Reducing the impact of systematic uncertainties with inference-aware summary statistics

Pablo de Castro and Tommaso Dorigo

INFN - Sezione di Padova, Via Marzolo 8, 35131 Padova - Italy

E-mail: pablo.de.castro@cern.ch , tommaso.dorigo@cern.ch

Abstract. Complex computer simulations are commonly required for accurate data modelling in many scientific disciplines, including experimental high-energy physics, making statistical inference challenging due to the intractability of the likelihood evaluation for the observed data. Furthermore, sometimes one is interested on inference drawn over a subset of the generative model parameters while taking into account model uncertainty or misspecification on the remaining nuisance parameters. In this work, we show how non-linear summary statistics can be constructed by minimising inference-motivated losses via stochastic gradient descent such that they provide the smallest uncertainty for the parameters of interest. As a use case, the problem of confidence interval estimation for the mixture coefficient in a multi-dimensional two-component mixture model (i.e. signal vs background) is considered, where the proposed technique outperforms summary statistics based on probabilistic classification, a commonly used alternative which does not account for the presence of nuisance parameters in the optimisation.

1. Introduction

Simulator-based inference is currently at the core of many scientific fields, such as population genetics, epidemiology, and experimental particle physics. In these situations the generative procedure implicitly defined in the simulation often lacks a tractable probability density $p(\boldsymbol{x}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the vector of model parameters. Given n experimental observations $D = \{\boldsymbol{x}_0, ..., \boldsymbol{x}_n\}$, a problem of special relevance for these disciplines is statistical inference on a subset of model parameters $\boldsymbol{\omega}$. This can be approached via likelihood-free inference algorithms such as Approximate Bayesian Computation (ABC) [1], simplified synthetic likelihoods [2] or density estimation-by-comparison approaches [3].

Because the relation between the parameters of the model and the data is only available via forward simulation, most likelihood-free inference algorithms tend to be computationally expensive due to the need of repeated simulations to cover the parameter space. When data are high-dimensional, likelihood-free inference can rapidly become inefficient, so low-dimensional summary statistics $\boldsymbol{t}(D)$ are used instead of the raw data for tractability. The choice of summary statistic for such cases becomes critical, given that naive choices might cause loss of relevant information and a corresponding degradation of the power of resulting statistical inference.

In many cases, such as particle collisions at the Large Hadron Collider (LHC), the nature of the generative model (i.e. a mixture of different processes) allows the treatment of the problem as signal (s) vs background (b) classification [4], when the task becomes the one of effectively estimating an approximation of $t_B(\boldsymbol{x}) = p_s(\boldsymbol{x})/(p_s(\boldsymbol{x}) + p_b(\boldsymbol{x}))$ by means of probabilistic classification. While the use of classifiers to learn a summary statistic can increase

