Computing and Software for Big Science (2021) 5:4
https://doi.org/10.1007/s41781-020-00049-5

ORIGINAL ARTICLE

Optimal Statistical Inference in the Presence of Systematic Uncertainties Using Neural Network Optimization Based on Binned Poisson Likelihoods with Nuisance Parameters

Stefan Wunsch[1,2] · Simon Jörger[1] · Roger Wolf[1] · Günter Quast[1]

Received: 5 May 2020 / Accepted: 7 December 2020

Abstract

Data analysis in science, e.g., high-energy particle physics, is often subject to an intractable likelihood if the observables and observations span a high-dimensional input space. Typically the problem is solved by reducing the dimensionality using feature engineering and histograms, whereby the latter allows to build the likelihood using Poisson statistics. However, in the presence of systematic uncertainties represented by nuisance parameters in the likelihood, an optimal dimensionality reduction with a minimal loss of information about the parameters of interest is not known. This work presents a novel strategy to construct the dimensionality reduction with neural networks for feature engineering and a differential formulation of histograms so that the full workflow can be optimized with the result of the statistical inference, e.g., the variance of a parameter of interest, as objective. We discuss how this approach results in an estimate of the parameters of interest that is close to optimal and the applicability of the technique is demonstrated with a simple example based on pseudo-experiments and a more complex example from high-energy particle physics.

Keywords Optimal statistical inference · Neural networks · High-energy particle physics

Introduction

Measurements in many areas of research like, e.g., high-energy particle physics, are typically based on the statistical inference of one or more parameters of interest defined by the likelihood $\mathcal{L}(D, \boldsymbol{\theta})$ with the observables $\boldsymbol{x} \in X \subseteq \mathbb{R}^d$ building the dataset $D = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_n\} \subseteq \mathbb{R}^{n \times d}$ and the parameters $\boldsymbol{\theta}$ of the statistical model. The likelihood would have to be evaluated for the dataset D spanning a high-dimensional input space, which is computationally expensive and typically unfeasible. The dimension of D can be reduced by the engineering of high-level observables and the usage of summary statistics. Analysts create high-level observables to reduce the dimension d of a single observation to k, ideally without losing information about the parameters $\boldsymbol{\theta}$. An example from high-energy particle physics is the usage of the invariant mass of a decay system instead of the kinematic properties of all its constituents. The dimension n of D can be reduced with the computation of a summary statistic, for which histograms are frequently used so that the statistical model can be expressed in form of a likelihood, based on Poisson statistics. The dimension is thus reduced from the number of observations n to the number of bins h in the histogram, whereby the analyst tries to optimize the trade-off between a feasible number of bins and the loss of information about the parameters of interest. Applying both methods, the initial dimension of $D \subseteq \mathbb{R}^{n \times d}$ is reduced to $\mathbb{R}^{h \times k}$.

This paper discusses an analysis strategy using machine learning techniques, by which the suboptimal performance introduced by the reduction of dimensionality can be avoided resulting in estimates of the parameters of interest $\boldsymbol{\mu} \in \boldsymbol{\theta}$ close to optimal, e.g., with a minimal variance. We put emphasis on the applicability of this approach to analysis strategies as

✉ Stefan Wunsch
stefan.wunsch@cern.ch

Simon Jörger
simon.joerger@cern.ch

Roger Wolf
roger.wolf@cern.ch

Günter Quast
guenter.quast@kit.edu

[1] Karlsruhe Institute of Technology, Institute of Experimental Particle Physics, Karlsruhe, Germany

[2] CERN, Geneva, Switzerland

Published online: 12 January 2021

used for Higgs boson analyses at the Large Hadron Collider (LHC) [1–4].

Section “Methods” presents the method in detail and “Related Work” puts the proposed technique in context of related work. Section “Application to a Simple Example Based on Pseudo-Experiments” shows the performance of the method with a simple example using pseudo-experiments of a two-component mixture model with signal and background and “Application to a More Complex Analysis Task Typical for High-Energy Particle Physics” applies the same approach to a more complex example from high-energy particle physics.

Methods

The method requires as input an initial dataset $D \subseteq \mathbb{R}^{n \times d}$ used for the statistical inference of the parameters of interest with n being the number of observations and d the number of observables. To simplify the statistical evaluation, we typically reduce the number of observables by the engineering of high-level observables. Besides manual crafting of such features, a suited approach taken from machine learning is using a neural network (NN) function $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\omega}) : \boldsymbol{x} \in X \subseteq \mathbb{R}^d \rightarrow \boldsymbol{f} \in F \subseteq \mathbb{R}^k$ with $\boldsymbol{\omega}$ being the free NN parameters. After application of the NN, we get a transformed dataset $D_{\mathrm{NN}} \subseteq \mathbb{R}^{n \times k}$ with k the number of output nodes of the NN architecture.

To reduce the dataset D_{NN} further, the number of observations n is compressed using a histogram. Histograms are widely used as a summary statistic since counts follow the Poisson distribution and therefore are well suited to build the statistical model of the analysis. For example in high-energy particle physics, many statistical models and well established methods for describing the statistical model and systematic uncertainties are based on counts and binned Poisson likelihoods. The resulting dataset is $D_{\mathrm{H}} \subseteq \mathbb{R}^{h \times k}$ using h bins for the k-dimensional histogram. The count operation for a single bin in the histogram can be written as $C = \sum_{i=1}^{n} S\big(\boldsymbol{f}(\boldsymbol{x}_i, \boldsymbol{\omega})\big)$ with

$$S\big(\boldsymbol{f}(\boldsymbol{x}_i, \boldsymbol{\omega})\big) = \begin{cases} 1, & \text{if } \boldsymbol{f} \text{ in the bin boundaries} \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

In order to propagate the gradient from the result of the statistical inference to the free parameters $\boldsymbol{\omega}$ of the NN, the histogram has to be differentiable. Since the derivative of S is ill-defined on the edges of the bin and otherwise zero, the gradient is not suitable for optimization. Therefore, we use a smoothed approximation of the gradient [5] shown in Fig. 1 for a one-dimensional bin. The approximation uses the similarity of S to a Gaussian function $\mathcal{G}$ normalized to $\max(\mathcal{G}) = 1$ with the standard deviation being the half-width of the bin. We replace only the gradient of the operation S and not the calculation of the count itself to keep the statistical model of the final analysis unchanged.

