# Robust deep learning: A case study

Victor Estrade, Cécile Germain, Isabelle Guyon, David Rousseau

# Robust deep learning: A case study

Victor Estrade, Cecile Germain, Isabelle Guyon, David Rousseau

Laboratoire de Recherche en Informatique

**Abstract.** We report on an experiment on robust classification. The literature proposes adversarial and generative learning, as well as feature construction with auto-encoders. In both cases, the context is domain-knowledge-free performance. As a consequence, the robustness quality relies on the representativity of the training dataset wrt the possible perturbations. When domain-specific a priori knowledge is available, as in our case, a specific flavor of DNN called Tangent Propagation is an effective and less data-intensive alternative.

**Keywords:** Domain adaptation, Deep Neural Networks, High Energy Physics

## 1 Motivation

This paper addresses the calibration of a classifier in presence of systematic errors, with an example in High Energy Physics.

An essential component of the analysis of the data produced by the experiments of the LHC (Large Hadron Collider) at CERN is the selection of a region of interest in the space of measured features associated to each collision, i.e., the variables for each particle collision or "event". Classifiers have 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 [1].

This introduces two kind of errors: *statistical* and *systematic*. When, as it is the case here, the data model is known, the statistical error essentially comes from the limited size of the training data. Coping with statistical error is at the core of classification theory and practice. Systematics are the "known unknowns" of the data model, in statistical parlance the *nuisance parameters* that coherently bias the training data, but which exact values are unknown. A typical example is the uncertainty on the value of a physical quantity that parameterizes the simulation.

Formally, for a family of classifiers parameterized by $\theta$ (e.g. the architecture and hyperparameters of a neural network), let $h(., \theta)$ be the score function of classifier $h$ and $Z$ be the nuisance parameter. Without specific action, $Z$ impacts the selection of $\theta$. This is exactly what we want to avoid: Robustness means that $h(., \theta)$ and $Z$ should be independent ($h$ should be *pivotal*). Of course, $h$ should also be a good classifier, (a constant classifier would be pivotal, but useless), which helps to situate robustness as a regularization objective.

## 2 Domain Adaptation

Learning with systematics fall under the theory of domain adaptation [4]. Implementations have to choose between two strategies: either a knowledge-free setting, where the invariances are discovered from the data; or the integration of prior knowledge. The knowledge-free adaptation can be supervised, with Generative Adversarial Networks [6],[7], or semi supervised with Domain Adversarial Networks [5]. It requires large training sets, representative of the nominal and perturbed data distributions. In the HEP context, the cost of precise simulations would be too high.

In the second case, the invariances describe the expected robustness properties typically as small geometric transforms in the feature space. The Tangent Propagation (TP) algorithm, proposed long ago [9] and recently revived [8], provides a principled method to integrate the invariance constraints into the learning of the data model with a classifier. With TP, the systematics are considered as a transformation $f(x, Z)$ of the input. The objective is to have $h(x, \theta) = h(f(x, Z), \theta)$, thus the model is regularized by : $\frac{\partial h(f(x,Z),\theta)}{\partial Z}$ i.e. the partial derivative of the classifier score wrt the nuisance parameter. As usual, a parameter noted $\lambda$ in the following, controls the tradeoff between the classification error and the regularization.

The derivative of the networks according to the nuisance parameter $Z$ of the transformation coming from the systematics. [9] implements this regularization by combining a classical Deep Neural Network (DNN) and a Jacobian network, to exploit the caracteristic of backpropagation as a differentiation engine. The regularization is computed by forward propagation of the tangent vectors $\frac{\partial f(x,Z)}{\partial Z}$ through the Jacobian network. As usual, a parameter controls the tradeoff between the classification error and the regularization.

Tangent Propagation makes it possible to integrate as many invariants as needed, each invariant requiring a forward and backward propagation of its tangent vector through the Jacobian network. If one does not have access to $\frac{\partial f(x,Z)}{\partial Z}$, these tangent vectors can be approximated with finite difference :

$$\frac{\partial f(x, Z)}{\partial Z} = \frac{f(x, 1 + \alpha) f(x, 1 - \alpha)}{2\alpha}$$

### 2.1 HEP systematics

The case we study here is measurement in HEP experiments . The objective is to estimate the number of Higgs particles produced in an HEP experiment. The proton-proton collisions create particles that are catch by the detector. From these particles it is possible to infer what happened during the collision, for example if a Higgs boson was created (then decayed in other particles) or not. A single collision is called an event and measurements on the detected particles are done to produce a vector of features to describe this event. The dataset is a simulation of a bunch of these feature vectors.

The objective is to measure the True Positives of a classifier trained to separate the signal events from the background events.

## 3 Experimental results

*The dataset* We use the dataset of the HiggsML challenge [2], `http://opendata.cern.ch/record/328?ln=en`. Data is split between training and test sets with 5-fold cross validation All training is performed at the nominal setting. The systematics are introduced in the **test set** only.

