Karlsruhe Institute of Technology

Optimal statistical inference in the presence of systematic uncertainties using neural network optimization based on binned Poisson likelihoods with nuisance parameters

Stefan Wunsch, Simon Jörger, Roger Wolf, Günter Quast
stefan.wunsch@cern.ch

KIT ETP / CERN EP-SFT

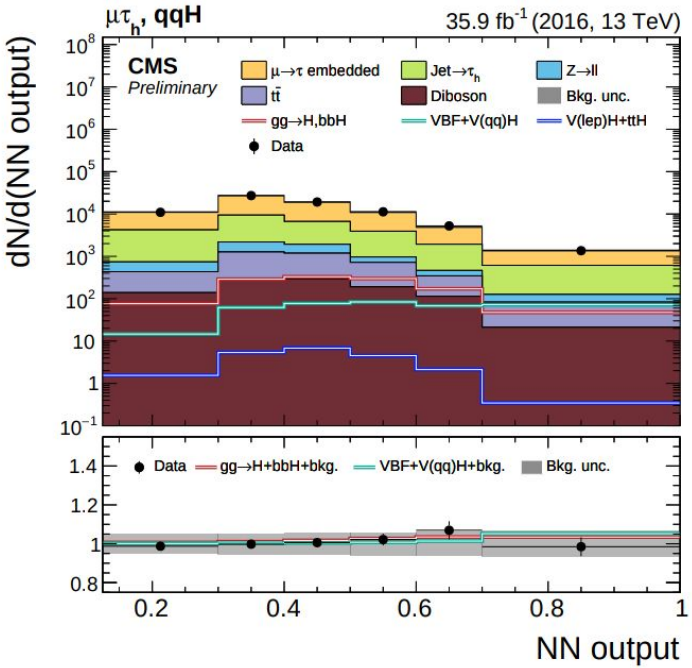
Introduction

- Machine learning is more and more often part of the very-end data analysis toolchain in HEP and other fields of science

- Often used are neural networks trained as classifiers
  - Separating signal(s) vs background(s)
  - Using cross entropy function as loss
  - Fit the NN output as discriminative variable

- **Why cross entropy seems to be a good choice?**

- **Is there a better or even optimal analysis strategy?**



Signal category from the CMS public analysis note HIG-18-032 used to measure the Higgs boson cross section

What is the cross entropy?

- The cross entropy is closely related to the definition of a (log) likelihood, e.g., for binary classification:

$$\log \mathcal{L} = \log \prod_i p(x_i) = \log \prod_i f_i^{y_i} \cdot (1-f_i)^{1-y_i} = \sum_i y_i \log f_i + (1-y_i) \log(1-f_i) = \mathrm{CE}$$

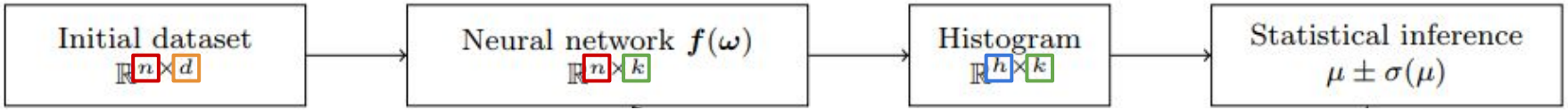
$$x_i = \text{Input}$$
$$f_i = \text{NN output}$$
$$y_i = \text{Label}$$

- It is possible to prove that a NN function trained on binary classification is a sufficient statistic to infer the signal strength μ in a two-component mixture model p(x | μ·s + b) without nuisance parameters (see the appendix in the INFERNO paper).

The cross entropy loss is optimal if the analysis takes only statistical uncertainties into account.

Can we do better if we include also systematic uncertainties in the loss?