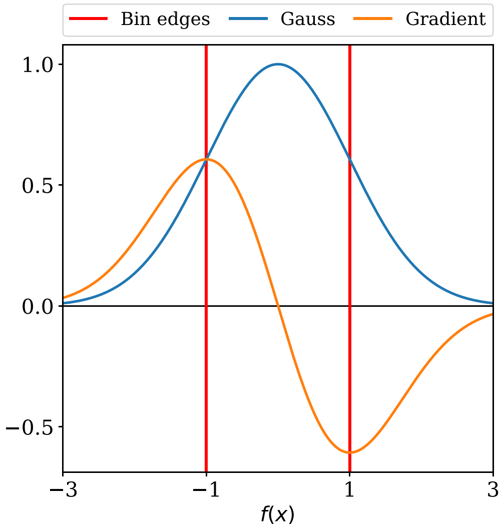

Fig. 1 The figure shows the approximation of the gradient of a single bin in a histogram with the gradient of a Gaussian G normalized to $\max(G) = 1$ with the standard deviation equal to the half-width of the bin

On top of the reduced dataset D_{H}, we build the statistical model using a binned likelihood $\mathcal{L}(D_{\mathrm{H}}, \boldsymbol{\theta})$ with $\boldsymbol{\theta}$ being the parameters of the statistical model. For a mixture model with the two processes signal $\boldsymbol{s}$ and background $\boldsymbol{b}$, the binned likelihood describing the statistical component is given by

$$\mathcal{L}(D_{\mathrm{H}}, \boldsymbol{\theta}) = \prod_{i=1}^{h} \mathcal{P}\big(d_i | \mu s_i + b_i\big) \tag{2}$$

with $\mathcal{P}$ being the Poisson distribution, $\boldsymbol{d}$ the observation and $\mu \in \boldsymbol{\theta}$ the parameter of interest scaling the expectation of the signal process $\boldsymbol{s}$.

Moreover, the formulation of the statistical model allows to implement systematic uncertainties by adding nuisance parameters to the set of parameters $\boldsymbol{\theta}$. For the model in Eq. 2, a single nuisance parameter η controlling a systematic variation $\boldsymbol{\Delta}$ of the expected bin contents results in

$$\mathcal{L}(D_{\mathrm{H}}, \boldsymbol{\theta}) = \prod_{i=1}^{h} \mathcal{P}\big(d_i | \mu s_i + b_i + \eta \Delta_i\big)\, \mathcal{N}(\eta) \tag{3}$$

with $\mathcal{N}$ being a standard normal distribution constraining the nuisance η. If the systematic variation is asymmetric, the additional nuisance term can be written as

$$\max(\eta, 0)\boldsymbol{\Delta}_{\mathrm{up}} + \min(\eta, 0)\boldsymbol{\Delta}_{\mathrm{down}} \tag{4}$$

or with any other differential formulation [6].

The performance of an analysis is measured in terms of the variance of the estimate for the parameters of interest, for example in our case the variance of the estimated signal strength μ. We built a differential estimate of the variance using the Fisher information [7] of the likelihood in Eq. 3 given by

$$F_{ij} = \frac{\partial^2}{\partial\theta_i \partial\theta_j} \left(-\log \mathcal{L}(D_{\mathrm{H}}, \boldsymbol{\theta})\right). \tag{5}$$

Because the maximum likelihood estimator is asymptotically efficient [8, 9], the variance of the estimates for $\boldsymbol{\theta}$ is asymptotically close to

$$V_{ij} = F_{ij}^{-1}. \tag{6}$$

Assuming the first diagonal element to correspond to the parameter of interest μ, without loss of generality, the loss function to optimize the variance of the estimate for μ with respect to the free parameters $\boldsymbol{\omega}$ of the NN function $\boldsymbol{f}$ is V_{00}.

To be independent of the statistical fluctuations of the observation, the optimization is performed on an Asimov dataset [1]. This artificial dataset replaces the observation $\boldsymbol{d}$ with the nominal expectation for $\boldsymbol{s}$ and $\boldsymbol{b}$ serving as representative for the median expected outcome of the analysis in presence of the signal plus background hypothesis.

Given the assumption that the dimensionality reduction performed by the NN together with the histogram is a sufficient statistic, the optimization can find a function for $\boldsymbol{f}$ that gives the best estimate for the parameter of interest μ, similar to a statistical inference performed on the initial high-dimensional dataset D with an unbinned likelihood.

A graphical overview of the proposed method is given in Fig. 2.

Related Work

The authors of [10] were first to develop an approach that also optimizes the parameters of an NN based on the binned Poisson likelihood of the analysis. They also identify the problem that a histogram has no suitable derivative but follow a different strategy to enable automatic differentiation replacing the counts with means of a softmax function. This approach is also followed by [11].

Likewise Ref. [12] estimates a count with the sum of the NN output values but uses an inclusive estimate of the significance as training objective, which results in an improved analysis objective in a search for new physics.

The strategy to allow a NN to find the best compression of the data has also been discussed in [13]. The authors show that the NN is able to learn a summary statistic that is a close approximation of a sufficient statistic, yielding a powerful statistical inference.

A related approach to training the NN on the statistical model of the analysis including systematic uncertainties is the explicit decorrelation against the systematic variation. For example, the idea has been discussed on the basis of an adversarial architecture [14–16], a penalty term based on distance correlation [17] and an approach penalizing the variation using approximated bin counts [5]. These strategies are not aware of the analysis objective such as the variance of a parameter of interest and therefore the decorrelation is subject to manual optimization. For a large number of nuisances, this optimization procedure is computational expensive and typically unfeasible.

Another approach to optimize the statistical inference is the direct estimation of the likelihood in the input space, which is typically carried out using machine learning techniques. Such methods intend to use the approximated likelihood in the input space for the statistical inference, which avoids the dimensionality reduction that is optimized with the proposed method. The technical difficulties to carry out these methods are discussed in [18, 19].

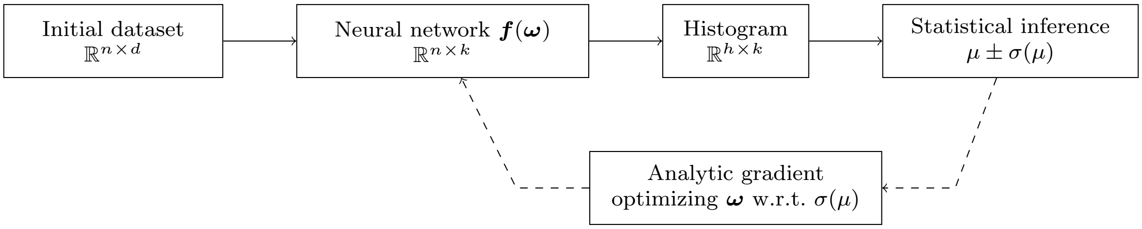

Fig. 2 Graphical overview of the proposed method to optimize the reduction of the dataset used for the statistical inference of the parameters of interest from end to end. The number of observables d in the initial dataset with n observations is reduced to a set of k observables by the neural network function $\boldsymbol{f}$ with the free parameters $\boldsymbol{\omega}$. The dataset is compressed further by summarizing the n observations using a k-dimensional histogram with h bins. Eventually the free parameters $\boldsymbol{\omega}$ are optimized with the variance of the parameter of interest μ as objective, which is made possible by an approximated gradient for the histogram

Application to a Simple Example Based on Pseudo-Experiments

A simple example based on pseudo-experiments and a known likelihood in the input space $\mathbb{R}^{n\times d}$ is used to illustrate our approach. The distributions of the signal and background components in the input space are shown in Fig. 3. We assume a systematic uncertainty on the mean of the background process modelled by the shifts $x_2 \pm 1$, representing the systematic variations in Eq. 4.