*Figure of merit* The figure of merit is not the classication accuracy, but a non-linear function of true and false positives related to error propagation in measurement [3]. Let $s_0$ and $b_0$ be the number of true and false postives at nominal, and $s_Z$ and $b_Z$ their counterparts with systematics at $Z$. The figure of merit is $\sigma_\mu = \sqrt{\Sigma_0^2 + \Sigma_Z^2}$, where $\Sigma_0 = \frac{\sqrt{s_0+b_0}}{s_0}$ is the statistical error and $\Sigma_Z = \frac{s_Z+b_Z-(s_0+b_0)}{s_0}$ the systematic error. Because this function is not additive in the examples, we use the regularized classification error as a proxy to train the classifier.

*Evaluation methodology* The baseline is a DNN without TP (or equivalently a TP-DNN with $\lambda = 0$). As TP constrains the architecture (softmax activations), we also include results for a standard (RELU-based) DNN. In order to make the comparison manageable, the dimensioning hyper parameters are identical for all architectures : 3 hidden layers of 40 neurons each. All networks were trained for 2000 iterations with a mini-batch size of 1024 and optimized with Adam method, and a learning rate of 0.01.

*Results* Figure 1a shows that TP consistently reduces the systematic error $\Sigma_Z$, by 20% on average near the minimum. The narrow confidence intervals support the significance of this result. For all architectures, $\Sigma_Z$ is very noisy. As a similar behavior is observed with gradient boosting, noisiness is probably intrinsic to the problem.

Figure 1b highlights the complex impact of TP on the statistical error $\Sigma_0$. The much wider confidence intervals with TP might be due to the limits of cross-validation. But, as the TP-NN is trained to ignore some variability, this might indicate that this variability crosses the class boundary, i.e. that the gap between the class manifolds is too small. Experimenting with the boostrap may help disentangling this two causes.

Figure 1c shows that TP has a net positive effect on $\sigma_\mu$. Other experiments (not reported here) show that this remains true for sensible ranges of $Z$ and $\lambda$. The impact of the slight increase of the statistical error is limited, as expect at very high threshold, the statistical error remains well below the systematic error, and will be even more negligible when added in quadrature.

## 4 Conclusion

The positive results of this preliminary work show that the tangent propagation approach can be effective to reduce the systematic error even in the extremely

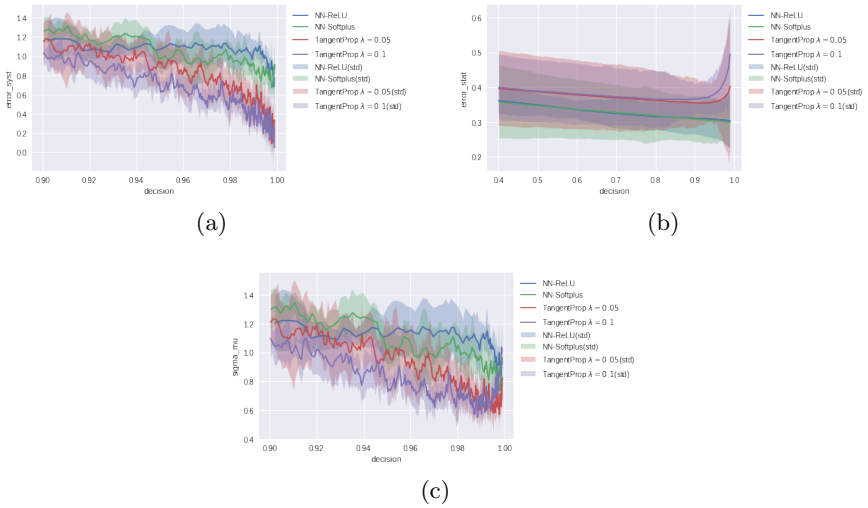


Fig. 1: Performance comparison for $Z = -1\%$. The values are the mean and standard deviation of the 5-fold cross validation. The decision threshold range corresponds to the constraints of physics analysis [1]

difficult HEP case. Further experiments comparing this methods with adversarial networks and ensemble methods are in progress. We will also refine the implementation with better hyper-parameter selection and explore the bootstrap. As systematics are pervasive in scientific measurements, we envision the creation of a *systematics challenge*, in the spirit of the AutoML challenge for future work.

## References


1. Claire Adam-Bourdarios, Glen Cowan, Ccile Germain, Isabelle Guyon, Balzs Kgl, and David Rousseau. The Higgs boson machine learning challenge. In *HEPML@ NIPS*, pages 19–55, 2014.
2. ATLAS Collaboration. Dataset from the ATLAS Higgs Boson Machine Learning Challenge 2014, 2014.
3. Roger Barlow. Systematic errors: Facts and fictions. In *Advanced Statistical Techniques in Particle Physics. Proceedings, Conference, Durham, UK, March 18-22, 2002*, pages 134–144, 2002.
4. Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Mach. Learn.*, 79(1-2):151–175, May 2010.
5. Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Franois Laviolette, Mario Marchand, and Victor Lempitsky. Domain-Adversarial Training of Neural Networks. *arXiv:1505.07818 [cs, stat]*, May 2015. arXiv: 1505.07818.

6. 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.
7. Gilles Louppe, Michael Kagan, and Kyle Cranmer. Learning to Pivot with Adversarial Networks. *arXiv:1611.01046 [physics, stat]*, November 2016. arXiv: 1611.01046.
8. Salah Rifai, Yann Dauphin, Pascal Vincent, Yoshua Bengio, and Xavier Muller. The Manifold Tangent Classifier. In *NIPS*, volume 271, page 523, 2011.
9. 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.