One step back: (Binned) data analysis in HEP



Dimensionality of the dataset: $R^{a \times b}$

n Number of events

d Number of observables (pt, mass, missing energy, …)

k Number of high-level observables (neural network output, invariant mass of the decay system, …)

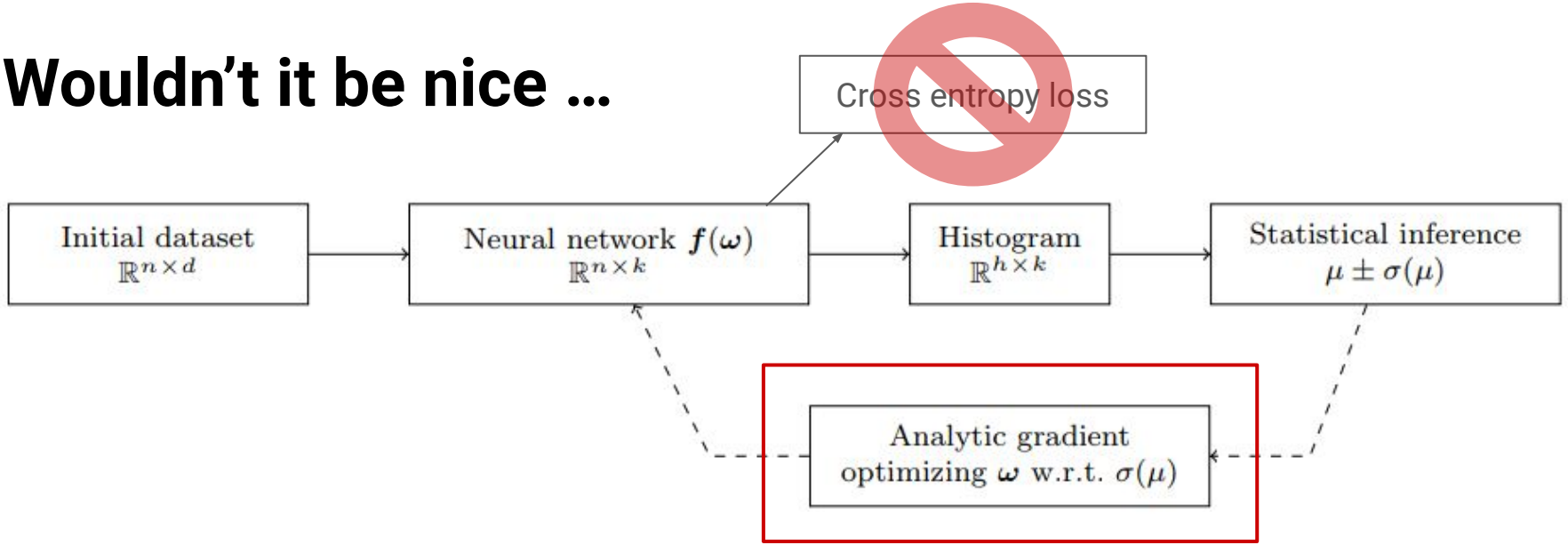
h Number of bins in the histogram

Statistical inference

Profile of the binned Poisson likelihood including all statistical and systematic uncertainties

This workflow covers typical analyses performed in CMS and ATLAS, e.g., the Higgs discovery.

Wouldn’t it be nice …



… if we could optimize directly on the objective of the statistical inference?

Instead of training on the cross entropy loss, we could optimize directly the objective of the analysis, e.g., the uncertainty of the estimated signal strength σ(μ).

Statistical inference

$$\mathcal{L}(D_{\mathrm{H}}, \boldsymbol{\theta}) = \prod_{i=0}^{h} \mathcal{P}\left(d_i | \mu s_i + b_i + \eta \Delta_i\right) \mathcal{N}\left(\eta\right)$$

P Poisson distribution
d Observation
s Signal expectation
b Background expectation

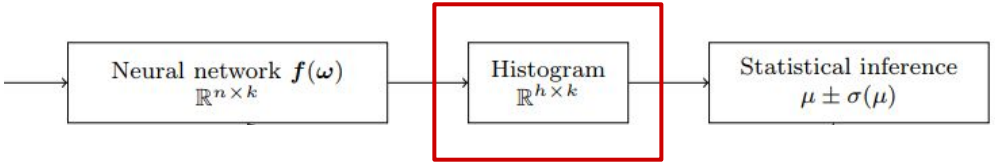
μ Signal strength modifier
η Nuisance parameter
Δ Systematic variation
N Normal distribution

$$F_{ij} = \frac{\partial^2}{\partial\theta_i \partial\theta_j} \left(-\log \mathcal{L}(D_{\mathrm{H}}, \boldsymbol{\theta})\right) \longrightarrow V_{ij} = F_{ij}^{-1}$$

F_{ij} Fisher information
V_{ij} Covariance matrix

- V_{ij} is the exact variance of the estimator, e.g., for μ, if the likelihood is parabolic
- Using Asimov data representing the median expected performance
- **Signal strength constraint $V_{00} = \sigma(\mu)^2$ used as objective for the NN optimization**

What is the problem?