The NN architecture is a fully-connected feed-forward network with 100 nodes in one hidden layer. The initialization follows the Glorot algorithm [20] and the activation function is a rectified linear unit [21]. The output layer has a single node with a sigmoid activation function.

We use eight bins for the histogram of the NN output and compute the variance of the estimate for the parameter of interest μ denoted by V_{00}. The operations are implemented using TensorFlow as computational graph library [22, 23] and we use the provided automatic differentiation and the Adam algorithm [24] to optimize the free parameters $\boldsymbol{\omega}$ with the objective to minimize V_{00}. The systematic variations $\boldsymbol{\Delta}$ can be implemented with reweighing techniques using statistical weights or duplicates of the nominal dataset with the simulated variations, whereas we chose the latter solution. Each gradient step is performed on the full dataset with 10^5 simulated events for each process. The dataset is split in half for training and validation, and all results are computed from a statistically independent dataset of the same size as the original one. The training is stopped if the loss has not improved for 100 gradient steps eventually using the model with the smallest loss on the validation dataset for further analysis. We found that the convergence is more stable if the model is first optimized only on the statistical part of the likelihood shown in Eq. 2 and therefore apply this pre-training for 30 gradient steps. We apply statistical weights to scale the expectation of signal and background to 50 and 10^3, respectively.

The best possible expected result in terms of the variance of the estimate for μ is given by a fit of the unbinned statistical model without dimensionality reduction. Alternatively, we can get an asymptotically close result by using a binned likelihood with sufficiently large number of bins in the two-dimensional input space. The latter approach with 20×20 equidistant bins in the range shown in Fig. 3 results in the profile shown in Fig. 4 with $\mu = 1.0^{+0.37}_{-0.35}$. The best-fit value of μ is always at 1.0 because of the used Asimov dataset. Further, we find the uncertainty of μ in all fits by profiling the likelihood [25] rather than using the approximation by the covariance matrix in Eq. 6. We obtain all results in this paper with validated statistical tools, RooFit and RooStats [26–28], such as used by most publications analyzing data of the LHC experiments.

The first comparison to this best-possible result is done by training the NN not on the variance of the estimate for μ, V_{00}, but on the cross entropy loss with signal and background weighted to the same expectation. This approach has been used in multiple analyses in high-energy particle

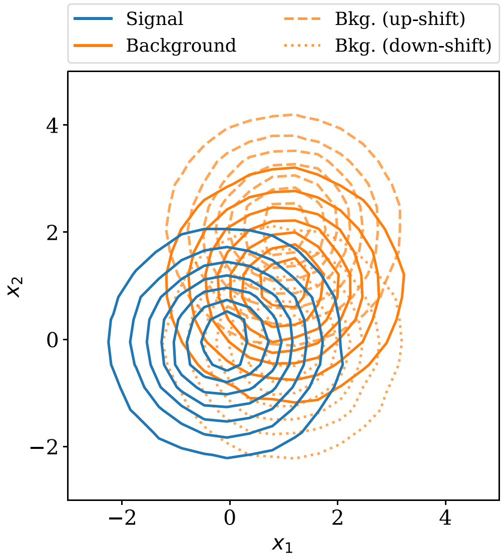

Fig. 3 Distribution of the signal and background components in the input space modelled by multivariate Gaussian distributions centered around $(0\ \ 0)$ and $(1\ \ 1)$ with the covariance matrix $\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$. We introduce a systematic variation that shifts the mean of the background component along x_2

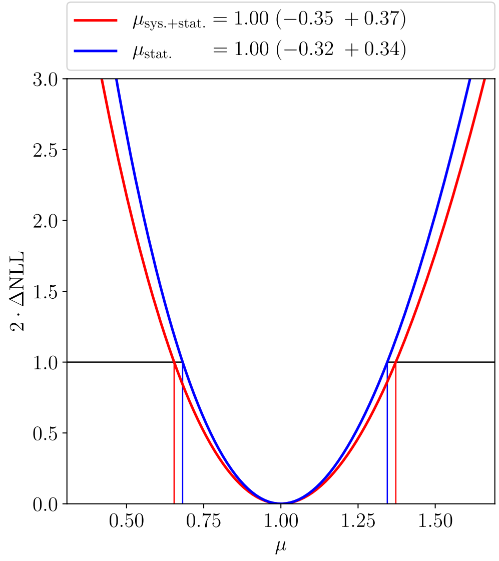

Fig. 4 Profile of the likelihood with (blue line) only the statistical uncertainty and (red line) the systematic uncertainty in addition for the likelihood defined in the two-dimensional input space spanned by x_1 and x_2 as given in Fig. 3

physics [29, 30]. The NN function $\boldsymbol{f}$ is a sufficient statistic - and therefore optimal - if no systematic uncertainties have to be considered for the statistical inference such as the likelihood in Eq. 2 [10]. The resulting function $\boldsymbol{f}$ is shown in the input space and by the distribution of the output in Fig. 5. The NN learns to project the two-dimensional space spanned by x_1 and x_2 on the diagonal, which is trivially the optimal dimensionality reduction in this simple example. If we apply the statistical model including the systematic uncertainty on the histograms in Fig. 5, the parameter of interest is fitted as $\mu = 1.0^{+0.45}_{-0.44}$ with an uncertainty worse by 19% than the best possible result obtained above, measured with the width of the total error bars.

As a consistency check for our new strategy described in section "Methods", we train the NN on the variance of the estimate for μ given by V_{00} in Eq. 6 but without adding the nuisance parameter η modelling the systematic uncertainty. The resulting NN function $\boldsymbol{f}$ in the input space, the distribution of the outputs and the profile of the likelihood are shown in Fig. 6. As expected, the plane of the function $\boldsymbol{f}$ in the input space is qualitatively similar, resulting with $\mu = 1.0^{+0.47}_{-0.46}$ in a comparable performance than the training on the cross entropy loss. It should be noted that the systematic uncertainty has been included again for the statistical inference.

