https://indico.cern.ch/event/648004/contributions/3051475/

Scan QR code for web version
Poster at XIIIth QCHS
3rd August 2018
Maynooth (Ireland)

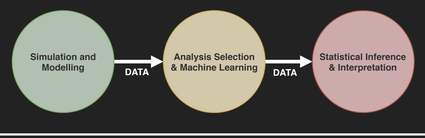
DIRECT LEARNING OF SYSTEMATICS-AWARE SUMMARY STATISTICS

Pablo de Castro (@pablodecm) and Tommaso Dorigo (@dorigo)
INFN - Sezione di Padova / AMVA4NewPhysics ITN

AMVA4NewPhysics has received funding from European Union's Horizon 2020 Programme under Grant Agreement number 675440

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.

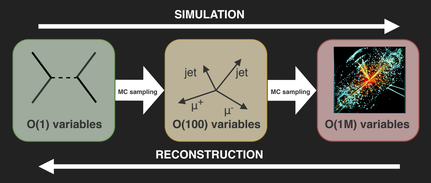


The ultimate aim of analyses is **powerful statistical inference** (i.e. interval estimation or hyphotesis testing) based on the observed data

IS THIS THE BEST WE CAN DO?

$\mathbf{p(x|model)}$ IS NOT KNOWN AT LHC EXPERIMENTS

Samples under different hypotheses can be simulated via complex physics-based MC but $p(\mathbf{x})$ cannot be directly evaluated → **LIKELIHOOD-FREE INFERENCE**



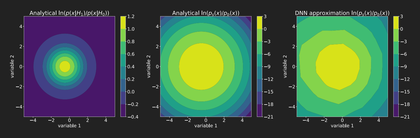
good approximations of $p(\mathbf{x})$ are unachievable due to curse of dimensionality

DIMENSIONALITY REDUCTION $\mathbf{R^n} \rightarrow \mathbf{R^{O(1)}}$ (SUMMARY STATISTIC)

KEEPING AS MUCH USEFUL INFORMATION FOR INFERENCE AS POSSIBLE

CLASSIFICATION AS A SURROGATE

Most machine learning classification algorithms approximate the likelihood ratio $p_s(\mathbf{x})/p_b(\mathbf{x})$ [1] (e.g. a DNN minimising cross entropy loss $-\sum_i k_i \log y_i$)

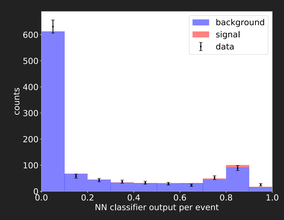


This can also be exploited to carry out inference directly [2]

CLASSIFIER-BASED INFERENCE

A trained ML classifer $d(\mathbf{x})$ is an uncalibrated approximation of $p_s(\mathbf{x})/p_b(\mathbf{x})$

How can it be used for statistical inference from observed data $\mathcal{D}$?



1-D → **cut or histogram to build a Poisson counts non-parametric likelihood**

$$\mathcal{L}(\mu) = \prod_{i \in bins} \text{Pois}(n_i | \mu \cdot s_i + b_i)$$

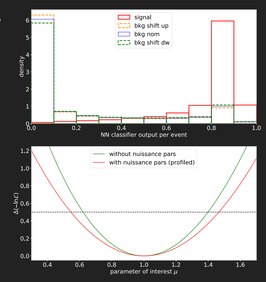
which can be used for further inference, such as measuring μ given observed $\mathcal{D}$

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 ν)

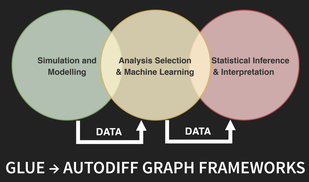
Causes a degradation of classifier-based inference, leading to larger measurement uncertainties



UPPER LIMIT OF ML USEFULNESS IN LHC ANALYSES [3]

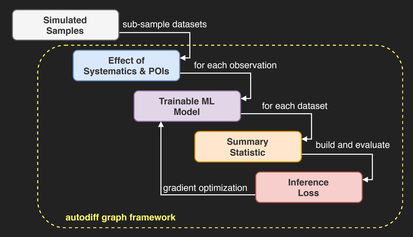
CAN WE PUT IT ALL TOGETHER?