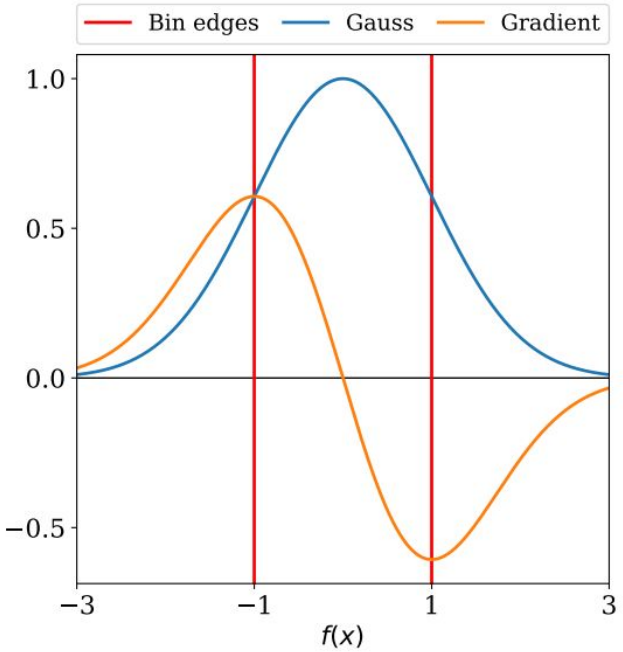


- NN optimization is based on automatic differentiation using the chain rule (aka backpropagation)

$$\frac{dz}{dx} = \frac{dz}{dy} \cdot \frac{dy}{dx}$$

- The bin function has a gradient, which is
  - zero in the bin
  - undefined on the edges
  - not suited for the backpropagation

- **Solution** Approximate the gradient of the bin function
  - Forward pass not changed
  - Gradient replaced by derivative of a Gauss function



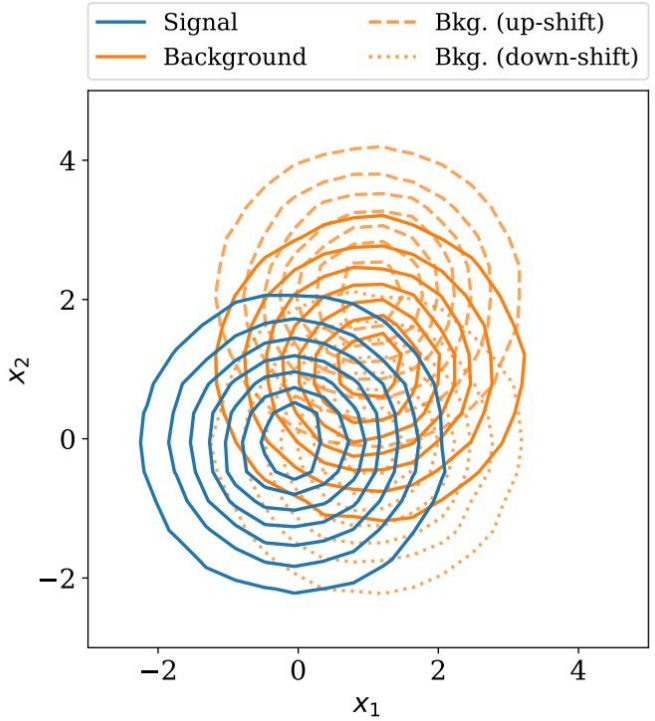
Simple example based on pseudo-experiments

- **Two processes** Signal and background
- **Two variables** x_1 and x_2
- **Systematic uncertainty** $x_2 \pm 1$ for the background process