When adding the nuisance parameter η to the likelihood, the training of the NN results in the function $\boldsymbol{f}$ shown in Fig. 7. The uncertainty of the parameter of interest is with the fit result $\mu = 1.0^{+0.39}_{-0.36}$ considerably decreased and lowers the residual difference to the optimal result from 19% to 4%.

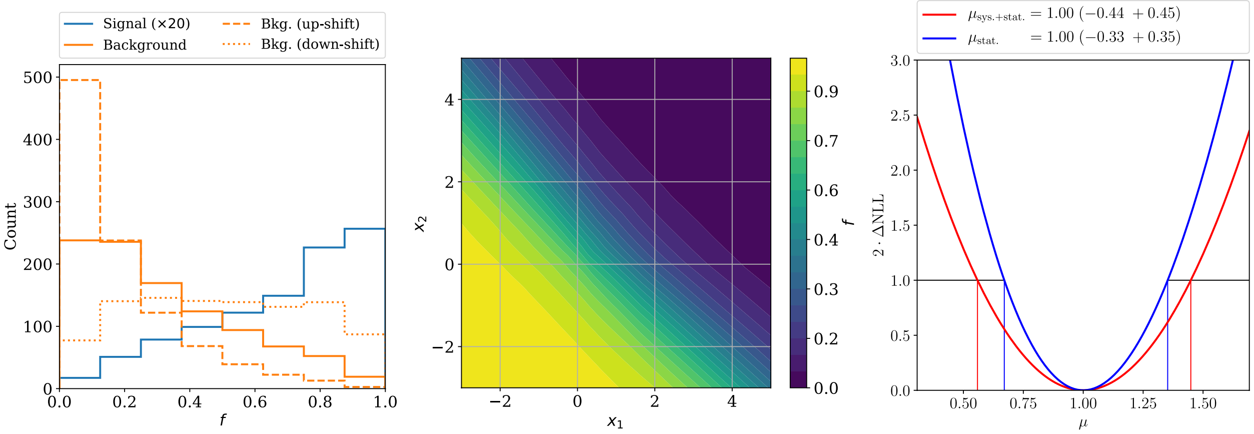

Fig. 5 Distributions of the NN output for the simple example consisting of signal, background, and systematic variation in the (left) input space spanned by x_1 and x_2 and (middle) value space, if the NN is trained on the classification of the two processes using the cross entropy loss. The likelihood profiles taking (red line) only the statistical uncertainty and (blue line) the statistical and systematic uncertainty into account for the final statistical inference of μ are shown on the right

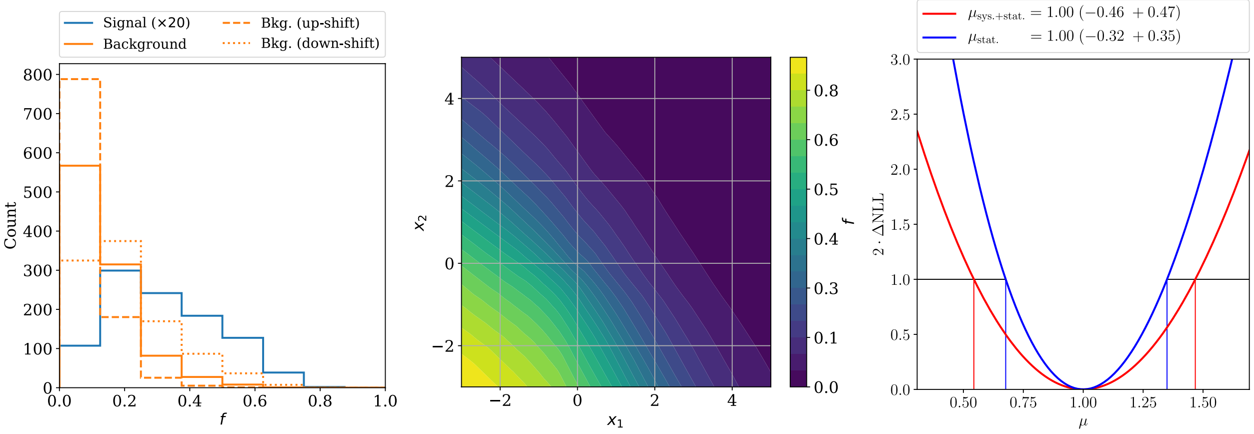

Fig. 6 Distributions of the NN output for the simple example consisting of signal, background, and systematic variation in the (left) input space spanned by x_1 and x_2 and (middle) value space, if the NN is trained on the variance of the signal strength V_{00} defined by the likelihood without the description of the systematic uncertainty. The likelihood profiles taking (red line) only the statistical uncertainty and (blue line) the statistical and systematic uncertainty into account for the final statistical inference of μ are shown on the right

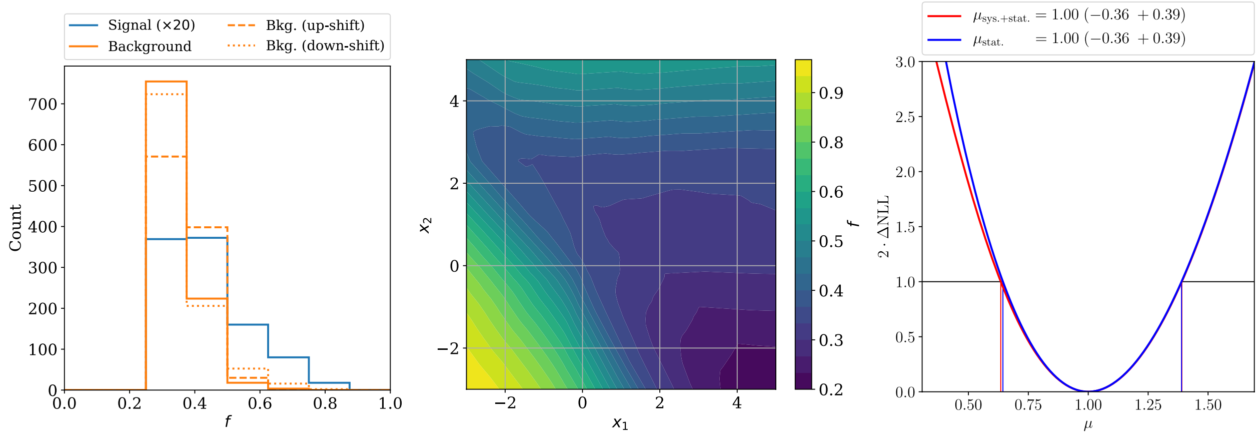

Fig. 7 Distributions of the NN output for the simple example consisting of signal, background, and systematic variation in the (left) input space spanned by x_1 and x_2 and (middle) value space, if the NN is trained on the variance of the signal strength V_{00} defined by the likelihood including the systematic uncertainty. The likelihood profiles taking (red line) only the statistical uncertainty and (blue line) the statistical and systematic uncertainty into account for the final statistical inference of μ are shown on the right

The function f in the input space in Fig. 7 shows that the training identified successfully the signal-enriched region with less contribution of the systematic uncertainty resulting in counts in the histogram yielding high signal statistics with a small uncertainty from the variation of the background process. Figure 7 shows also that the NN function is decorrelated against the systematic uncertainty because the profile of the likelihood changes only little if we remove the systematic uncertainty from the statistical model. The proposed method shares this feature with other approaches for decorrelation of the NN function such as discussed in section "Related work". The difference is that the strength of the decorrelation is not a hyperparameter but controlled by the higher objective V_{00}, which enables us to find directly the best trade-off between statistical and systematic uncertainty contributing to the estimate of μ. The correlation of the parameter of interest μ to the parameter η controlling the systematic variation is reduced from 64% for the training on the cross entropy loss to 13% for the training on the variance of the parameter of interest V_{00}.

