Graphical Abstract

Application of Neural Networks to the discovery of new physics

Óliver Partida,Pere Masjuan

Highlights

Application of Neural Networks to the discovery of new physics

Óliver Partida,Pere Masjuan

- Research highlights item 1
- Research highlights item 2
- Research highlights item 3

Application of Neural Networks to the discovery of new physics

Óliver Partida[a,c,*,1] (Researcher), Pere Masjuan[b,d]

[a] Elsevier B.V., Radarweg 29, 1043 NX Amsterdam, The Netherlands

ARTICLE INFO

Keywords:
standard model
B-physics
neural networks
conditional adversarial networks

ABSTRACT

This template helps you to create a properly formatted LaTeX manuscript.
`\beginabstract ... \endabstract` and `\begin{keyword} ... \end{keyword}` which contain the abstract and keywords respectively.
Each keyword shall be separated by a `\sep` command.

1. Introduction

Over the last few years, many observables related to the flavor-changing neutral-current transitions $b \to sl^+l^-$ have exhibited deviations from SM expectations. These anomalies can be classified in two sets: $b \to s\mu\mu$ related to observables testing only muonic transitions, called Lepton Flavor Dependent (LFD), and Lepton-Flavor Universality Violating (LFUV) anomalies that correspond to deviations in observables comparing muonic and electronic transitions. In 2013, using the 1 fb^{-1} dataset, the LHCb experiment measured the basis of optimized observables for $B \to K^*\mu\mu$, observing the so-called P_5' anomaly, i.e., a sizable 3.7σ discrepancy between the measurement and the SM prediction in one bin for the angular observable P_5'(Figure 1). A new discrepancy in the ratio $R_K^* = Br(B \to K^*\mu\mu)/Br(B \to K^*ee)$(Figure 2) was also observed by LHCb, hinting at the violation of Lepton Flavor Universality (LFU) and suggesting that deviations from the SM are predominantly present in $B \to K\mu^+\mu^-$ transitions but not in $B \to Ke^+e^-$ ones. In order to evaluate the significance and coherence of these deviations, a global model-independent fit is the most efficient tool to determine if they contain patterns explained by New Physics(NP). The starting point is an effective Hamiltonian in which heavy degrees of freedom (the top quark, the W and Z bosons, the Higgs and any heavy new particle) are integrated out in short-distance Wilson coefficients C_i, leaving only a set of operators $\mathcal{O}_i$ describing the physics at long distances:

$$\mathcal{H}_{eff} = -\frac{4G_F}{\sqrt{2}} V_{tb} V_{ts}^* \sum_i C_i \mathcal{O}_i.$$

This approach is valid if all NP degrees of freedom have masses well above the energy scale of the observables of interest, a well-motivated assumption due to the lack of observations in direct searches, both at the Large Hadron Collider (LHC) as well as in low-energy experiments. The observables measured by the experimental collaborations can be written in terms of the Wilson coefficients. Current analysis of anomalies in flavour physics are based on a linear regression of a χ^2 function. After taking into account correlation between theoretical predictions and experimental observables the χ^2 is built and minimized. In [1] a global fit is attempted to all available $b \to sl^+l^-$ data in a model-independent way allowing for different patterns of NP. They showed that the NP hypothesis is preferred over the SM by 5σ in a general case when NP can enter SM-like operators and their chirally-flipped partners and that LFU violation is favored with respect to LFU at the 3-4 σ level.

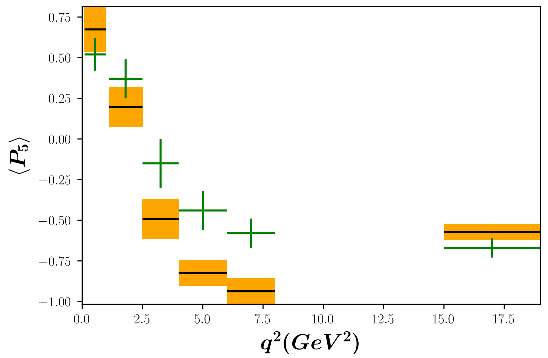

Figure 1: $P_5 = B \to K^*\mu\mu$. SM predictions(orange boxes) and experimental values(green crosses).