- Systematic variation can be implemented as
  - Reweighting on histogram level
  - Simulation on input level (done here)

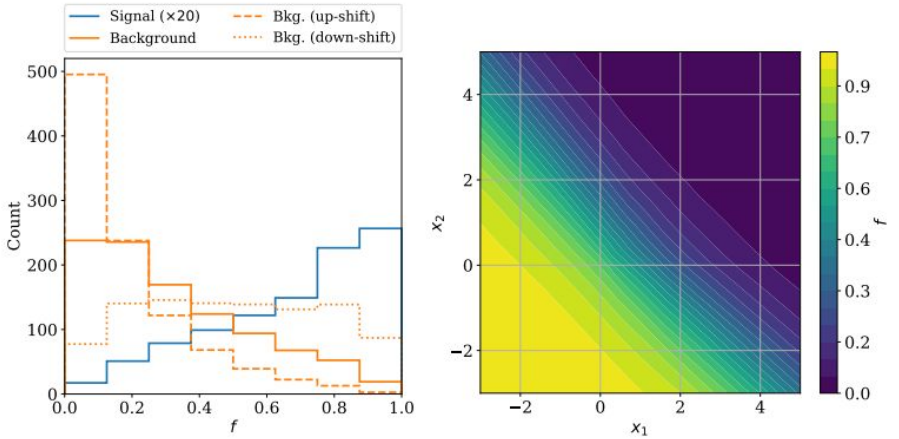
- **Architecture**
  - Fully connected feed-forward network
  - 1 hidden layer with 100 nodes
  - ReLU and sigmoid activation

- Use likelihood evaluated on 100k events for each gradient step

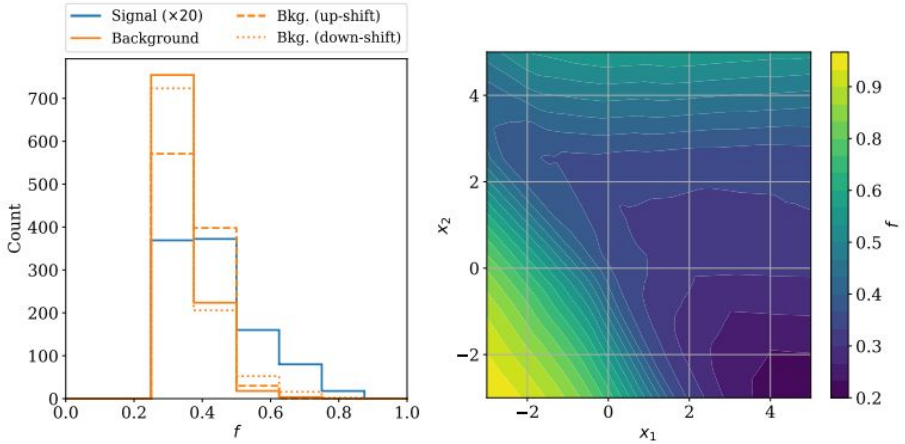


Comparison of the neural network functions

- Training on NLL loss (V_{00}) reduces the impact of the systematic variation in signal enriched bins
- Neural network function in the input space shows mitigation of the phase space with high impact of the systematic



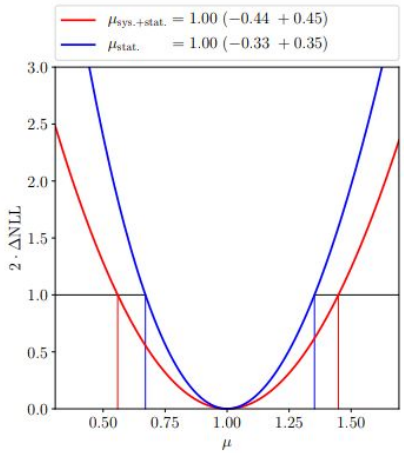
Training on cross entropy (CE) loss



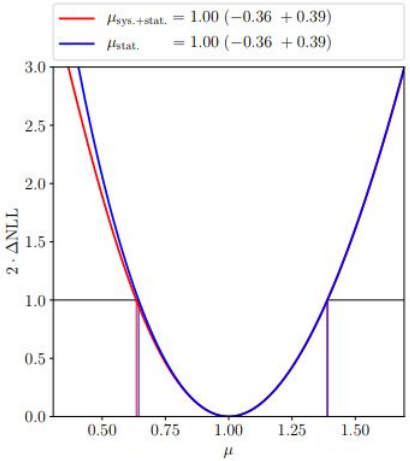
Training on NLL loss

Is it optimal?