Application to a More Complex Analysis Task Typical for High-Energy Particle Physics

In this section, we apply the proposed method to a problem typical for data analysis in high-energy particle physics at the LHC. We use a subset of the dataset published for the Higgs boson machine learning challenge [31, 32] extended by a systematic variation. The goal of the challenge is to achieve the best possible significance for the signal process representing Higgs boson decays to two tau leptons overlaid by the background simulated as a mixture of different physical processes [31]. We pick from the dataset four variables, namely `PRI_met`, `DER_mass_vis`, `DER_pt_h` and `DER_deltaeta_jet_jet` and select only events, which have all of these features defined. In addition to the event weights provided with the dataset, we scale the signal expectation with a factor of two. The final dataset has 244.0 and 35140.1 (106505 and 131480) weighted (unweighted) events for the signal and background process, respectively. The systematic uncertainty in the dataset is assumed as a 10% uncertainty on the missing transverse energy implemented with the transformation $\texttt{PRI\_met} \cdot (1.0 \pm 0.1)$ and propagated to the other variables using reweighing. The distributions of the variables including the systematic variations are shown in Figs. 8, 9, 10. The NN is trained only on three of the four variables, excluding the missing transverse energy. The systematic variations propagated to the remaining variables are thus correlated via a hidden variable, representing a more complex scenario than the simple example in section "Application to a simple example based onpseudo-experiments". We split the dataset using one third for training and validation of the NN, and two thirds for the results presented in this paper. The NN architecture and the training procedure are the same as implemented for the simple example in section "Application to a simple example based onpseudo-experiments" with the difference that we apply a standardization of the input ranges following the rule $(x - \bar{x})/\sigma(x)$ with the mean $\bar{x}$ and standard variation $\sigma(x)$ of the input x (Fig. 11).

An (asymptotically) optimal result as derived for the previous example is not available since the likelihood in the input space is not known. Instead we use the training on the cross entropy loss as reference with $\mu = 1.0^{+0.69}_{-0.68}$. Using V_{00} as training objective, but without the implementation of the systematic variations of the input distributions in the loss function, the result for the signal strength

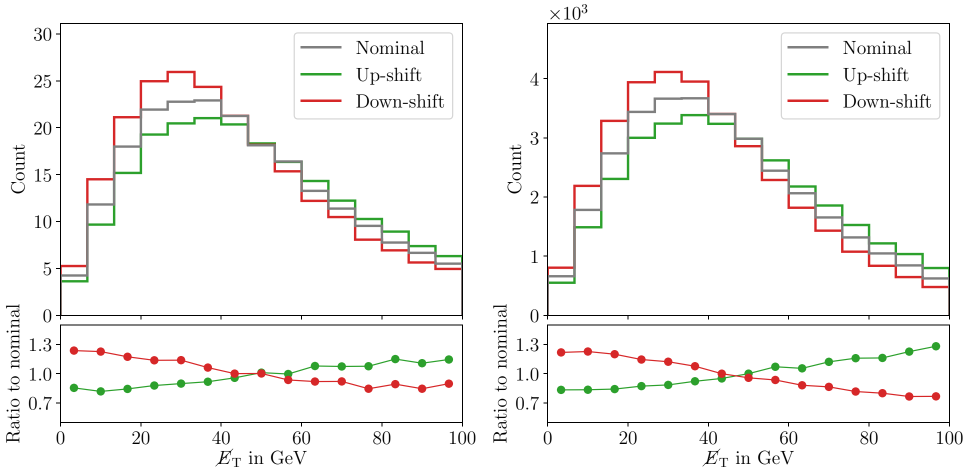

Fig. 8 Distribution of the missing transverse energy (PRI_met) for the (left) signal and (right) background process

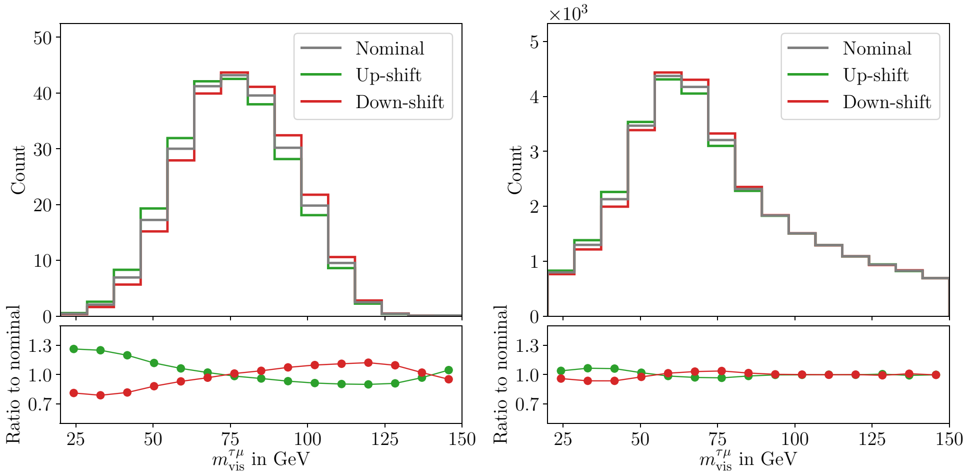

Fig. 9 Distribution of the visible mass of the di-tau system (DER_mass_vis) for the (left) signal and (right) background process