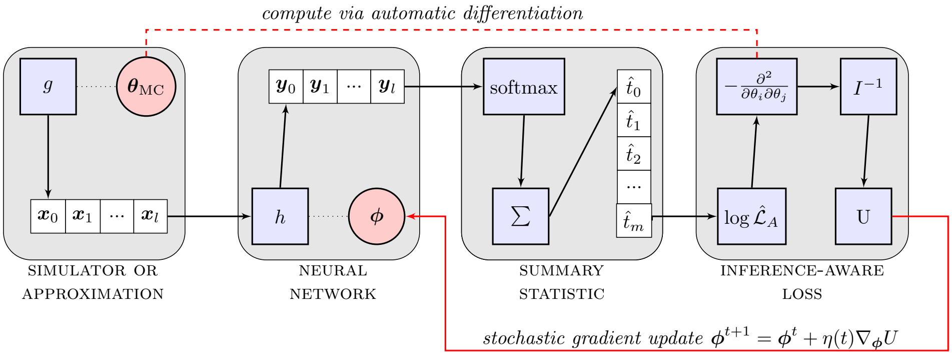

compute via automatic differentiation
g
$\boldsymbol{\theta}_{\mathrm{MC}}$
$\boldsymbol{x}_0$ $\boldsymbol{x}_1$ $\cdots$ $\boldsymbol{x}_l$
SIMULATOR OR APPROXIMATION
$\boldsymbol{y}_0$ $\boldsymbol{y}_1$ $\cdots$ $\boldsymbol{y}_l$
h
$\boldsymbol{\phi}$
NEURAL NETWORK
softmax
$\sum$
$\hat{t}_0$ $\hat{t}_1$ $\hat{t}_2$ $\cdots$ $\hat{t}_m$
SUMMARY STATISTIC
$-\frac{\partial^2}{\partial\theta_i\partial\theta_j}$
I^{-1}
$\log\hat{\mathcal{L}}_A$
U
INFERENCE-AWARE LOSS
stochastic gradient update $\boldsymbol{\phi}^{t+1} = \boldsymbol{\phi}^t + \eta(t)\nabla_{\boldsymbol{\phi}} U$

Figure 1: Learning inference-aware summary statistics (see text for details).

the discovery sensitivity, the simulations used to generate the samples which are needed to train the classifier often depend on additional uncertain parameters (commonly referred as nuisance parameters). These nuisance parameters are not of immediate interest but have to be accounted for in order to make quantitative statements about the model parameters based on the available data. Classification-based summary statistics cannot easily account for those effects, so their inference power is degraded when nuisance parameters are finally taken into account.

In this work, a new machine learning method referred to as Inference-Aware Neural Optimisation (INFERNO) [5] is presented, that constructs non-linear summary statistics directly optimising the expected amount of information about the subset of parameters of interest, by taking into account the effect of nuisance parameters. The optimisation procedure is carried out iteratively by stochastic gradient descent (SGD) [6] using small subsets of available simulated data. The learned summary statistics can be used to perform robust and efficient classical or Bayesian inference from the observed data, so they can be readily applied in place of current classification-based or domain-motivated summaries in current scientific data analyses.

2. Method

In this section, a machine learning technique to learn non-linear sample summary statistics is described. The method seeks to minimise the expected variance of the parameters of interest obtained via a non-parametric simulation-based synthetic likelihood. A graphical description of the technique is depicted on Fig. 1. The parameters of a neural network are optimised by SGD within an automatic differentiation framework, where the considered loss function accounts for the details of the statistical model as well as the expected effect of nuisance parameters.

The family of summary statistics $\boldsymbol{t}(D)$ considered in this work is composed by a neural network model applied to each dataset observation $\boldsymbol{h}(\boldsymbol{x};\boldsymbol{\phi}) : \mathbb{R}^d \rightarrow \mathbb{R}^m$, whose parameters $\boldsymbol{\phi}$ will be learned during the training phase. The neural network $\boldsymbol{h}(\boldsymbol{x}_i;\boldsymbol{\phi})$ will reduce the dimensionality from the d-dimensional inputs to the m-dimensional outputs and will effectively define the summary statistic transformation. The next step is to map observation outputs to a dataset summary statistic, which will in turn be calibrated and optimised via a non-parametric likelihood $\mathcal{L}(D;\boldsymbol{\theta},\boldsymbol{\phi})$ created using a set of l simulated observations $G_{\mathrm{MC}} = \{\boldsymbol{x}_0, ..., \boldsymbol{x}_l\}$, generated at a certain instantiation of the simulator parameters $\boldsymbol{\theta}_{\mathrm{MC}}$.

In experimental high-energy physics, histograms of observation counts are the most commonly used non-parametric density estimator because the resulting likelihoods can be expressed as the product of Poisson factors. A naive sample summary statistic can be built from the output of the neural network by simply assigning each observation $\boldsymbol{x}$ to a bin corresponding to the cardinality of the maximum element of $\boldsymbol{h}(\boldsymbol{x};\boldsymbol{\phi})$, which can in turn be used to build the following likelihood, where the expectation for each bin is taken from the simulated sample G_{MC}:

$$\mathcal{L}(D;\boldsymbol{\theta},\boldsymbol{\phi}) = \prod_{i=0}^{m} \mathrm{Pois}\left(t_i(D;\boldsymbol{\phi}) \mid \left(\frac{n}{l}\right) t_i(G_{\mathrm{MC}};\boldsymbol{\phi})\right) \tag{1}$$