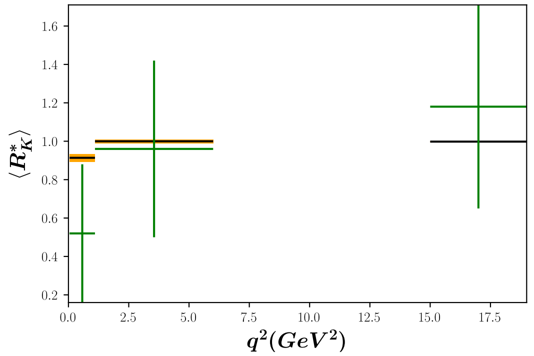

Figure 2: $R_K^* = Br(B \to K^*\mu\mu)/Br(B \to K^*ee)$. Ratios with different lepton at the final state, R_K ratios, are a clear indication of the Lepton Flavor Universality Violation something which is not expected from the theory.

cvr_1@tug.org.in (Ó. Partida)
www.cvr.cc, cvr@sayahna.org (Ó. Partida)
ORCID(s): 0000-0001-7511-2910 (Ó. Partida)

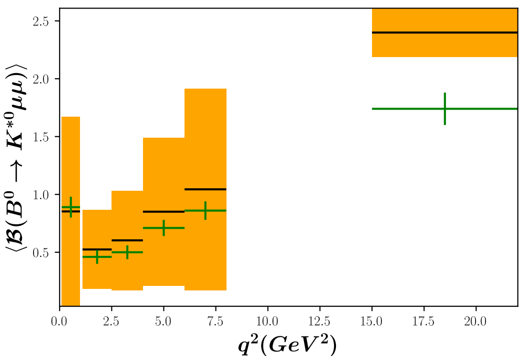

Figure 3: $\mathcal{B}(B^0 \to K^{*0}\mu\mu)$. Branching ratios.

2. Machine learning techniques: Neural Networks And Generative Adversarial Networks

2.1. Generative vs. Discriminative models

Generative and discriminative models are the two main types of models used to solve machine learning classification problems. In classification tasks we are usually interested in learning the conditional distribution $p(C_k|\boldsymbol{x})$, where x is the observation and C_k is the corresponding class assigned to this observation, and then use this conditional distribution to make class assignments. Generative models tackle this problem by first determining the class-conditional distributions $p(\boldsymbol{x}|C_k)$ and the prior class probabilities $p(C_k)$ and the using Bayes' theorem

$$p(C_k|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|C_k)p(C_k)}{p(\boldsymbol{x})}$$

to find the posterior class probabilities $p(C_k|\boldsymbol{x})$. It is also possible to model the joint distribution $p(\boldsymbol{x}, C_k)$ directly and normalize to obtain the posterior probabilities. Data from the input space can be generated by sampling from the learned model.

Discriminative classifiers, on the other hand, model the posterior $p(y|x)$ directly or learn a direct map from inputs x to class labels. Discriminative models that use probability distributions to solve the classification problem are called probabilistic discriminative models. For example, in the case of two classes, the posterior probability

$$p(C_1|x) = \frac{p(\boldsymbol{x}|p(C_1)p(C_1)}{p(\boldsymbol{x}|p(C_1)p(C_1) + p(\boldsymbol{x}|p(C_2)p(C_2)}$$

can be rewritten as

$$\frac{1}{1 + exp(-a)} = \sigma(a),$$

where

$$a = ln\frac{p(\boldsymbol{x}|p(C_1)p(C_1)}{p(\boldsymbol{x}|p(C_2)p(C_2)}$$

and $\sigma(a)$ is the logistic sigmoid function. For some specific conditional distributions $p(\boldsymbol{x}|p(C_k)$ the argument of the sigmoid function is a linear function of the inputs $\boldsymbol{x}$

$$a = \boldsymbol{w}^{\intercal}\boldsymbol{x} + \boldsymbol{w_0}.$$

In discriminative modeling we can use maximum likelihood to directly find the parameters w.