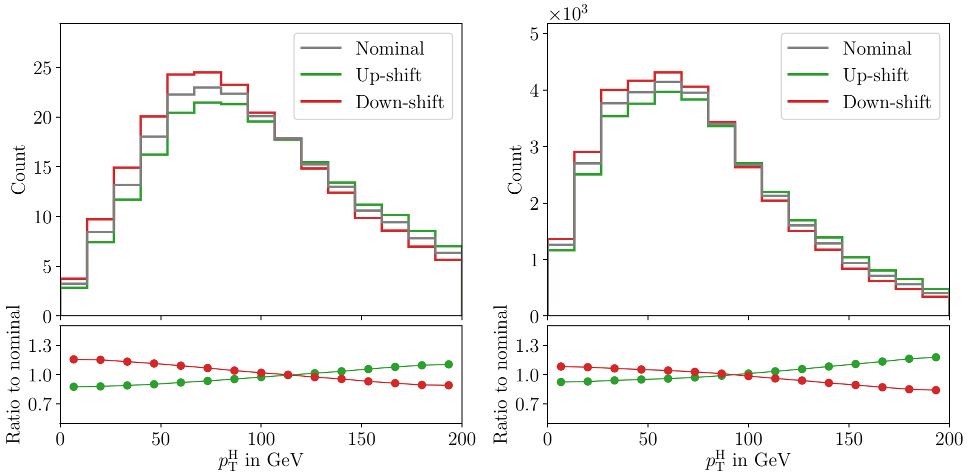

Fig. 10 Distribution of the transverse momentum built from the vector sum of the hadronic tau, the muon and the missing transverse momentum (DER_pt_h), used as an estimate of the transverse momentum of the reconstructed Higgs boson candidate, for the (left) signal and (right) background process

$\mu = 1.0^{+0.65}_{-0.64}$ shows a similar uncertainty compared to this reference. However, using the full likelihood from Eq. 3 as training objective, the signal strength is fitted with $\mu = 1.0^{+0.61}_{-0.60}$. The inclusion of the systematic variations yields an improvement in terms of the uncertainty on μ of 12% compared to the training on the cross entropy loss. The histograms and profiles of the likelihood used for extracting the results are shown in Figs. 12, 13, 14. For the assessment of the distributions of the NN output, it should

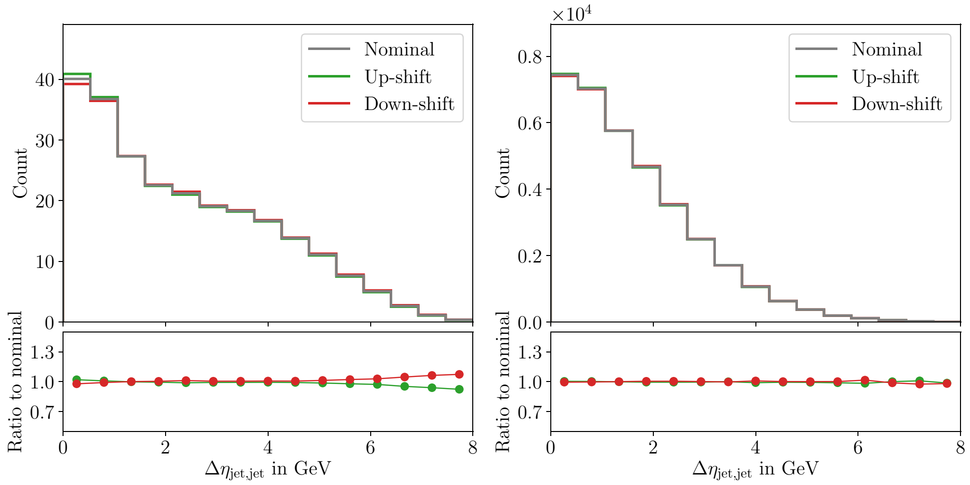

Fig. 11 Distribution of the absolute difference in the pseudorapidity of the two leading jets (`DER_deltaeta_jet_jet`) for the (left) signal and (right) background process

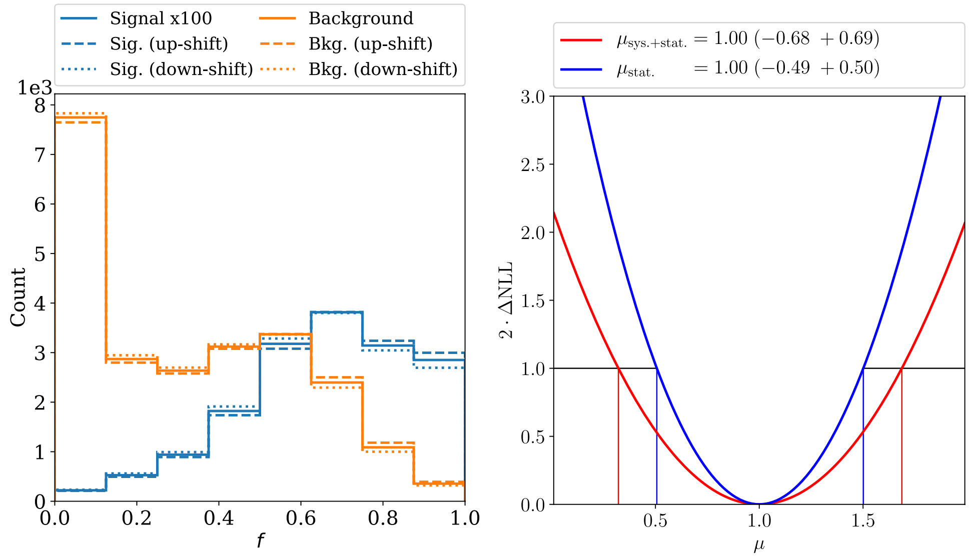

Fig. 12 Distribution of the NN output for the more complex example of section "Application to a more complex analysis task typical for high-energy particle physics", if the NN is trained on the classification of the two processes using the cross entropy loss. The likelihood profiles taking (red line) only the statistical uncertainty and (blue line) the statistical and systematic uncertainty into account for the final statistical inference of μ are shown on the right

