

# Higgs Boson Machine Learning Project

Ettore Fincato, Hannah Sansford, Harry Tata

January 5, 2022

## Abstract

Write abstract here.

## 1 Introduction

### 1.1 Background

The Higgs boson can *decay* through various different processes, producing other particles in the process. In physics, one calls a decay into specific particles a *channel*. Until fairly recently, the Higgs boson had been seen only in boson pair decay channels. It is now of importance to seek evidence on the decay into *fermion* pairs, specifically *tau-leptons* or *b-quarks*, and to measure their characteristics [2]. The ATLAS experiment [1] was the first to report evidence of the *H* to tau-tau channel and the goal of this report is to improve on this analysis.

### 1.2 Overview

The Atlas experiment at CERN provided simulated data that was used to optimise the analysis of the Higgs boson. In the Large Hadron Collider (LHC), proton bunches are accelerated in both directions on a circular trajectory. This results in some of the protons colliding as the bunches cross the ATLAS detector (called an *event*), which produces hundreds of millions of proton-proton collisions per second. The particles resulting from each event are detected by sensors and, from this raw data, certain real-valued features are estimated [2].

Most of the uninteresting events (called the *background*) are discarded using a real-time multi-stage cascade classifier. However, many of the remaining events represent known processes that are also known as *background*. Our aim is to find the region of the feature space in which there is a significant excess of events compared to what known background processes can explain (called *signal*).

Once the region has been fixed, the significance of the excess is determined using a statistical test. If the probability that the excess has been produced by background processes falls below a pre-determined limit, the new particle is deemed to be discovered.

## 2 Problem Formulation

Let  $\mathcal{D} = \{(\mathbf{x}_1, y_1, w_1), \dots, (\mathbf{x}_n, y_n, w_n)\}$  be the training set, where  $\mathbf{x}_i \in \mathbb{R}^d$  is a  $d$ -dimensional feature vector,  $y_i \in \{\text{b}, \text{s}\}$  is the label, and  $w_i \in \mathbb{R}^+$  is a non-negative weight. Let  $\mathcal{S} = \{i : y_i = \text{s}\}$  and  $\mathcal{B} = \{i : y_i = \text{b}\}$  be the index sets of signal and background events respectively, and let  $n_s = |\mathcal{S}|$  and  $n_b = |\mathcal{B}|$  be the number of simulated signal and background events.

The simulated dataset also includes importance weights for each event. Since the objective function (5) depends on the *unnormalised sum* of weights, in order to make the setup invariant to the *numbers* of simulated events  $n_s$  and  $n_b$ , the sum across each set (test/training) and each class (signal/background) is set to be fixed, i.e.,

$$\sum_{i \in \mathcal{S}} w_i = N_s \quad \text{and} \quad \sum_{i \in \mathcal{B}} w_i = N_b. \quad (1)$$

These normalisation constants  $N_s$  and  $N_b$  are simply the *expected total number* of signal and background events, respectively, during the time interval of the data taking. The individual weights are then proportional to the conditional densities,

$$p_s(\mathbf{x}_i) = p(\mathbf{x}_i|y = s) \quad \text{and} \quad p_b(\mathbf{x}_i) = p(\mathbf{x}_i|y = b),$$

divided by the instrumental densities  $q_s(\mathbf{x}_i)$  and  $q_b(\mathbf{x}_i)$ , i.e.,

$$w_i \propto \begin{cases} p_s(\mathbf{x}_i)/q_s(\mathbf{x}_i), & \text{if } y_i = s. \\ p_b(\mathbf{x}_i)/q_b(\mathbf{x}_i), & \text{if } y_i = b. \end{cases} \quad (2)$$

Now, let  $g : \mathbb{R}^d \rightarrow \{b, s\}$  be a classifier. Let the *selection region*  $\mathcal{G} = \{\mathbf{x} : g(\mathbf{x}) = s\}$  be the set of points classified as signal, and let  $\hat{\mathcal{G}}$  denote the *index set* of points that  $g$  classifies as signal, i.e.,

$$\hat{\mathcal{G}} = \{i : f(\mathbf{x}_i) \in \mathcal{G}\} = \{i : g(\mathbf{x}_i) = s\}.$$

Then we have that

$$s = \sum_{i \in \mathcal{S} \cap \hat{\mathcal{G}}} w_i \quad (3)$$

is an unbiased estimator of the expected number of signal events selected by  $g$ , and similarly,

$$b = \sum_{i \in \mathcal{B} \cap \hat{\mathcal{G}}} w_i \quad (4)$$

is an unbiased estimator of the expected number of background events selected by  $g$ . Alternatively,  $s$  and  $b$  are the *unnormalised* true and false positive rates, respectively.

High-energy physicists suggest the use of the *approximate median significance* (AMS) objective function defined by

$$\text{AMS} = \sqrt{2 \left[ (s + b + b_{\text{reg}}) \ln \left( 1 + \frac{s}{b + b_{\text{reg}}} \right) - s \right]} \quad (5)$$

to optimise the selection region for discovery significance, where  $b_{\text{reg}}$  is a regularisation term suggested to be set to  $b_{\text{reg}} = 10$ . Hence, our aim is simply to train a classifier  $g$  based on the training data  $\mathcal{D}$  with the goal of maximising the AMS on some unseen test data.

## References

- [1] Evidence for Higgs Boson Decays to the  $\tau^+\tau^-$  Final State with the ATLAS Detector. Technical report, CERN, Geneva, Nov 2013. All figures including auxiliary figures are available at <https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/CONFNOTES/ATLAS-CONF-2013-108>.
- [2] Claire Adam-Bourdarios, Glen Cowan, Cecile Germain, Isabelle Guyon, Balazs Kegl, and David Rousseau. Learning to discover: the higgs boson machine learning challenge.