ML Challenge 6

Jacob Murphy, Julie Roche
Ohio University Physics and Astronomy Department

August 2020

1 Introduction

The calorimeter used for DVCS in Hall A consists of an assembly of 13x16 lead-glass blocks. Each block reads out a voltage signal produced as a result of an initial photon hitting the block; it is sampled every 1 ns and stored into a 128 integer array. The shape-versus-time of the signal produced by a block is a characteristic of the lead-block material and is well known. Using this, a recursive two steps analysis system based on χ^2 minimization is used in understanding the precise time of arrival of the high energy photon (better than 1 ns) as well as its total energy (better than 2%). The arrival time (t_0) of the signal and the overall amplitude (A) vary event per event. Figure 1 shows a large single pulse measured by a block. For most events, only one high energy photon hits the calorimeter at the time. Complications to this analysis scheme arise from the events for which multiple photons hit the calorimeter within the 128 ns time frame, as seen in Figure 2.

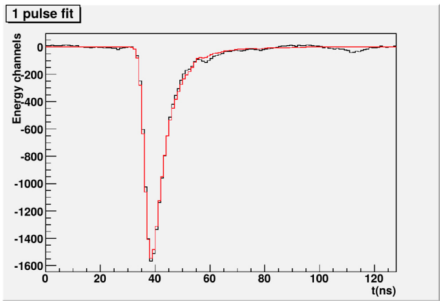

Fig. 1: Clean large signal from a block. The black points are the data while the red trace is the offline fitted signal.

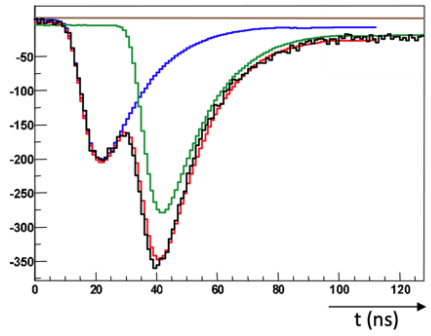

Fig. 2: Typical two pulse signal. The data are the black line. The total fitted signal is in red and is the sum of two pulses of different amplitudes A arriving at different times (blue 20ns and green 40ns).

2 Goal

The goal is to design and train a model such that given an input of 128 signals, can output the amplitude and arrival time of any potential pulses. The maximum possible number of pulses is 2, meaning this output consists of 4 values: A1, A2, t1, t2. Where A is the amplitude of a pulse and t is the arrival time of the pulse, or time of the peak.

Amplitude is always taken as a positive value for real data. For example, an Amplitude of 400 would mean a peak going down to -400 Energy Channels. Arrival time is the time to the pulse peak since the start of the signal. The first signal is at time 0 ns, and the last is at time 127 ns. Should a pulse be non-existent, its amplitude should output as 0 and arrival time as -1. The definition of pulse 1 vs pulse 2 is not a priority

for this problem, so pulse 1 may have the larger arrival time and bigger or smaller amplitude, and pulse 2 could be the only pulse in the array of signals.

The model will be expected to accept a 128-signal input and always output the 4 values A1, A2, t1, and t2. The size of the input will not vary, though the number of pulses in the input will be either 0, 1, or 2. The model should be able to recognize how many pulses there are and return 0 amplitude and -1 arrival time for non-existent pulses. In other words, a 0-pulse event should output A1 = A2 = 0, and t1 = t2 = -1. A 1-pulse event should output either A1 = 0 and t1 = -1 or A2 = 0 and t2 = -1.

3 Materials

All materials are available **here**. Individual copies can be made available by emailing Jacob Murphy at jm443918@ohio.edu.

The training set will consist of rows of 133 comma-separated values. The first 5 are the baseline value, A1, A2, t1, t2. The baseline value is the average value of the signal without pulses. It is not required to be found but rather given to potentially help with training. The remaining 128 values are the signal values from time 0 ns to time 127 ns. The training set contains a little over 190k events.

The test set will be a CSV file with only the 128 signal inputs. Aside from missing the first 5 values, the format will be identical to the training set. Output is expected to be of a similar form to the training set: a CSV file with 4 columns (A1, A2, t1, t2) and a row for every event.

4 Judging Criteria

On Wednesday, November 4th at noon, the test set will be released. Participants will have 48 hours, or until Friday, November 6th at noon, to make a submission. Submissions should be sent to Jacob Murphy at jm443918@ohio.edu and must consist of all scripts used for training and testing, along with the results from the test set in a CSV file. The results should contain a row for every event and 4 columns for A1, A2, t1, and t2 (in that order). Any submission not based on ML only will be disqualified (eg. no pre-processing with traditional methods). There are no restrictions for ML methods.

For each value (A1, A2, t1, t2), the submission will be compared to the true values (those found with traditional methods) and the sum of the difference squared will be taken. For these values, each pulse will be compared to both true pulse values and the lower difference will be taken. In other words, if the submitted pulse 1 matches better with the true pulse 2, that difference will be used for the sum. Similarly, the number of pulses found will be compared to the true value and the sum of the difference squared will be taken. Number of pulses will be determined by non-zero amplitudes in conjunction with corresponding time values not equal to -1.

The individuals or teams with the lowest sum in each of the 5 categories (A1, A2, t1, t2, pulse number) will receive 3 points. The next lowest in each category will receive 2, and then 1. Ties in these categories will receive equal points, with the next rank being skipped (ie two 1st places followed by 3rd, or two 2nd places followed by no score). The highest total score will win, with ties being broken by the scores for pulse number.