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



GLUE → AUTODIFF GRAPH FRAMEWORKS

END-TO-END DIFFERENTIABILITY FOR LHC ANALYSES



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

MODELLING THE EFFECT OF SYSTEMATIC UNCERTAINTIES

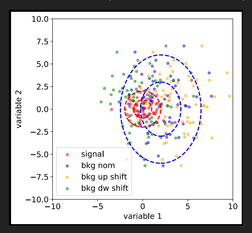
Differentiable approximation of the effect of parameters of interest θ and nuisance parameters ν

A non-linear function that depends on the details of problem and transforms event features/weights

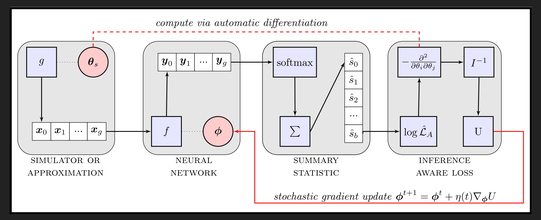
- jet/tau/muon energy scale
- PDF/QCD uncertainties

Could also depend on simulation/latent variables, such the event category (S/B)

Simple example: shift for background observations (i.e. $\mathbf{x}' = \mathbf{x} + \nu \cdot \mathbf{s}$)

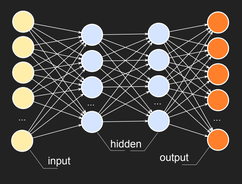


INFERNO: INFERENCE-AWARE NEURAL OPTIMISATION



check arxiv.org/abs/1806.04743 [4] for a detailed mathematical description

TRAINABLE PARAMETRIZED MODEL

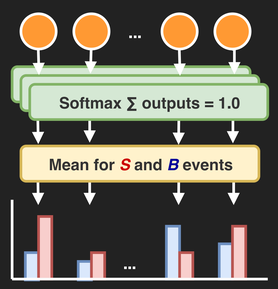


Any (deep) neural network will do

Could in principle re-use the same tweaks, techniques and architectures as for standard supervised deep learning

A small two-hidden layer MLP (10 units each, ReLU activation, glorot_normal initd) has been used for examples here.

NEURAL NETWORK OUTPUT → SUMMARY STATISTIC



We can approximate a histogram-like summary statistic from the NN output applying softmax for each event and summing over each dataset

$$\mathcal{L}(\theta, \nu) = \prod_{i \in bins} \text{Pois}(n_i | \alpha_s s_i + \alpha_b b_i)$$

The likelihood depends both on the neural network parameters and the statistical model parameters

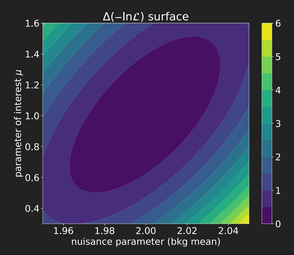
INFERENCE-MOTIVATED LOSS FUNCTION

If we expand the negative log-likelihood around minimum (e.g. Asimov $n_i = \alpha_s s_i + \alpha_b b_i$), due to Cramér-Rao bound:

$$\text{covariance} \geq \mathbf{H^{-1}}(-\ln \mathcal{L})$$

which can be computed via autodiff. Can use as loss function directly the variance bound on the parameters of interest

$$\text{loss} \approx \text{Var}(\mu) \quad \text{(expected)}$$

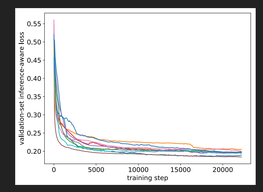


SYNTHETIC EXAMPLE IMPLEMENTATION

Applied 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 mini-batch size are critical hyper-parameter.



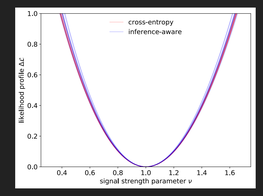
COMPARISON WITH CLASSIFICATION-BASED APPROACH

Expected signal strength uncertainties computed using the validation set (average of 10 random initialisations):

- cross-entropy: 0.444 ± 0.003
- **inference-aware: 0.437 ± 0.008**

Early results are encouraging but additional study of the technique needed.

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 systematics-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 simulators with nuisance parameters (2008) (available at http://cds.cern.ch/record/1099977).
2. K. Cranmer, J. Pavez, G. Louppe, Approximating Likelihood Ratios with Calibrated Discriminative Classifiers. arXiv [stat.AP] (2015), (available at http://arxiv.org/abs/1506.02169).
3. G. Louppe, M. Kagan, K. Cranmer, Learning to Pivot with Adversarial Networks. arXiv [stat.ML] (2016), (available at http://arxiv.org/abs/1611.01046).
4. P. de Castro, T. Dorigo, INFERNO: Inference-Aware Neural Optimisation. arXiv [stat.ML] (2018), (available at http://arxiv.org/abs/1806.04743).
5. J. Brehmer, G. Louppe, J. Pavez, K. Cranmer, Mining gold from implicit models to improve likelihood-free inference. arXiv [stat.ML] (2018), (available at http://arxiv.org/abs/1805.12244).