High-statistics and GPU Accelerated Data Analysis

Philipp Eller (for the IceCube-Gen2 Collaboration)
Penn State University, Dept. of Physics, 104 Davey Lab, 10802 PA University Park
E-mail: pde3@psu.edu

Abstract. We present methods to perform high statistics data analyses to investigate fundamental neutrino properties in large volume neutrino detectors, fast and with modest computational resources. The introduced measures are threefold: speeding up computations using graphics processors, evaluating the underlying physics processes on a grid instead of treating every event individually and lastly applying smoothing methods to quantities obtained from Monte Carlo simulations. We show that with our method we can get reliable analysis results using significantly less simulation than what is usually needed, and that the timing to run an analysis with our method is independent of sample size.

1. Introduction
Proposed very large volume neutrino telescopes (VLVnTs), such as PINGU [1, 2] or ORCA [4], will collect unprecedented amounts of atmospheric neutrinos (in a typical event selection, several hundred thousand events a year). Said neutrinos can be used to perform various measurements, such as the atmospheric oscillation angle $\theta_{23}$, the mass squared splitting $\Delta m_{31}^2$ and its sign (neutrino mass hierarchy, NMO) and the rate of $\nu_\tau$ appearance just to name a few. Also limits on non-standard interactions, sterile neutrinos and other exotic phenomena can be extracted.

What all of these analyses have in common, is that they compare distributions of simulated neutrino events with given model parameters to the observed data. The typical procedure is to perform a maximum likelihood analysis of the reconstructed energy and arrival direction of the observed neutrinos to find the maximum likelihood estimators of the model parameters. To do so, a large parameter space needs to be explored, and every time the distributions of the reconstructed event properties (templates) generated to compare them to the data. The distributions that are compared to the data should be of significantly better statistical precision (typically an order of magnitude) than the data itself.

The canonical way of achieving this is by simulating large enough samples, we refer to this in the following as direct histogramming. A way to alleviate statistical fluctuations of otherwise too small samples is to use smoothing techniques, such as kernel density estimation (KDE) [5] instead of simple histograms, we refer to this as direct KDE. Our proposed method will be explained in the following and referred to as PISA. The used techniques and methods are explained in much greater detailed in [3].

2. The PISA staged approach
Instead of directly using simulated data by putting the events into a histogram or KDE to obtain the final distributions that can directly be compared to the data, we take a different route. The simulated data is rather used to characterize the experiments properties, here its
acceptance and resolutions. These quantities together with the atmospheric neutrino flux and its flavor oscillation probabilities are used to generate high-statistics samples of reconstructed event properties in the following way: The initial flux (1) multiplied together with the oscillation probabilities (2) and the experiment’s acceptance (3) yield truth-level distributions (see figure 1). In a convolution with point spread and energy resolution function (4) these spectra are translated into the space of reconstructed variables.

Figure 1: Illustration of the PISA staged approach for obtaining event templates, here for simplicity using a characterization in one dimension (energy) only. Steps 1, 2, and 3 are in true energy \( E_{\text{true}} \); the product of these yields the expected event distribution (lower left). Smearing this spectrum with energy-dependent energy resolution functions (step 4) gives the reconstructed event rate spectrum (lower right). The different horizontal axes on panel 4 are used to show resolution kernels of different true energies. Note that the dotted green line in step 2 shows a hypothetical change of oscillation parameters, affecting only stage 2.

2.1. Validation

Table 1 shows, that in the limit of dense enough grids on which the individual properties are evaluated this method is equivalent to direct histogramming. When using 320 x 320 grid points or more, the maximum \( \chi^2_{\text{max}} \) distance of all bins is more than an order of magnitude smaller than the average \( \langle \chi^2 \rangle = 1.0 \) that is expected when comparing to data. In the next section it will be demonstrated that this way is faster when using large sample sizes. But more importantly, this approach allows to apply smoothing to quantities that are well understood and behaved.

| Grid \((M \times N)\)          | 40 x 40 | 80 x 80 | 160 x 160 | 320 x 320 | 640 x 640 | 1280 x 1280 |
|--------------------------------|---------|---------|-----------|-----------|-----------|-------------|
| \( \langle \chi^2 \rangle \)   | 0.01067 | 0.00253 | 0.00060   | 0.00014   | 0.00003   | 0.00001     |
| \( \chi^2_{\text{max}} \)     | 1.45906 | 0.46930 | 0.19718   | 0.04974   | 0.00634   | 0.00172     |