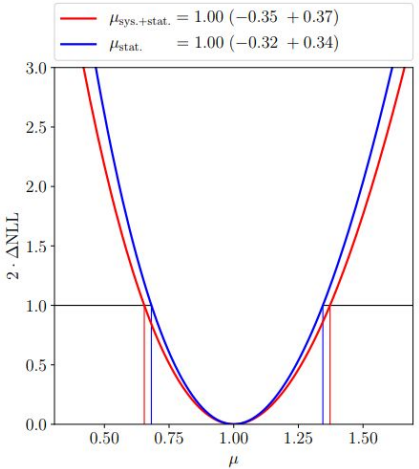
Cross entropy (CE) training
(binned fit)



NLL training
(binned fit)



Optimal result
(unbinned fit)



- Shown are profiles of the likelihood with Asimov data (expected results)
- NLL compared to CE reduces σ(μ) by 16%
- Optimal result given by unbinned fit in the 2D input space
- Residual difference in σ(μ) between NLL and optimal result is 4%
- NLL compared to CE reduces correlation of μ to η from 64% to 13%

NLL training results in an analysis strategy which is close to optimal

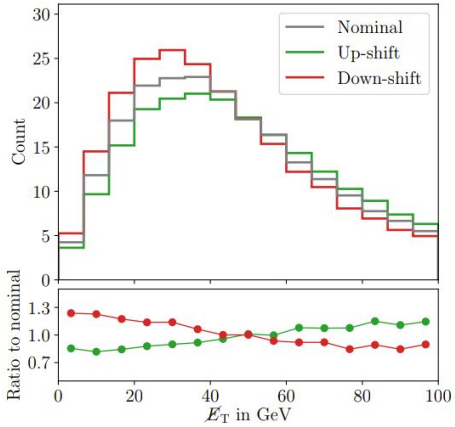
More complex example typical for HEP analysis

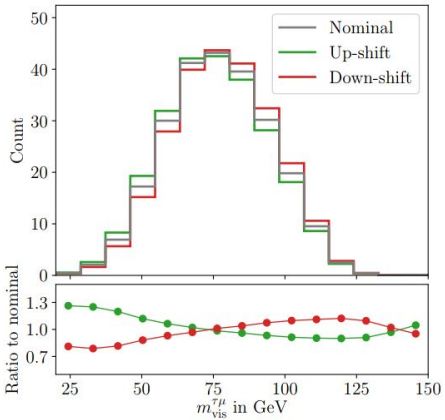
- Dataset from the Kaggle Higgs challenge with two processes containing signal and a mixture of backgrounds

- **Enhanced by a systematic variation**
  - Introduced a ±10% shift of the missing transverse energy
  - Propagated to all other variables via reweighting

- Using only three variables as input of the NN
  - Visible mass of the Higgs system
  - Transverse momentum of the Higgs system
  - Absolute difference in the pseudorapidity of the two leading jets
  - Missing transverse energy explicitly not included to create a more complex scenario