2.2. Neural Networks

Usually linear models for regression and classification are based on linear combinations of fixed nonlinear basis functions

$$y(\boldsymbol{x}, \boldsymbol{w}) = f\left(\sum_{j=1}^{M} w_j \phi_j(\boldsymbol{x})\right),$$

where $f(.)$ is a nonlinear activation function in the case of classification and the identity in the case of regression. For example in polynomial regression when there is only one input variable the set of basis function $\phi_j(\boldsymbol{x})$ take the form:

$$[1, x, x^2, x^3...]$$

One limitation of polynomial basis functions is that they are global functions of the input variable so input space regions are not independent. This problem can be alleviated by fitting different polynomials to different regions of space leading to *spline functions*. Neural networks make basis functions depend on adjustable parameters that are *learned* along network coefficients w_j during training. In the basic neural network model with just one hidden layer the output of each basis function is the result of applying a nonlinear function $h(.)$ to a linear combination of the input variables $x_1, ..., x_D$:

$$a_j = \sum_{i=1}^{D} w_{ji} x_i x_{j0}$$

$$z_j = h(a_j).$$

These values are again linearly combined and transformed using an appropriate function to give the final output:

$$a_k = \sum_{i=1}^{D} w_{kj} z_j w_{k0}$$

$$y_k = g(a_k).$$

For regression problems g is the identity so $y_k = a_k$. For binary classification g is the sigmoid function:

$$g(a) = \frac{1}{1 + exp(-a).}$$

2.3. Conditional Generative Adversarial Networks

Generative Adversarial Networks (GANs) are classified within the group of generative models. They consists of two adversarial models: a generative model G that captures the data distribution, and a discriminative model D that estimates the probability that a sample came from the training

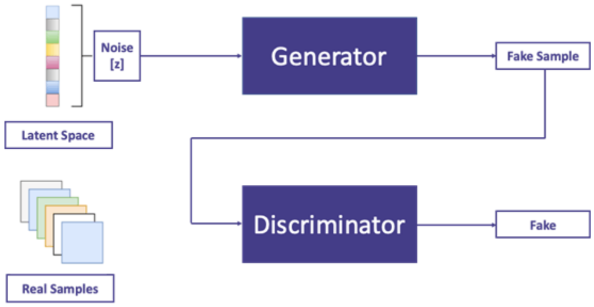

Figure 4: Generator and Discriminator as blocks of the GAN network.

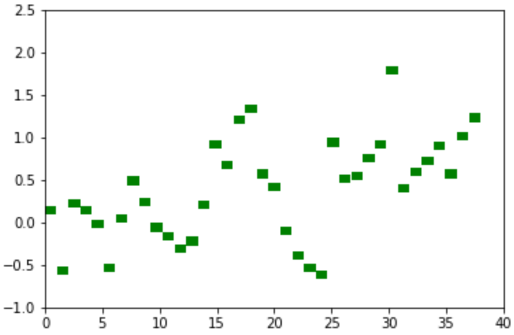

Figure 5: Experimental observables values.

data rather than G. The generator's G loss during training depends on the probability of the discriminative model of making an error by assigning the class *real* to *fake* data, that is, data generated by the generator. Other generative models such as Gaussian Mixture Models are mostly based on maximum likelihood estimate however maximum likelihood estimation may not represent the complexity of the actual data distribution and cannot learn the high-dimensional data distributions. The generator in a GAN model takes a fixed-length random vector $z \sim p(z)$ as input and generates a sample $x \sim P^*_{data}(x)$ in the domain where $P^*_{data}(x)$ is the true distribution.
Conditional Generative Adversarial Networks are an extension of GANs where a conditional setting is applied. The generator receives a new input along the random variable z which adds extra information about the sample to be generated. Similarly the discriminator is trained to classify samples that incorporate this new information.

3. Data preparation and implementation

In this work we have included 37 observables. Observable is referred to a measurement of either an angular observable, a branching ratio, or a ratio, in a particular energy bin. We randomly generated 2500 model coefficients pairs $(C9, C10)$ by generating 50 values for $C_9 \in [-2, 0]$ and 50 values for $C_{10} \in [0, 1]$ and forming the Cartesian product $(C9, C10) : C_9 \times C_{10}$. For each pair all 37 predicted observable values were arranged in a 37-dimensional vector.