where $t_i(D;\boldsymbol{\phi})$ are the sum of observations for which the maximum is at the bin i and the n/l factor accounts for the different number of observations in the simulated samples. In the above construction, the chosen family of summary statistics is non-differentiable due to the *argmax* operator, so gradient-based updates for the parameters cannot be computed. To work around this problem, a differentiable approximation $\hat{\boldsymbol{t}}(D;\boldsymbol{\phi})$ is considered. This function is defined by means of a *softmax* operator $\hat{t}_i(D;\boldsymbol{\phi}) = \sum_{x\in D} e^{f_i(\boldsymbol{x};\boldsymbol{\phi})/\tau}\ \sum_{j=0}^{m} e^{f_j(\boldsymbol{x};\boldsymbol{\phi})/\tau}$, where the temperature hyper-parameter τ will regulate the softness of the operator. Similarly, let us denote by $\hat{\mathcal{L}}(D;\boldsymbol{\theta},\boldsymbol{\phi})$ the differentiable approximation of the non-parametric likelihood obtained by substituting $\boldsymbol{t}(D;\boldsymbol{\phi})$ with $\hat{\boldsymbol{t}}(D;\boldsymbol{\phi})$. Instead of using the observed data D, the value of $\hat{\mathcal{L}}$ may be computed when the observation for each bin is equal to its corresponding expectation based on the simulated sample G_{MC}, which is commonly denoted as the Asimov likelihood [7] $\hat{\mathcal{L}}_A$. By taking the negative logarithm and expanding in $\boldsymbol{\theta}$ around $\boldsymbol{\theta}_{\mathrm{MC}}$, we can obtain the Fisher information matrix [8] for the Asimov likelihood:

$$I(\boldsymbol{\theta})_{ij} = \mathbb{E}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\left(-\log\hat{\mathcal{L}}_A(\boldsymbol{\theta};\boldsymbol{\phi})\right)\right] \tag{2}$$

which can be computed via automatic differentiation if the simulation function $g(\boldsymbol{\theta}_{\mathrm{MC}})$ or an approximation of the effect of varying $\boldsymbol{\theta}$ over the simulated dataset G_{MC} are differentiable.

The inverse of the Fisher information can be used as an approximate estimator of the expected covariance matrix of the parameters $\boldsymbol{\theta}$ for an unbiased estimator. In Bayesian terminology, this approach is referred to as the Laplace approximation [9]. The loss function used for stochastic optimisation of the neural network parameters $\boldsymbol{\phi}$ can be any function of the inverse of the Fisher information matrix at $\boldsymbol{\theta}_{\mathrm{MC}}$, depending on the ultimate inference aim. The diagonal elements $I^{-1}_{ii}(\boldsymbol{\theta}_{\mathrm{MC}})$ correspond to the expected variance of each of the ϕ_i under the approximation mentioned before, so if the aim is efficient inference about one of the parameters $\omega_0 = \theta_k$ a candidate loss function is $U = I^{-1}_{kk}(\boldsymbol{\theta}_{\mathrm{MC}})$ which corresponds to the expected width of the confidence interval for ω_0 accounting also for the effect of the other nuisance parameters in $\boldsymbol{\theta}$.

3. Experiments

In this section, we first study the effectiveness of the inference-aware optimisation in a synthetic mixture problem where the likelihood is known. We then compare our results with those obtained by standard classification-based summary statistics. The code for reproducing the results is available online [10], using TensorFlow [11] and TensorFlow Probability [12, 13].

