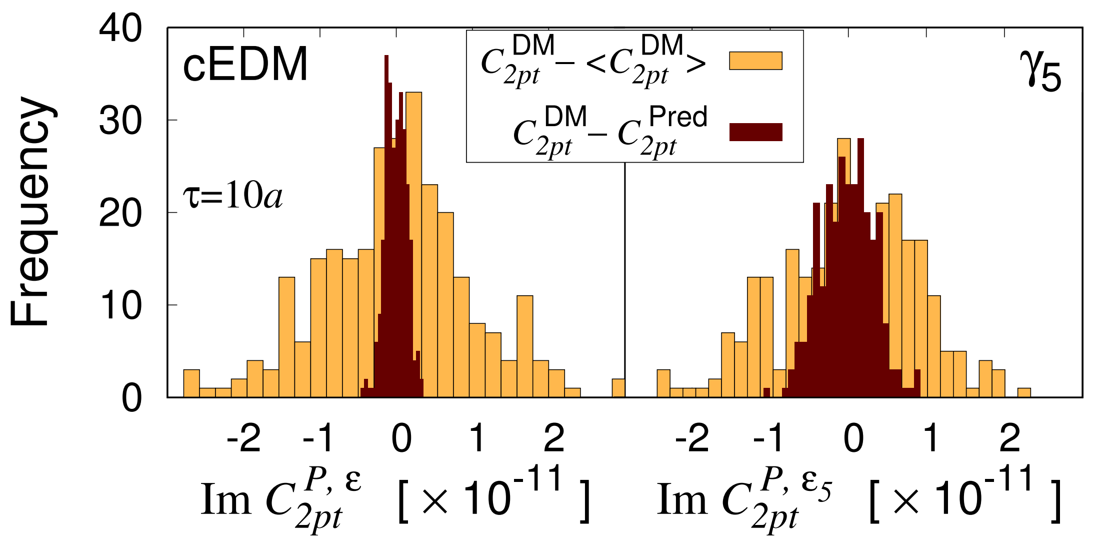
1. Variational methods: Answer is the minimum of the energy functional; ML provides alternatives to try.
2. Correlated observables: Statistical bias subtraction, ML provides observables with high correlation.
3. Variance reduction: ML provides clusters with small ingroup variance.
4. Contour deformation: ML provides contours that reduce variance.
5. Monte Carlo: ML provides unbiased samples for metropolis.
6. Autoencoders: ML provides an efficient encoding of data.

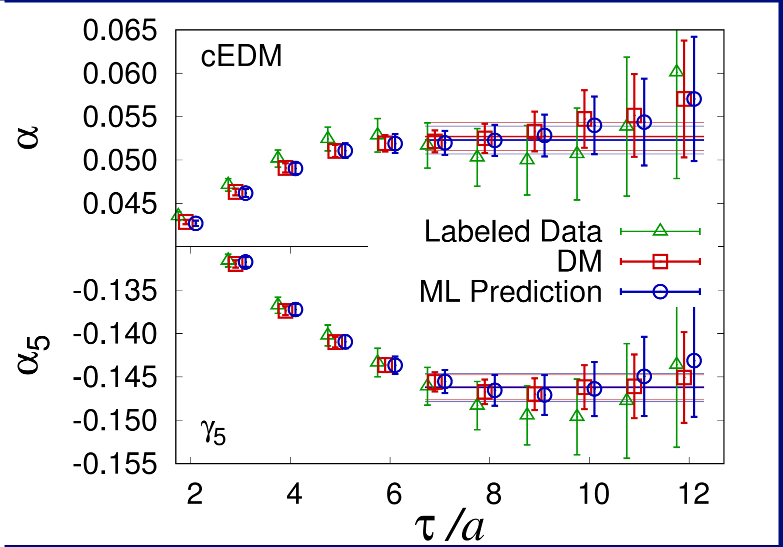
In all these cases, failure of ML does not change answer, only efficiency.