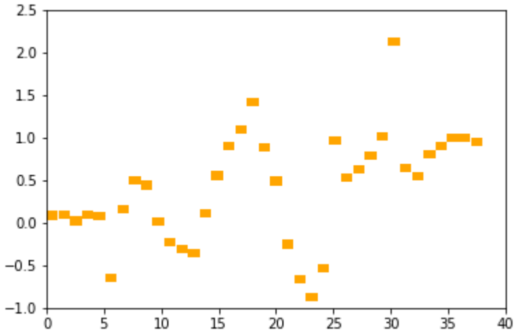

Figure 6: Observables values generated by GAN generator.

Obs1	...	Obs37

This basic training dataset, a 2D array $\boldsymbol{X}$ of dimensions $(2500, 37)$, is slightly modified to be used in each of the experiments.

3.1. Neural Network

In this experiment we train a neural network to *learn* the mapping from a set of observable bin values to C_9, C_{10} coefficients.

3.1.1. Dataset

To each of the samples in the basic dataset we add a label. This label is a 2D vector that contains the C_9, C_{10} coefficients that generated the sample.

3.1.2. Architecture

3.2. Conditional Adversarial Network

3.2.1. Dataset

A function $F(q^2) = a_0 + a_1 q^2 + a_2 (q^2)^2$ is added to each observable bin to account for the hadronic uncertainties emerging from form factor contributions. A different set of coefficients is chosen for each observable. The coefficients of the $F(q^2)$ function and the $(C9, C10)$ values are concatenated to the original 37-dimensional vector:

Obs1	...	Obs37	C_9	C_{10}	a_0	a_1	a_2

A Conditional Generative Adversarial Network is trained on this dataset. The conditional information fed to both the generator and the discriminator along each training sample is the $\log_2 \chi^2$ value corresponding to the specific selection of $C_9, C_1 0, a_0, a_1, a_2$ coefficients present in the sample. The distribution of $\log_2 \chi^2$ values is shown in Figure7.

The generator should be able to generate, after training, a particular sample with a selected χ^2 value. Our assumption in this work has been that this trained generator should also be able to generate a sample with a $\chi^2 = 0$. This sought after sample should contain 37 observables bin values close enough to the experimental values and the GANs predicted coefficients $C_9, C_{10}, a_0, a_1, a_2$.

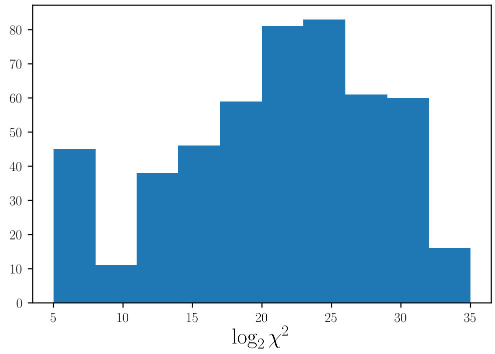

Figure 7: Training set χ^2 distribution

3.2.2. GAN architecture

The GAN have been built using TensorFlow, and open-source software library developed by Google and Keras, an open-source neural-network library written in Python that allows you to build different Neural Network architectures in terms of layers. Training GAN models is hard, model parameters might oscillate, destabilize and never converge, the generator might collapse producing limited varieties of samples, the discriminator can get so successful that the generator gradient vanishes and learns nothing, an unbalance between the generator and discriminator produces overfitting and last but not least GAN are highly sensitive to hyper-parameter selections.

4. Results

5. Conclusions

CRediT authorship contribution statement

Óliver Partida: Conceptualization of this study, Methodology, Software.

References

[1] Capdevila, B., Crivellin, A., Descotes-Genon, S., Matias, J., Virto, J., 2017. Patterns of new physics in $b \to s\ell^+\ell^-$ transitions in the light of recent data. arXiv:1704.05340.