To demonstrate the usage of the proposed approach, a three-dimensional mixture example with two components $p(\boldsymbol{x}|\mu,r,\lambda) = (1-\mu)f_b(\boldsymbol{x}|r,\lambda) + \mu f_s(\boldsymbol{x})$ is considered. One component will be referred as background $f_b(\boldsymbol{x}|r,\lambda)$ and the other as signal $f_s(\boldsymbol{x})$; their probability density functions are taken to correspond respectively to:

$$f_b(\boldsymbol{x}|r,\lambda)=\mathcal{N}\left((x_0,x_1)\,\middle|\,(2+r,0),\begin{bmatrix}5 & 0\\ 0 & 9\end{bmatrix}\right)Exp(x_2|\lambda) \quad f_s(\boldsymbol{x})=\mathcal{N}\left((x_0,x_1)\,\middle|\,(1,1),\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\right)Exp(x_2|2) \tag{3}$$

so that (x_0,x_1) are distributed according to a multivariate normal distribution while x_2 follows an independent exponential distribution both for background and signal.

In this toy problem, we consider a case where the underlying model predicts that the total number of observations are Poisson distributed with a mean $s+b$, where s and b are the expected number of signal and background observations. Thus the following parametrisation will be more convenient for building sample-based likelihoods:

$$p(\boldsymbol{x}|s,r,\lambda,b) = \frac{b}{s+b} f_b(\boldsymbol{x}|r,\lambda) + \frac{s}{s+b} f_s(\boldsymbol{x}). \tag{4}$$

If the probability density is known, but the expectation for the number of observed events depends on the model parameters, the likelihood can be extended [14] with a Poisson count term as: $\mathcal{L}(s,r,\lambda,b) = \text{Pois}(n|s+b)\prod^n p(\boldsymbol{x}|s,r,\lambda,b)$ which will be used to provide an optimal inference baseline when benchmarking the different approaches. Another quantity of relevance is the conditional density ratio, which would correspond to the optimal classifier (in the Bayes risk sense) separating signal and background events in a balanced dataset (equal priors):

$$t_B(\boldsymbol{x}|r,\lambda) = \frac{f_s(\boldsymbol{x})}{f_s(\boldsymbol{x}) + f_b(\boldsymbol{x}|r,\lambda)}. \tag{5}$$

While the synthetic nature of this example allows one to rapidly generate training data on demand, a training dataset of 200,000 simulated observations has been considered, in order to study how the proposed method performs when training data is limited. Half of the simulated observations correspond to the signal component and half to the background component. The latter has been generated using $r = 0.0$ and $\lambda = 3.0$. A validation holdout from the training dataset of 200,000 observations is only used for computing relevant metrics during training and to control over-fitting. The final figures of merit that allow to compare different approaches are computed using a larger dataset of 1,000,000 observations.

The statistical model described above has up to four unknown parameters: the expected number of signal observations s, the background mean shift r, the background exponential rate in the third dimension λ, and the expected number of background observations. The effect of the expected number of signal and background observations s and b can be easily included in the computation graph by weighting the signal and background observations. Instead the effect of r and λ, both nuisance parameters that will define the background distribution, is more easily modelled as a transformation of the input data $\boldsymbol{x}$. In particular, r is a nuisance parameter that causes a shift on the background along the first dimension and the effect of λ can be modelled by multiplying x_2 by the ratio between the λ_0 used for generation and the one being modelled.

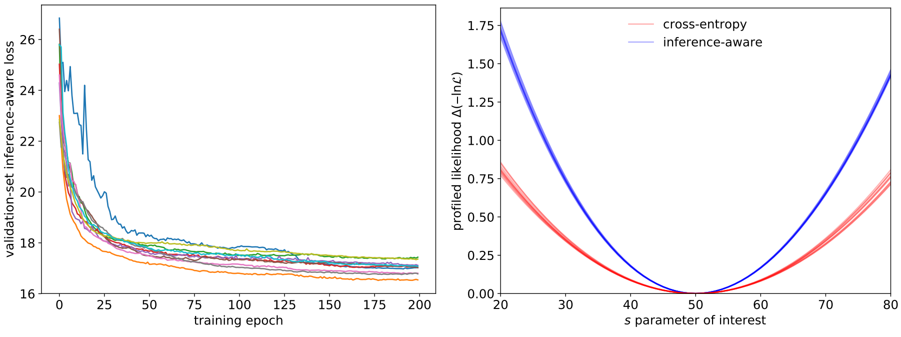
For this problem, we are interested in carrying out statistical inference on the parameter of interest s. The performance of inference-aware optimisation will be compared with classification-based summary statistics for a series of inference benchmarks based on the synthetic problem described above that vary in the number of nuisance parameters considered and their constraints, as shown in Table 1. For Benchmark 0 no nuisance parameters are considered, so the classification approach is expected to provide near optimal summary statistics. The rest of the benchmarks correspond to the presence of nuisance parameters, differing among them in their number and constrains. The main figure of merit will be the expected uncertainty in the parameter of interest s for the inference problem defined for each benchmark and conditioned on the true value of the parameters of $s = 50$, $r = 0.0$, $\lambda = 3.0$ and $b = 1000$.

A supervised machine learning model can be trained to discriminate signal and background, considering parameters r and λ fixed, as a way to obtain a variable transformation that is informative about the mixture fraction or s. The output of such a model are class probabilities c_b and c_s given an observation $\boldsymbol{x}$, where the latter will asymptotically tend to the optimal classifier from Eq. 5 given enough data and a flexible enough model. The classification output is a powerful learned feature that can be used as a summary statistic; but its construction ignores the effect of the nuisance parameters.

When using classification-based summary statistics, the construction of a summary statistic does depend on the presence of nuisance parameters, so the same model is trained independently

Table 1: Definition of the different statistical inference benchmark problems that will be considered when comparing different techniques to obtain summary statistics.

	Benchmark 0	Benchmark 1	Benchmark 2	Benchmark 3	Benchmark 4
interest pars	1 (s)	1 (s)	1 (s)	1 (s)	1 (s)
nuisance pars	0 (all fixed)	1 (r)	2 (r and λ)	2 (r and λ)	3 (r, λ and b)
r (bkg shift)	0.0 (fixed)	free (init 0.0)	free (init 0.0)	$\mathcal{N}(\lambda\|0.0, 0.4)$	$\mathcal{N}(\lambda\|0.0, 4.0)$
λ (bkg exp rate)	3.0 (fixed)	3.0 (fixed)	free (init 3.0)	$\mathcal{N}(\lambda\|3.0, 1.0)$	$\mathcal{N}(\lambda\|3.0, 1.0)$
b (bkg normalisation)	1000 (fixed)	1000 (fixed)	1000 (fixed)	1000 (fixed)	$\mathcal{N}(b\|1000, 100)$



(a) inference-aware training loss **(b)** profile-likelihood comparison

Figure 2: Dynamics and results of inference-aware optimisation: (a) square root of inference-loss (i.e. approximated uncertainty of the parameter of interest) as a function of the training step for 10 different random initialisations of the neural network parameters; (b) profiled likelihood around the expectation value for the parameter of interest s of 10 trained inference-aware models and 10 trained CE loss based models. All results correspond to Benchmark 2.

of the benchmark considered. For the approach presented in this work, inference-aware neural optimisation, the effect of the nuisance parameters and their constraints can be taken into account during training. Hence, 5 different training procedures for INFERNO will be considered, each one using as training loss function the final expected uncertainty of the actual inference problem for each of the benchmarks listed in Table 1. These models are denoted by the name INFERNO followed the corresponding benchmark number.

The same basic network architecture is used both for cross-entropy (CE) $L_{\text{CE}} = \sum_i y_i \log \hat{y}_i$ and inference-aware training: two hidden layers of 100 nodes followed by rectified linear unit (ReLU) [6] activations. The number of nodes on the output layer is two when classification proxies are used, while for inference-aware classification the number of output nodes corresponds to the dimensionality of the sample summary statistics. For the experiments shown in this work using the INFERNO technique, an output size of $m = 10$ has been used. The final layer is followed by a softmax activation function and a temperature $\tau = 0.1$ for inference-aware learning to ensure that the differentiable approximations are close to the true values.

In Fig. 2a, the training dynamics of inference-aware optimisation are shown by the validation loss, which corresponds to the approximate expected variance of parameter s, as a function of the step for 10 random-initialised instances of the INFERNO model corresponding to Benchmark 2. All inference-aware models were trained during 200 epochs with SGD using mini-batches of 2000 observations and a learning rate $\gamma = 10^{-6}$. All the model initialisations converge to summary statistics that provide low variance for the estimator of s.

To compare with alternative approaches, the profiled likelihood [15] obtained both with the

Table 2: Expected uncertainty on the parameter of interest s for each of the inference benchmarks considered using a CE trained model, a INFERNO model customised for each problem, the optimal classifier $t_B(\boldsymbol{x}|r = 0.0, \lambda = 3.0)$ and the analytical likelihood.

	Benchmark 0	Benchmark 1	Benchmark 2	Benchmark 3	Benchmark 4
NN classifier	$14.99^{+0.02}_{-0.00}$	$18.94^{+0.11}_{-0.05}$	$23.94^{+0.52}_{-0.17}$	$21.54^{+0.27}_{-0.05}$	$26.71^{+0.56}_{-0.11}$
INFERNO 0	$\mathbf{15.51^{+0.09}_{-0.02}}$	$18.34^{+5.17}_{-0.51}$	$23.24^{+6.54}_{-1.22}$	$21.38^{+3.15}_{-0.69}$	$26.38^{+7.63}_{-1.36}$
INFERNO 1	$15.80^{+0.14}_{-0.04}$	$\mathbf{16.79^{+0.17}_{-0.05}}$	$21.41^{+2.00}_{-0.53}$	$20.29^{+1.20}_{-0.39}$	$24.26^{+2.35}_{-0.71}$
INFERNO 2	$15.71^{+0.15}_{-0.04}$	$16.87^{+0.19}_{-0.06}$	$\mathbf{16.95^{+0.18}_{-0.04}}$	$16.88^{+0.17}_{-0.03}$	$18.67^{+0.25}_{-0.05}$
INFERNO 3	$15.70^{+0.21}_{-0.04}$	$16.91^{+0.20}_{-0.05}$	$16.97^{+0.21}_{-0.04}$	$\mathbf{16.89^{+0.18}_{-0.03}}$	$18.69^{+0.27}_{-0.04}$
INFERNO 4	$15.71^{+0.32}_{-0.06}$	$16.89^{+0.30}_{-0.07}$	$16.95^{+0.38}_{-0.05}$	$16.88^{+0.40}_{-0.05}$	$\mathbf{18.68^{+0.58}_{-0.07}}$
Optimal classifier	14.97	19.12	24.93	22.13	27.98
Analytical likelihood	14.71	15.52	15.65	15.62	16.89

INFERNO-based and classifier-based summary statistics, accounting for the effect of nuisance parameters defined in Benchmark 2, are shown in Fig. 2b. The expected uncertainty of the trained models are used for subsequent inference on the value of s can be estimated from the profile width when $\Delta\mathcal{L} = 0.5$. The average width for the profile likelihood with inference-aware, 16.97 ± 0.11, can be compared with the one obtained by uniformly binning the output of classification-based models in 10 bins, 24.01 ± 0.36. The models based on CE loss were trained during 200 epochs using a mini-batch size of 64 and a fixed learning rate of $\gamma = 0.001$.

A more complete study of the improvement provided by the INFERNO training procedure is provided in Table 2, where the median and 1-sigma percentiles on the expected absolute uncertainty on s are provided for 100 random-initialised instances of each model. In addition, results for 100 random-initialised CE neural network models trained as previously indicated, the optimal (Bayes) classifier $t_B(\boldsymbol{x}|r = 0.0, \lambda = 3.0)$ from Eq. 5, and the analytical likelihood-based inference are also included. The median expected uncertainties shown in Table 2, with the exception of the analytical likelihood which was based on the extended analytical likelihood, were obtained by constructing a binned likelihood for the resulting histograms of the two mixture components. The binned models for the inference problems with nuisance parameters were obtained by interpolating the histograms when the nuisance parameters are varied one standard deviation in each direction, which is a common practice in high-energy physics statistical inference.