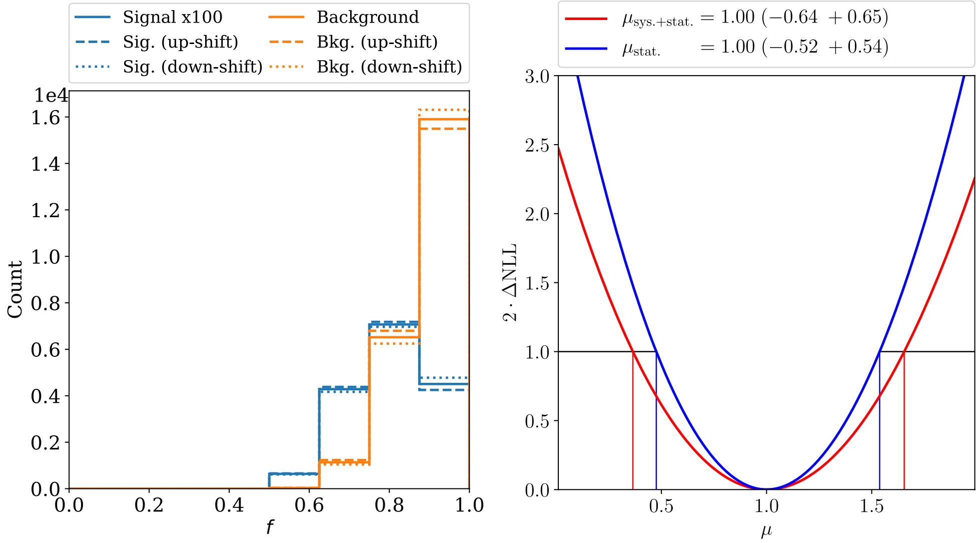

Fig. 13 Shown on the left is the distribution of the NN output in the Higgs example for signal, background and the systematic variation if the NN is trained on the variance of the signal strength V_{00} defined by the likelihood without the description of the systematic uncertainty. The likelihood profiles taking (red line) only the statistical uncertainty and (blue line) the statistical and systematic uncertainty into account for the final statistical inference of μ are shown on the right

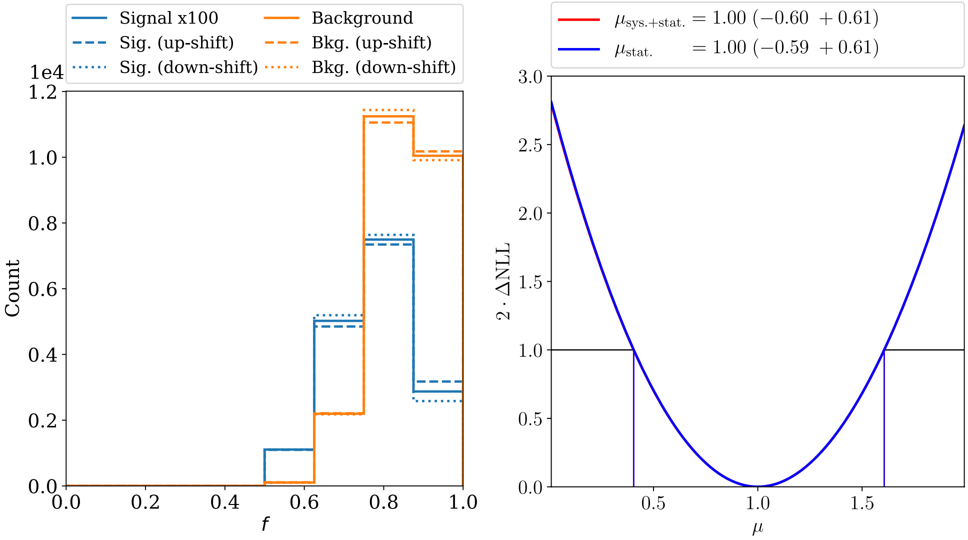

Fig. 14 Shown on the left is the distribution of the NN output in the Higgs example for signal, background and the systematic variation if the NN is trained on the variance of the signal strength V_{00} defined by the likelihood including the systematic uncertainty. The likelihood profiles taking (red line) only the statistical uncertainty and (blue line) the statistical and systematic uncertainty into account for the final statistical inference of μ are shown on the right

be noted that in contrast to the training based on the cross entropy loss, for the training based on V_{00} no preference is given for signal (background) events to obtain values close to 1 (0). Similar to the result from the simple example in section "Application to a simple example based on pseudo-experiments", the profiles of the likelihood for all scenarios show that the training on V_{00} removes the dependence on the systematic uncertainty yielding a smaller variance on μ. On the other hand, the training on the cross entropy optimizes best the estimate of μ in the absence of systematic uncertainties, as expected from our previous discussion. With the proposed strategy, the NN function learns to decorrelate against the systematic uncertainty, visible in the correlation of the signal strength μ to the parameter η controlling the systematic variation, which drops from 69% for the training on the cross entropy to 4% for the training on the variance of the parameter of interest V_{00}, based on the full likelihood information as given in Eq. 3.

To improve the estimate of μ for the approach with the NN trained on the cross entropy loss, a possible strategy could be to increase the number of histogram bins to exploit better the separation between the signal and background process. Figure 15 shows the development of the performance with the number of bins for the training on the cross entropy loss and the training on the likelihood via V_{00}. The training on the cross entropy loss results in an estimate of μ with a mean correlation to the nuisance parameter η of 66% and a falling uncertainty in μ with an average distance of 0.18 between the result for taking only the statistical uncertainties and statistical and systematic uncertainties into account for

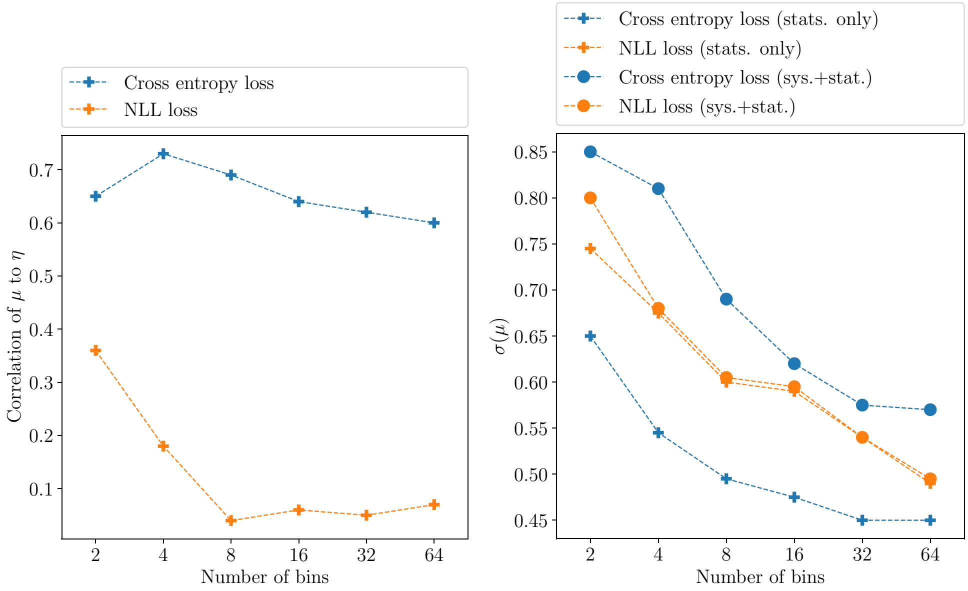

Fig. 15 Development of the (left) correlation between η and μ and (right) the variance of μ ($\sigma(\mu)$) with the number of histogram bins for the training based on the cross entropy loss or V_{00}

the statistical inference of μ. In contrast, the strategy with the NN trained on V_{00} shows a reduction of the correlation between μ and η of 0.35 when moving from two to eight bins for the input histogram for the statistical inference. The estimate remains robust against the systematic variation for all tested configurations, yielding a smaller variance for the estimate of μ compared to the training on the cross entropy loss. The average distance between the inference using only the statistical part of the likelihood and the full statistical model is 0.01. Including the systematic uncertainty in the inference, the comparison of the estimate of μ between the training based on V_{00} and the training based on the cross entropy shows an improved variance of μ by 0.07 on average, yielding a stable average improvement of 10%.

