DIRECT LEARNING OF SYSTEMATICS-AWARE SUMMARY STATISTICS

Pablo de Castro (@pablodecm) and Tommaso Dorigo (@dorigo)

INFN - Sezione di Padova / AMV4NewPhysics ITN

MOTIVATION: THE HEPML THREE-BODY PROBLEM

The three main analysis components only share processed data, each step is
carried out independently, without considering the remaining other two.

MODELLING UNCERTAINTIES DEGRADE INFERENCE

Simulations are imperfect, mainly due to the limited information of the
system being modelled
Lack of knowledge for inference accounted by additional unknown
parameters (nuisance parameters)
Causes a degradation of classifier-based inference, leading to larger measurement uncertainties

INFERN: INFERENCE-AWARE NEURAL OPTIMISATION

Applies on 2D Gaussian two-component mixture toy dataset, with unknown
background mean in one of the coordinates = one nuisance parameter
Loss is non-decomposable, because it is dataset-based instead of event-
based
Seems to converge independently on the initialization. Learning rate and/or batch size are critical hy-per-parameter.

SYNTHETIC EXAMPLE IMPLEMENTATION

UPPER LIMIT OF ML USEFULNESS IN LHC ANALYSES

CAN WE PUT IT ALL TOGETHER?

Embed some of the knowledge about modelling and statistical inference such
as systematic uncertainties in the dimensionality-reduction step

TRAINABLE PARAMETRIZED MODEL

Any (deep) neural network will do
Could in principle re-use the same

architecture and learn from density

FUNCTIONAL DISTANCE

E[true] − E[fit] ≥ 0

Comparing the true likelihood

with the mean of a set of

Gaussian random variables.

NEURAL NETWORK OUTPUT + SUMMARY STATISTIC

END-TO-END DIFFERENTIABILITY FOR LHC ANALYSES

Within this general framework, several approaches are possible, focus here is
direct learning of systemsatics-aware summary statistics.

COMPARISON WITH CLASSIFICATION-BASED APPROACH

BETTER/EQUAL THAN CLASSIFICATION

CONCLUSIONS AND PROSPECTS

Presented a machine learning approach that directly optimises an inference-
guided non-decomposable loss accounting for the effect of model
uncertainties
Flexibility of current autoDiff frameworks allows the inclusion of nuisance
parameters effect (via derivatives) over the training batches
The application of this approach and comparison with alternatives [5] to a
realistic systematic-dominated benchmark (e.g. systematic extended Higgs dataset) could shed some light on its real-world usefulness
Working on an update to the paper to be released together with TensorFlow
implementation code with more involved examples.

REFERENCES

1. R. Neal, Others, Computing likelihood functions for high-energy physics
experiments when distributions are defined by simulations with nuisance
2. K. Cranmer, J. Pavez, G. Louppe, Approximating Likelihood Ratios with
http://arxiv.org/abs/1506.02180];
5. J. Brehmer, G. Louppe, J. Pavez, K. Cranmer, Mining gold from implicit