Selected topics from physics aware ML

Mateusz Ploskon, LBL
A Landscape of AI in Science

Data Analytics
- Classification
- Regression
- Clustering
- Dimensionality Reduction

Inverse problems
- Model reconstruction
- Parameter estimation
- Denoising

Surrogate models
- Approximate expensive simulations
- Approximate experiments
- Fill in missing models in simulations

Design and control
- Optimize design of experiments
- Control instruments
- Navigate state spaces
- Learn from sparse rewards

From AI4Sci / DOE
Physics-awareness – what is it?

• Think: **ML obeying conservation laws** (symmetries, invariances) – **incorporate physics into ML setup objective** (e.g. loss function)

• Benefits:
  • Trainable with simulated data;
  • Improving understanding of uncertainties and/or short comings of the modelling;
  • Expose non-physical components / noise
  • Build tools to improve simulators by working with **real** data (unsupervised learning) => if done in a well controlled way allows to gain knowledge => impact on analytic solutions / theory / modelling;
  • Towards explainable ML: eventually learn physics from the data alone
  • …
  • *note: in general, this is more than just apply “standard” ML techniques*
ML areas in physical sciences

Some challenges (and thus opportunities!) – national lab driven:

1. The biggest scientific datasets with complex, high-dimensional phase space
   • Challenging pattern recognition; benefits from **physics-aware learning**
   • Strict requirements on **uncertainty quantification / interpretability**

2. High-fidelity, first-principles simulations / theory
   • Can scaffold / exploit simulations with NNs for precise **likelihood-free inference**
   • Often too slow and need to be accelerated with **generative models**

3. Simulation-based inference is complemented by **data-driven learning**
   • **Anomaly detection** is becoming a key across the area
   • Often require fast inference (trigger), feedback (accelerator control), and/or environment awareness (hazards/safety)

Compilation Ben Nachman, MP
1. The biggest scientific datasets with complex, high-dimensional phase space
   - **Particle tracking**, noise mitigation, calibration, particle/jet/event classification, ...
   - Uncertainty quantification, interpretable observables/learning, robust learning...

2. High-fidelity, first-principles simulations / theory
   - Unfolding (deconvolution), **parameter estimation**, strong force dynamics, ...
   - Generative models for calorimeter emulation and cosmology; accelerator simulation

3. Simulation-based inference is complemented by data-driven learning
   - **Searching for new particles and forces**, mixed data/simulation labels, ...
   - **Accelerator stability / control**, learning in the presence of radioactive hazards, ...


Compilation Ben Nachman, MP
Recent example of **explainable**, physics aware ML

Explainable machine learning of the underlying physics of high-energy particle collisions


... proof-of-concept of our White Box AI approach using a Generative Adversarial Network (GAN) which learns from a DGLAP-based parton shower Monte Carlo event generator.

From “final state” particles learn internal workings of QCD – e.g. Altarelli-Parisi splitting function

Learning from final state distribution the internal splitting probability density \( f \)

Extract physics from the inner elements of the network – the “white box” to look at

\[
F(p^{in}) \rightarrow \{p_i\}
\]

\[
F(p^{in} \otimes f^n(z^k, \theta^k, \phi^k)) \rightarrow \{p_i\}
\]

Methodology extendable to other areas – e.g. modelling of nuclear reactions

\( z = \text{fraction of momentum carried by a particle in a 1->2 split} \)
Recent example of explainable, physics aware ML

Explainable machine learning of the underlying physics of high-energy particle collisions


... proof-of-concept of our White Box AI approach using a Generative Adversarial Network (GAN) which learns from a DGLAP-based parton shower Monte Carlo event generator.

From “final state” particles learn internal workings of QCD – e.g. Altarelli-Parisi splitting function

Learning from final state distribution the internal splitting probability density

Methodology extendable to other areas → e.g. modelling of nuclear reactions

Z = fraction of momentum of the original parton carried by a final state particle

f(z^k, θ^k, φ^k)

z = fraction of momentum carried by a particle in a 1->2 split
Other GAN applications

CaloGAN: Simulating 3D High Energy Particle Showers in Multi-Layer Electromagnetic Calorimeters with Generative Adversarial Networks
https://arxiv.org/abs/1712.10321

ForSE: a GAN based algorithm for extending CMB foreground models to sub-degree angular scales

CosmoGAN: creating high-fidelity weak lensing convergence maps using Generative Adversarial Networks
https://arxiv.org/abs/1706.02390
New horizons - challenges

• Explainable AI / interpretable ML
• Automated discovery
  • Physics laws inference or event of interest detection directly from data
  • Interesting discussion: https://www.frontiersin.org/articles/10.3389/frai.2020.00025/full
• Uncertainties evaluation
  • e.g.: Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty - https://arxiv.org/abs/2011.07586
• Learning on ensemble basis not only ‘event-based’
  • e.g.: E Pluribus Unum Ex Machina: Learning from Many Collider Events at Once - https://arxiv.org/abs/2101.07263
• Graph based ML – relational mapping
  • e.g.: A Comprehensive Survey on Graph Neural Networks - https://arxiv.org/abs/1901.00596
  • ... there is an increasing number of applications where data are generated from non-Euclidean domains and are represented as graphs with complex relationships and interdependency between objects...
Selected problems in physics aware ML

• Availability of models (& simulators)
  • Need for supervised vs. unsupervised (or partially supervised) learning

• Availability of data
  • ‘simple’ statistics – precision of measurements
    • Small variance – ‘mode collapse’ problem (also could be present in simulated sets)

• Quality of data
  • Purely labelled? not-labelled? Noisy or inconsistent?
  • Sufficient knowledge of uncertainties?
Extra Slides
Recent example of explainable, physics aware ML

**Explainable machine learning of the underlying physics of high-energy particle collisions**


... proof-of-concept of our White Box AI approach using a **Generative Adversarial Network (GAN)** which learns from a DGLAP-based parton shower Monte Carlo event generator.