Example 1: Correlated observable

$\langle O\rangle = \langle O_s\rangle + \langle O - O_s\rangle$

$N^2/N_s^2 = \Delta(O - O_s)/\Delta O_s$



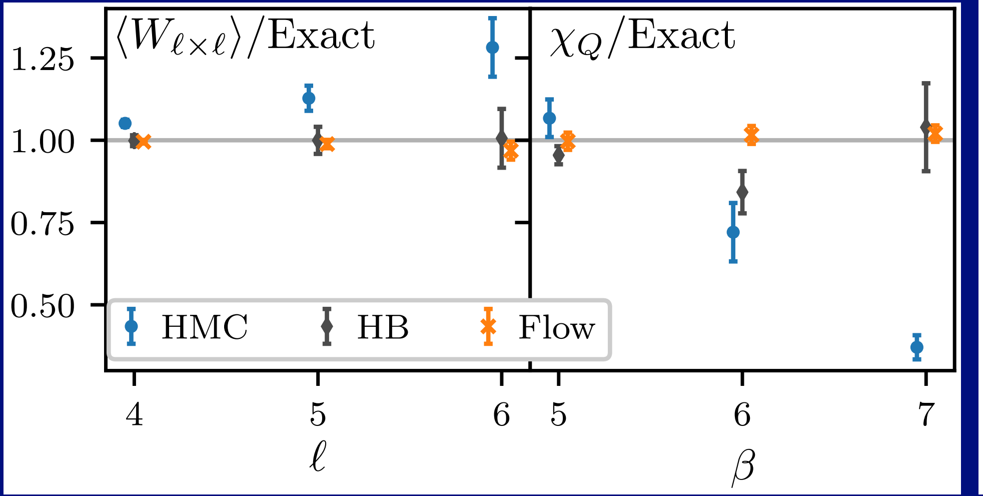


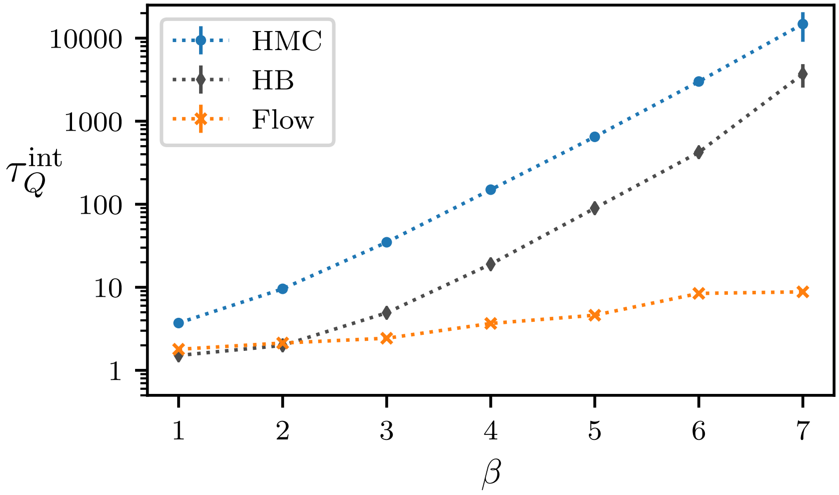
Example 2: Gauge Generation

Normalizing flows:

$$p_X\ (X = f\ (Z)) = p_Z(Z)\ J(Z,X)$$

Choose f such that p_X is simple and J(Z,X) is easily calculable.





arXiv:2003.06413

Conclusions

1. Machine Learning is the new frontier in using automation.
2. It takes the idea of black box prediction seriously.
3. Such high-dimensional interpolation is very powerful.
4. Large volume of data and requirement of low processing time is forcing us into using these methods.
5. Used properly, they do not lead to any compromise on integrity of the results.
6. It is an area of active research, both in experimental and theoretical research.

FIN