Table 1: Average and maximal \( \chi^2 \) deviations between non-smoothed staged approach and direct histogramming, for different grid point densities in \((E_{\text{true}}, \cos \theta)\), using an MC dataset of \(10^6\) events.
2.2. Smoothing

Instead of applying smoothing techniques to the final event distributions (as shown in direct KDE), we take the advantage that the underlying properties (acceptance and resolutions) of the experiment are inherently easier to deal with – the shape of the acceptance for example is completely decoupled from rapidly changing oscillation probabilities. We apply a simple spline smoothing to the acceptance and adaptive KDEs to the resolution functions. So this method is general and does not make assumptions of a parametric form of the resolution functions. A comparison of final level distributions obtained with histogramming and with the PISA method are compared for different sample sizes in figure 2 A more qualitative comparison of final level event distributions are shown in figure 3. These are obtained from a generic toy detector model, and therefore the underlying true distributions are known and also shown.

3. Results

Running a typical analysis (here the determination of the neutrino mass hierarchy) gives an idea of the severity of lacking statistical precision in the templates. Figure 4 shows the result one would obtain using a given sample size of simulated events and using a given technique\textsuperscript{1}. For direct histogramming, relatively large sample sizes need to be used to obtain unbiased, realistic answers. The KDE applied to the final events can certainly help here, but it will be shown later that this techniques (apart from being inferior to our method) is extremely slow. The proposed PISA method outperforms the other two methods across the board and even for the smallest tested sample sizes, the obtained answers are still compatible with truth.

\textsuperscript{1} The result is shown in terms of $\chi^2$ difference between the best fit of both neutrino mass orderings to a given pseudo data trail.
Figure 3: Final level templates used for the example data analysis. The reference distributions (truth) obtained directly for a toy detector model parameterizations are shown in panel (a). Given the same sample of $10^4$ events the estimated distributions using histograms are shown in panel (b), using KDEs in panel (c), and using the PISA staged approach with a (400 x 400) grid in panel (d).

3.1. Benchmark

In figure 5 we show the average time necessary to generate a template, for a typical analysis this is done many times. While the direct histogramming method scales roughly linear with sample size, the direct KDE gets slower faster than linear as the sample size growth, making it impractical for usage with sample sizes much larger than $\approx O(10^5)$. The proposed PISA method is by construction independent of input sample size (excluding initial start-up time which is growing with sample size, but negligible when many (typically thousands) of templates need to be generated).

Furthermore, we implemented all of the methods, or part thereof, to run on graphics co-processors (GPUs) to achieve speed ups. The gain for the two direct methods is in the order of a magnitude faster, for the PISA approach not all steps have yet been ported to harness the GPUs power and therefore the speed improvement are less.

However, the PISA method is as fast or faster then the other method when sample sizes are considered that deliver useful physics answers as outlined in the previous section.

4. Summary

For high statistics data analyses, as needed to analyze the data from new large neutrino telescopes, high-precision expected event distributions from given model assumptions are needed to compare to the data. The challenge is to keep the computational burden low to facilitate data analysis. While simple histogramming of events suffers from large inherent statistical fluctuations, state for the art smoothing techniques like KDE can help, but at the same time
Figure 4: Estimated sensitivity ($\sqrt{\Delta \chi^2}$) to the NMO vs. sample size for direct histogramming, direct KDE, and the proposed PISA staged smoothing methods applied to multiple (between 50 and 200) statistically independent toy MC datasets. Vertical lines indicate central 68% quantiles. The dashed horizontal line shows the significance obtained from truth templates based on the parametric toy detector model. The staged approach outperforms the other methods—both in terms of bias and variance—for sample sizes through $3 \cdot 10^6$, with direct histogramming only matching the staged approach using $10^7$ samples. Note that no data points exist for direct KDE and sample sizes above $3 \cdot 10^5$, as computational processing times become impractically large. Also note that direct histogramming is off-scale high for fewer than $3 \cdot 10^4$ events (mean values are indicated to the right of the corresponding markers).

can be slow. Our approach uses simulated events and smoothing techniques to characterize generic detector properties. The expected distributions of neutrino event properties are then constructed from these generic detector properties. The main advantages are higher precision due to reduced fluctuations and better computational performance.

References
[1] M. G. Aartsen et al. Letter of intent: the Precision IceCube Next Generation Upgrade (PINGU). 2017.
[2] M. G. Aartsen et al. PINGU: a vision for neutrino and particle physics at the South Pole. J. Phys., G44(5):054006, 2017.
[3] M. G. Aartsen et al. Computational Techniques for the Analysis of Small Signals in High-Statistics Neutrino Oscillation Experiments. 2018.
[4] S. Adrian-Martinez et al. