Adversarial learning to eliminate systematic errors: a case study in High Energy Physics

Victor Estrade[1], **Cécile Germain**[1], **Isabelle Guyon**[1], **David Rousseau**[2]

1. LRI, TAU, UPSud. 2. IN2P3, LAL.
Université Paris Saclay, France

Abstract

Making the region selection procedure used in High Energy Physics analysis robust to systematic errors is a case of supervised domain adaptation. This paper proposes a benchmark that captures a simple but realistic case of systematic HEP analysis, in order to expose the issue to the wider community. The benchmark makes easy to conduct an experimental comparison of the recent adversarial knowledge-free approach and a less data-intensive alternative.

1 Introduction

An essential component of the analysis of the data produced by the experiments of the LHC (Large Hadron Collider) at CERN is a procedure for the selection of a region of interest in the space of measured features. Multivariate classification has become the standard tool to optimize the selection region. In the case of discovery and measurement of a new particle such as the Higgs boson, by definition no real labeled data are available.The classifier has to be trained on simulated data [2].

This introduces two kind of errors: *statistical* and *systematic*. When the data distribution is well-defined, that is the test and training are samples of iid random variables, the only source error is statistical, originating from from the finite size of the training data and the model capacity limitations.

Systematics are the "known unknowns" of the data distribution, in statistical parlance the *nuisance parameters* that coherently bias the training data, but which exact value is not known. A typical example is the uncertainty on the value of a physical quantity that parameterizes the simulation. Its effects cannot be followed through combination of errors algebra but must be estimated.

Assuming the nuisance parameters have first been optimally constrained, the goal is to optimize the tradeoff between statistical and systematic error. So far, the learning techniques exploited in High Energy Physics (HEP) analysis target the minimization of the statistical error only. Minimizing the systematic error is addressed a posteriori. Given the amount of work involved, optimizing the two errors within the same learning procedure would significantly streamline the analysis.

The goal of this paper is threefold:

- propose a realistic but easy to use benchmark including data and the figure of merit (sec. 2);
- position the systematics problem with respect to the relevant ML contexts, adversarial learning and domain adaptation (sec. 3);
- present an experimental performance evaluation (sec. 4).

2 A benchmark for systematics in HEP analysis

The benchmark. Our benchmark addresses the *measurement* problem, for instance of a cross section. Simulations for the various sources of systematics are abundant, but strictly private to the

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

HEP experiments, and anyway in HEP-specific (and cryptic) formats. In order to create a reasonably realistic playground that does not require any physics knowledge, we started with the Higgs challenge dataset [7], which is publicly available, well documented [2] and nearly identical to the official Atlas simulations used for the first evidence of Higgs boson decaying to τ lepton pairs.

We augmented it with an easy to use software[1] that calculates the impact of the Tau Energy Scale (TES) parameter on the simulation.

We consider a counting experiment in the signal region determined by the classifier. The labels are binary: *signal* (S) or *background* (B). As usual, the classifier learns a predictive model for the labels conditional on the observed values The quantities of interest are the weighted true and false positive counts where the w_i are the weights and t is the classification threshold:

$$s = \sum_{S, score_i > t} w_i \qquad \text{and} \qquad b = \sum_{B, score_i > t} w_i.$$

How nuisance parameters work. HEP experimental papers focusing on measurement of a quantity typically end with

$$\text{measurement} = m \pm \sigma_{\text{stat}} \pm \sigma_{\text{syst}},$$

where σ_{stat} and σ_{syst} are the statistical and systematic uncertainties.

The systematic uncertainty comes from different sources but in most cases it can be described as a Nuisance Parameter modifying one external input to the data. In general case, it impacts the features of both signal and background, as well as the weights. We focus here on the Tau Energy Scale which is a nuisance parameter affecting in a consistent way several features, both for signal and background. The Tau Energy Scale is a calibration uncertainty of the energy of the tau particle. It is a scaling gaussian uncertainty which value range between 1 to 3%. Several features depend of the tau energy in a non linear way, so that for one given event there is a 100% correlated uncertainty on several other features. However the magnitude of the uncertainty on the other variables vary from event to event.

The figure of merit. The measurement μ is the measured cross-section divided by the expected cross-section in a given model; it is proportional to the measured number of events in the final state. There is only one nuisance parameter, denoted by Z, which is 0 in the nominal case (Tau Energy Scale is 1). The figure of merit is a non-linear function of the true and false positives that derives from error propagation [1]. Let s_0 and b_0 be the number of true and false positives measured at nominal, and s_Z and b_Z their counterparts with systematics at Z. The figure of merit is the relative error $\sigma_\mu/\mu = \sqrt{\sigma_{\text{sta}}^2 + \sigma_{\text{sys}}^2}$, with

$$\sigma_{\text{sta}} = \frac{\sqrt{s_0 + b_0}}{s_0} \qquad \text{and} \qquad \sigma_{\text{sys}} = \frac{s_Z + b_Z - s_0 - b_0}{s_0} \tag{1}$$

Possible extensions. The Higgs Boson dataset of UCI [3] is analogous to the Higgs Challenge dataset. Due to the difference in the final state and to the simplifications of the UCi dataset, the features are not identical. However, `higgsml.py` could easily be adapted to scale the lepton energy scale, even if in reality it is known at least one order of magnitude better.

3 Adversarial learning and Systematics

Learning with systematics is related to domain adaptation [5], in the sense that the target data distribution is not accessible. However, classical domain adaptation addresses the semi-supervised setting [9], where for instance, data, but not or few labels are available for the target distribution. On the contrary, learning with systematics is fully supervised: with simulations, at training time we have all the labels, and even the values of the nuisance parameter if a way to use these can be figured out.

Data Augmentation. This naive approach simply trains on a mix of data generated in the adequate nuisance parameter range; with a sufficiently large training set, and sufficient classifier capacity, the

[1] available at https://github.com/victor-estrade/higgsdata

training algorithm should discover the invariant manifold in the data space. The systematic error in the physics sense reduces to the statistical error in the learning theory sense. Can we do better than that? The alternative is to steer the supervised learning procedure towards embedding domain adaptation into learning representations that do not contain domain-specific information.

Adversarial learning. The general GAN (Generative Adversarial Network) framework [10] is to go beyond the maximum likelihood paradigm, when the objective function itself must be learned from the data [15]. For instance, this is the case with adversarial privacy-preserving learning, where the relationship of features with the protected information cannot be posited a priori [8].

The Pivot Adversarial Network [11] exemplifies this approach for the HEP case. Informally, the Pivot Network is a GAN where the generated data is the distribution of the classification score, and the real-world data the distribution of the ground truth labels. It thus enjoys the same theoretical optimality results. The counterpart is to require large training sets to be representative of the full range of the data distribution, from nominal to perturbed.

Tangent Propagation. Another approach to learning invariant representations in the supervised case posits the impact of the systematics as coherent geometric transforms in the feature space. The systematics are considered being a differentiable transformation $f(x, Z)$ of the input. The model is explicitly regularized by the partial derivative of the classifier score wrt the nuisance parameter [14]: the smaller the derivative, the less sensitive the classifier. This approach departs from the pure adversarial setting, as the objective function is made explicit, although its optimization cannot be realized by classical gradient descent. It has been shown effective on top of unsupervised representation learning [13] [12]. In our case, it requires much less data than any adversarial setting.

4 Performance comparison

Experimental setup. We compared the two adversarial algorithms, Pivot (PAN) and Tangent Propagation (TP), a DNN with data augmentation method (DA), and a plain DNN (NN) as the baseline.

In order to make the comparison manageable, the dimensioning hyper parameters are identical for all DNN architectures, with 3 hidden layers of 120 neurons each, selected with grid search to minimize the statistical error. Similarly, the networks were all trained for 10000 iterations with a mini-batch size of 1024 and optimized with Adam method; the learning rate value is 0.001. Softplus and ReLU activations were tried and gave the same results.

The NN is trained on the nominal dataset only. Pivot and Data Augmentation are trained on a mix of skewed data, with Z drawn from a normal distribution $\mathcal{N}(0, k10^{-2})$, with k can be 1, 3 or 5. TP is trained on the nominal dataset with the tangent vector initialized by the finite difference method.

In all cases, the systematics are introduced in the test set with a fixed skewing, in order to reflect the real world situation where the nuisance parameter is well defined, although unknown.

Results. We report the errors and the overall figure of merit σ_μ/μ not only at the optimum, but along the decision threshold t in a a physically plausible region (expressed by the fraction of rejected events), in order to capture its behavior along the sensitivity/specificity trade-off.

The values are the mean and standard deviation of a 5-fold cross validation. This standard deviation is a rough estimate of the generalization error and should not be confused with σ_{sta} and σ_{sys}, which are the individual values). By lack of space, the results are shown for $Z = 3\%$ only, but they are consistent within the $[-5\%, 5\%]$ range.

Figure 1-left quantifies the impact of systematics on the decision threshold t, by considering the baseline NN, which is not systematics aware. Even with the crude threshold selection shown in figure 1-left, just moving the threshold from 86% to 97.5% slashes the overall error from 1.80 to 0.53; systematics-aware methods should improve on this result, with a much narrower potential of gain bounded below by $\min \sigma_{\text{sta}} \sim 0.2$.

Tangent propagation was decidedly unsuccessful, worse than the baseline (figure 1 - right). A more detailed analysis shows that enforcing the directional insensitivity is always detrimental: it increases not only the statistical error, as was expected, but the systematic error too. The class distributions are

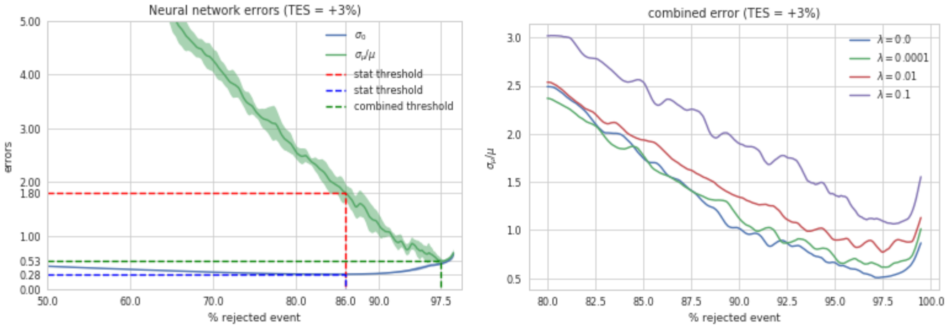

Figure 1: Left: Systematics-aware threshold selection. Right: Tangent Propagation and NN

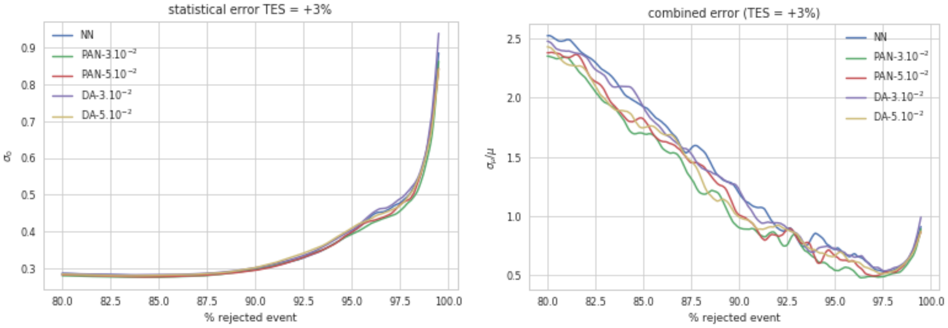

Figure 2: PAN and Data Augmentation

extremely overlapping. In [6] terms, the class manifolds are so extremely wrinkled that the geometric analogy is not effective.