From “final state” particles learn internal workings of QCD - Altarelli-Parisi splitting function, the ordering variable of the shower, and the scaling behavior.

- Two neural networks: the generator ("forger") and discriminator ("detective")
- Simultaneously optimize both, causing both to be in competition with each other (Nash equilibrium)
- Inner workings (splitting kernels) of the NNs forced to produce physical splittings
General scope ML4Sci(ence)

**Data**
- Experimental design
- Data curation and validation
- Compressed sensing
- Facilities operation and control

**Learning**
- Physics informed
- Reinforcement learning
- Adversarial networks
- Representation learning and multi-modal data
- “Foundational math” of learning

**Scalability**
- Algorithms, complexity and convergence
- Levels of parallelization
- Mixed precision arithmetic
- Communication
- Implementation on accelerated-node hardware

**Assurance**
- Uncertainty quantification
- Explainability and interpretability
- Validation and verification
- Causal inference

**Workflow**
- Edge computing
- Compression
- Online learning
- Federated learning
- Infrastructure
- Augmented intelligence
- Human-computer interface

From AI4Sci / DOE
AI Science Applications: One per Planet

**AI Enabled Design Workflows**
(tell me what to make)

...materials, polymers, organisms...

**AI Enabled Experimental Workflows**
(tell me how to make it)

...self-driving labs, synthesis search...

**AI Enabled Scientific Comprehension**
(tell me what it means)

- data Sets
- literature
- science “news”
- strategy

Cleaned
Updated
Annotated
Aggregated
Interpreted

→ Insight?
Ai4Sci Subtopics

Science and Engineering

- High Energy Physics
  - Cosmology and Astrophysics
  - Particle Physics
  - Accelerator Design and Control
- Nuclear Physics
  - Particle Identification
  - Particle Tracking
  - Control
- Fusion
  - Feature Tracking
  - Steering
  - Design and Control
- Energy Sciences
  - Materials and Chemistry Modeling
  - Photon/Neutron Science
  - Electron Microscopy
- Earth Sciences
  - Climate & Carbon
  - Subsurface
  - Water
- Biological Sciences
  - Environmental Biology
  - Imaging
  - Syn Biology
  - Health
- Energy Technologies
  - Engineering and Manufacturing
  - Smart Grid/Urban
  - Wind and Solar Mobility
- Computing Sciences
  - AI for hardware / software
  - AI for networks / computing facilities

Crosscuts

- Foundations
- Data Lifecycle
- Software
- Hardware
- Computing Facilities
Augmented Simulations

- Design
  - Biology
  - Materials
  - Chemistry
  - Devices
  - Batteries
  - Drugs

- Control
  - Accelerators
  - Reactors
  - Experiments
  - Simulation
  - Energy
  - Mobility

- Science and Math Comprehension
  - Physics
  - Biochemistry
  - Mathematics

- Generative Models
  - Cosmology
  - Biodesign
  - Detector Simulations

- Inverse Problems
  - Spectra2Structures
  - Image2Phase

- Multimodal Learning
  - Text
  - Images
  - Waveforms

- Decision-Making
  - Research Priorities
  - The Next Problem
  - Risk Assessment

From AI4Sci / DOE
ML at WANDA

• 2020 report – this talk was not to repeat it...

We strongly encourages entire ND community to embrace the advances that AI/ML tools can have for your work!

Emulation of expensive theory and application models for fast inference

Original

Emulated

Simulation

Emulation

Sobes

Schunck

Estimation of missing covariance information through AI/ML

AI/ML-augmented identification of regions to improve nuclear data

Neural Networks

Supervised Learning

Generative Modeling

Reinforcement Learning

Deep Q Learning

Bayesian Optimization
A new paradigm for data-driven, model-agnostic new physics searches at colliders is emerging, and aims to leverage recent breakthroughs in anomaly detection and machine learning. In order to develop and benchmark new anomaly detection methods within this framework, it is essential to have standard datasets. To this end, we have created the LHC Olympics 2020, a community challenge accompanied by a set of simulated collider events. Participants in these Olympics have developed their methods using an R&D dataset and then tested them on black boxes: datasets with an unknown anomaly (or not). This paper will review the LHC Olympics 2020 challenge, including an overview of the competition, a description of methods deployed in the competition, lessons learned from the experience, and implications for data analyses with future datasets as well as future colliders.
Food for a thought: Automated data mining

• Leverage NLP? – an example from other field...

Unsupervised word embeddings capture latent knowledge from materials science literature

Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A. Persson, Gerbrand Ceder & Anubhav Jain

Nature 571, 95–98 (2019) | Cite this article

https://www.nature.com/articles/s41586-019-1335-8

Named Entity Recognition and Normalization Applied to Large-Scale Information Extraction from the Materials Science Literature | Journal of Chemical Information and Modeling

https://pubs.acs.org/doi/10.1021/acs.jcim.9b00470

https://github.com/materialsintelligence/matscholar