Except for Benchmark 0, the confidence intervals obtained using INFERNO-based summary statistics are narrower than those using classification and tend to be much closer to those expected when using the true model likelihood for inference. Unsurprinsingly, the results for Benchmark 0, when no nuisance parameters are considered and thus the mixture components are perfectly known, show that classification-based summaries in this simplified setting can outperform the INFERNO technique.

The analytical likelihood, which amounts to use the true generative likelihood for inference, can be thought of an upper bound for the likelihood-free setting because it most effectively uses all the information of the data to constrain all the model parameters. The relative improvement over classification increases when more nuisance parameters are considered. As shown in Table 2, the constraining power of the summary statistic generated by INFERNO is stronger when it is constructed to solve the corresponding inference question, i.e. the training based on other benchmarks is sub-optimal. Thus the results also seem to suggest the inclusion of the detailed information about the inference problem in the INFERNO technique leads to comparable or better results than its omission.

4. Conclusions

In this work we have described a new approach for building non-linear summary statistics for likelihood-free inference that directly minimises the expected variance of the parameters of interest, which is considerably more effective than the use of classification surrogates when nuisance parameters are present. The application of INFERNO to non-synthetic examples, such as the systematic-extended Higgs dataset [16], are left for future studies.

Acknowledgments

Pablo de Castro would like to thank Daniel Whiteson, Peter Sadowski and the other members of the ML for HEP group at UCI for the initial feedback and support of the idea presented in this paper, as well as Edward Goul for his interest when the project was in early stages. The authors would also like to acknowledge Gilles Louppe and Joeri Hermans for some useful discussions directly related to this work.

This work is part of a more general effort to develop new statistical and machine learning techniques to use in high-energy Physics analyses within within the AMVA4NewPhysics project, which is supported by the European Union's Horizon 2020 research and innovation programme under Grant Agreement number 675440. CloudVeneto is also acknowledged for the use of computing and storage facilities provided.

References

[1] Beaumont M A, Zhang W and Balding D J 2002 *Genetics* **162** 2025–2035

[2] Wood S N 2010 *Nature* **466** 1102

[3] Cranmer K, Pavez J and Louppe G 2015 *arXiv:1506.02169*

[4] Adam-Bourdarios C, Cowan G, Germain C, Guyon I, Kgl B and Rousseau D 2015 *Proceedings of the NIPS 2014 Workshop on High-energy Physics and Machine Learning* pp 19–55 URL `http://proceedings.mlr.press/v42/cowa14.html`

[5] de Castro P and Dorigo T 2019 *Comput. Phys. Commun.* **244** 170–179

[6] Goodfellow I, Bengio Y and Courville A 2016 *Deep Learning* (MIT Press) `http://www.deeplearningbook.org`

[7] Cowan G, Cranmer K, Gross E and Vitells O 2011 *The European Physical Journal C* **71** 1554

[8] Fisher R A 1925 *Mathematical Proceedings of the Cambridge Philosophical Society* **22** 700725

[9] Laplace P S 1986 *Statistical Science* **1** 364–378

[10] de Castro P and Dorigo T 2018 Code and manuscript for the paper "INFERNO: Inference-Aware Neural Optimisation" `https://github.com/pablodecm/paper-inferno`

[11] Abadi M *et al.* 2015 TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems URL `https://www.tensorflow.org/`

[12] Tran D, Kucukelbir A, Dieng A B, Rudolph M, Liang D and Blei D M 2016 *arXiv:1610.09787*

[13] Dillon J V, Langmore I, Tran D, Brevdo E, Vasudevan S, Moore D, Patton B, Alemi A, Hoffman M and Saurous R A 2017 *arXiv:1711.10604*

[14] Barlow R 1990 *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **297** 496–506

[15] Tanabashi M, Hagiwara K, Hikasa K, Nakamura K, Sumino Y, Takahashi F, Tanaka J, Agashe K, Aielli G, Amsler C *et al.* 2018 *Physical Review D* **98** 030001

[16] Estrade V, Germain C, Guyon I and Rousseau D 2017 *Deep Learning for Physical Sciences workshop at NIPS*