- Otherwise, same setup than for the simple example

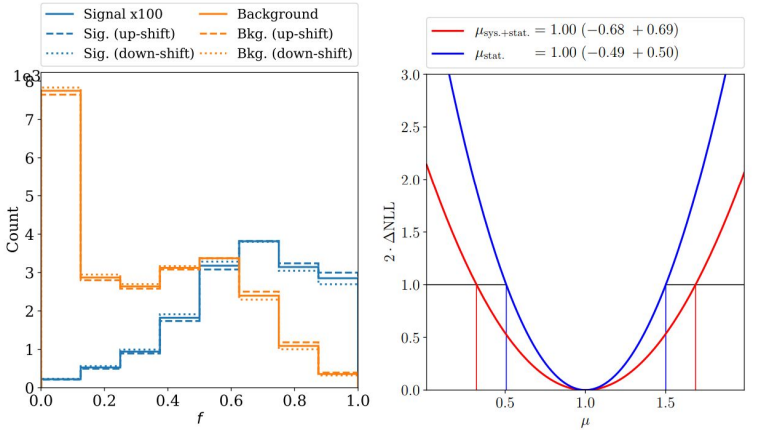




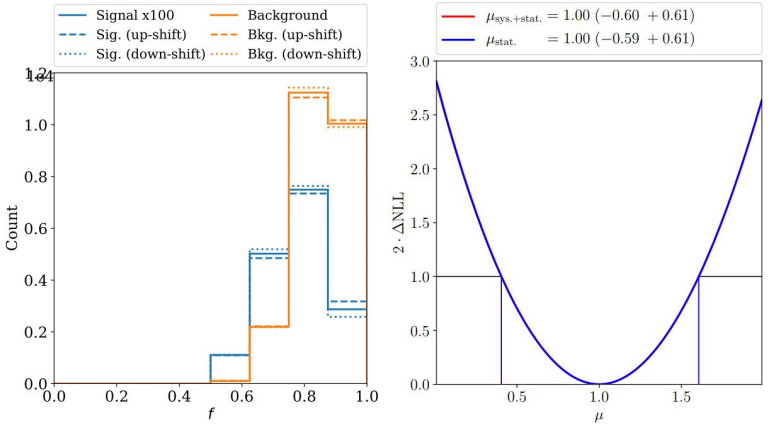
CE vs NLL loss

- Shown are profiles of the likelihood with Asimov data (expected results)
- NLL compared to CE reduces σ(μ) by 12%
- Not possible to compare to optimal result since unbinned likelihood is not known
- NLL compared to CE reduces correlation of μ to η from 69% to 4%

The proposed approach successfully optimized the analysis fully automatically



Training on cross entropy (CE) loss



Training on NLL loss

Further information and related work

- Full paper (preprint) available on arXiv: https://arxiv.org/abs/2003.07186

Optimal statistical inference in the presence of systematic uncertainties using neural network optimization based on binned Poisson likelihoods with nuisance parameters

Stefan Wunsch · Simon Jörger · Roger Wolf · Günter Quast

- Related publication using a similar technique to reduce the dependence of the NN function to systematics in the input space: https://arxiv.org/abs/1907.11674

Original Article | Open Access | Published: 23 February 2020

Reducing the Dependence of the Neural Network Function to Systematic Uncertainties in the Input Space

Stefan Wunsch, Simon Jörger, Roger Wolf & Günter Quast

Computing and Software for Big Science 4, Article number: 5 (2020) | Cite this article

- INFERNO discusses a similar approach but uses a sum over a softmax as summary statistic and is therefore only useable for likelihood free inference: https://arxiv.org/abs/1806.04743

INFERNO: Inference-Aware Neural Optimisation

Pablo de Castro
INFN - Sezione di Padova
pablo.de.castro@cern.ch

Tommaso Dorigo
INFN - Sezione di Padova
tommaso.dorigo@cern.ch

- Related publication with similar approach like INFERNO: https://arxiv.org/abs/1802.03537

Automatic physical inference with information maximizing neural networks

Tom Charnock, Guilhem Lavaux, and Benjamin D. Wandelt
Phys. Rev. D **97**, 083004 – Published 13 April 2018

Summary

- **Proposal of a novel approach to optimize data analysis based on binned likelihoods**
  - System is fully analytically differentiable thanks to an approximated gradient for the histogram
  - Use objective of the analysis directly for the optimization, e.g., constraint of the signal strength modifier

- **Simple example based on pseudo-experiments proves that the strategy finds an optimal solution**
  - Successful integration of information about systematic uncertainties in the optimization of the neural network

- **Feasibility study in a more complex example typical for HEP analysis**
  - Approach supports integration of a statistical model typical to HEP analysis, e.g., such as done by HistFactory or combine
  - Possible to include systematic variations defined on histogram level by reweighting techniques