It should be noted that in practice the granularity of the binning is limited by the statistics of data and the simulation. Limited statistical precision in the simulation is usually taken into account by introducing dedicated systematic uncertainties in the statistical model that typically degrade the performance of the analysis for a large number of bins.

Summary

We have presented a novel approach to optimize statistical inference in the presence of systematic uncertainties, when using dimensionality reduction of the dataset and likelihoods based on Poisson statistics. Neural networks and in particular the differential approximation for the gradient of a histogram enables us to optimize directly the variance of the estimate of the parameters of interest in consideration of the nuisance parameters representing the systematic uncertainties of the measurement. The proposed method yields an improved performance for data analysis influenced by systematic uncertainties in comparison to conventional strategies using classification-based objectives for the dimensionality reduction. The improvements are discussed using a simple example based on pseudo-experiments with a known likelihood in the input space and we show that the technique is able to perform a statistical inference close to optimal by leveraging the given information about the systematic uncertainties. The applicability of the method for more complex analyses is demonstrated with an example typical for data analyses in high-energy particle physics. Future fields of studies are the application of the proposed method on analyses with many parameters in the statistical model and the evaluation of other possible differential approximations for the gradient of a histogram.

Acknowledgements We thank Lorenzo Moneta and Andrew Gilbert for helpful discussions and feedback, which greatly improved the manuscript.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Cowan G, Cranmer K, Gross E, Vitells O (2011) Asymptotic formulae for likelihood-based tests of new physics. Eur Phys J C 71(2):1554
2. The ATLAS and CMS collaborations (2011) Procedure for the LHC Higgs boson search combination in summer 2011. Technical report, ATL-PHYS-PUB-2011-011, CMS NOTE 2011/005
3. The CMS collaboration (2012) Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. Phys Lett B 716(1):30
4. The ATLAS collaboration (2012) Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. Phys Lett B 716(1):1
5. Wunsch S, Jörger S, Wolf R, Quast G (2020) Reducing the dependence of the neural network function to systematic uncertainties in the input space. Comput Softw Big Sci, 4(1)
6. Conway JS (2011) Incorporating nuisance parameters in likelihoods for multisource spectra. arXiv preprint: arXiv:1103.0354
7. Fisher RA (1925) Theory of statistical estimation. Math Proc Cambridge Philos Soc 22(5):700–725
8. Cramér H (1999) Mathematical methods of statistics, vol 9. Princeton University Press, Princeton
9. Rao CR (1992) Information and the accuracy attainable in the estimation of statistical parameters. In: Breakthroughs in statistics. Springer, pp 235–247
10. De Castro P, Dorigo T (2019) INFERNO: Inference-Aware Neural Optimisation. Comput Phys Commun 244:170–179
11. Heinrich L, Simpson N (2020) pyhf/neos: initial zenodo release. Zenodo
12. Elwood A, Krücker D (2018) Direct optimisation of the discovery significance when training neural networks to search for new physics in particle colliders. arXiv preprint: arXiv:1806.00322
13. Charnock T, Lavaux G, Wandelt BD (2018) Automatic physical inference with information maximizing neural networks. Phys Rev D 97(8):083004
14. Louppe G et al (2017) Learning to pivot with adversarial networks. In: Advances in Neural Information Processing Systems. 982
15. Shimmin C, Sadowski P, Baldi P, Weik E, Whiteson D, Goul E, Søgaard A (2017) Decorrelated jet substructure tagging using adversarial neural networks. Phys Rev D 96(7):074034
16. Estrade V, Germain C, Guyon I, Rousseau D (2018) Systematics aware learning: a case study in High Energy Physics. In: ESANN

2018 - 26th European Symposium on Artificial Neural Networks, Bruges, Belgium

17. Kasieczka G, Shih D (2020) Disco fever: Robust networks through distance correlation. arXiv preprint: arXiv:2001.05310
18. Cranmer K, Brehmer J, Louppe G (2019) The frontier of simulation-based inference. arXiv preprint: arXiv:1911.01429
19. Cranmer K, Pavez J, Louppe G (2015) Approximating likelihood ratios with calibrated discriminative classifiers. arXiv preprint: arXiv:1506.02169
20. Glorot X, Bengio Y (2010) Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp 249–256
21. Glorot X et al (2011) Deep sparse rectifier neural networks. In: Proceedings of the fourteenth international conference on artificial intelligence and statistics. pp 315–323
22. Abadi M, Agarwal A, Barham P et al (2016) Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint: arXiv:1603.04467
23. Dillon JV, Langmore I, Tran D, Brevdo E, Vasudevan S, Moore D, Patton B, Alemi A, Hoffman MD, Saurous RA (2017) Tensorflow distributions. CoRR arXiv:abs/1711.10604
24. Kingma D, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint: arXiv:1412.6980
25. James F (2006) Statistical methods in experimental physics. World Scientific Publishing Company, Singapore
26. Antcheva I, Ballintijn M, Bellenot B et al (2009) ROOT - A C++ framework for petabyte data storage, statistical analysis and visualization. Comput Phys Commun 180(12):2499–2512
27. Moneta L, Belasco K, Cranmer KS, Kreiss S, Lazzaro A, Piparo D, Schott G, Verkerke W, Wolf M (2010) The RooStats project. In: 13th International Workshop on Advanced Computing and Analysis Techniques in Physics Research (ACAT2010), SISSA, PoS(ACAT2010)057
28. Verkerke W, Kirkby D (2003) The RooFit toolkit for data modeling
29. The CMS collaboration (2019) Search for $\mathrm{t\bar{t}H}$ production in the $\mathrm{H} \rightarrow \mathrm{b\bar{b}}$ decay channel with leptonic $\mathrm{t\bar{t}}$ decays in proton-proton collisions at $\sqrt{s} = 13$ tev. JHEP 03:026
30. The CMS collaboration (2019) Measurement of Higgs boson production and decay to the $\tau\tau$ final state. CERN
31. Adam-Bourdarios C, Cowan G, Germain C, Guyon I, Kégl B, Rousseau D (2014) The Higgs boson machine learning challenge. In: HEPML@NIPS. 9–55
32. The ATLAS collaboration (2014) Dataset from the ATLAS Higgs Boson Machine Learning Challenge 2014. CERN Open Data Portal

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.