Figure 2 compares PAN, DA and the baseline. The variance of training distribution (Z is drawn from a normal distribution $\mathcal{N}(0, k10^{-2})$), thus k is an hyper parameter of these algorithms. PAN seems to have a small advantage over DA. However, the confidence intervals (not shown for clarity) are too wide to conclude. In order to take into account this variability, we used a summary indicator: the area under the σ_μ/μ curve computed for each of the test cross-validation dataset. For each dataset, we get one vector of indicators, which defines a ranking of the methods. We check the significance of these rankings with the Wilcoxon rank sum test. At the 95% confidence level, the only positive result is the superiority of PAN-3 over plain NN; at 90% confidence level, the same is true for PAN-5. However, data augmentation schemes and pivot cannot be ranked.

5 Conclusion

Modeling uncertainties in HEP analysis is notoriously complicated: as Barlow [4] states: *There is a widespread myth that when errors are systematic, the standard techniques for errors do not apply, and practitioners follow prescriptions handed down from supervisor to student.*

This paper has presented a benchmark that captures a simple but realistic case of systematic HEP analysis, in order to expose the issue to the wider community. From the same dataset, and with a similar procedure, the robustness to different kind of systematics can be evaluated, like the jet energy resolution or a mismodelling of background composition

The problem consists of learning a representation that is insensitive to the perturbations induced by the nuisance parameters. The need for the adversarial techniques assuming a completely knowledge free approach has been questioned. The paper shows that the non separability of the classes invalidates this approach. The adversarial approach implemented in Pivot improves over a non systematic aware Neural Network. However, more work is needed to reinforce the statistical significance of this result.

References

[1] Georges Aad et al. Measurements of the higgs boson production and decay rates and constraints on its couplings from a combined atlas and cms analysis of the lhc pp collision data at $\sqrt{s} = 7$ and 8 tev. *JHEP*, 08:045, 2016.

[2] Claire Adam-Bourdarios, Glen Cowan, Cécile Germain, Isabelle Guyon, Balázs Kégl, and David Rousseau. The Higgs boson machine learning challenge. In *HEPML@ NIPS*, pages 19–55, 2014.

[3] P. Baldi, P. Sadowski, and D. Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nat Commun*, 5, 07 2014.

[4] R. Barlow. Systematic Errors: facts and fictions. *ArXiv High Energy Physics - Experiment e-prints*, 2002.

[5] Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Journal of Machine Learning*, (1-2):151–175, 2010.

[6] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, 2013.

[7] ATLAS collaboration. Dataset from the atlas higgs boson machine learning challenge 2014. `http://opendata.cern.ch/record/328`.

[8] Harrison Edwards and Amos Storkey. Censoring Representations with an Adversary. In *International Conference in Learning Representations (ICLR2016)*, 2016.

[9] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-Adversarial Training of Neural Networks. *arXiv:1505.07818 [cs, stat]*, May 2015. arXiv: 1505.07818.

[10] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Networks. *arXiv:1406.2661 [cs, stat]*, June 2014. arXiv: 1406.2661.

[11] Gilles Louppe, Michael Kagan, and Kyle Cranmer. Learning to Pivot with Adversarial Networks. *arXiv:1611.01046 [physics, stat]*, November 2016. arXiv: 1611.01046.

[12] Salah Rifai, Yann Dauphin, Pascal Vincent, Yoshua Bengio, and Xavier Muller. The Manifold Tangent Classifier. In *NIPS*, volume 271, page 523, 2011.

[13] Salah Rifai, Pascal Vincent, Xavier Muller, Xavier Glorot, and Yoshua Bengio. Contractive auto-encoders: Explicit invariance during feature extraction. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 833–840, 2011.

[14] Patrice Y. Simard, Bernard Victorri, Yann LeCun, and John S. Denker. Tangent Prop - A Formalism for Specifying Selected Invariances in an Adaptive Network. In John E. Moody, Stephen Jose Hanson, and Richard Lippmann, editors, *NIPS*, pages 895–903. Morgan Kaufmann, 1991.

[15] David Warde-Farley and Ian Goodfellow. Adversarial perturbations of deep neural networks. In Tamir Hazan, George Papandreou, and Daniel Tarlow, editors, *Perturbation, Optimization and Statistics*, pages 1–32. MIT Press, 2016.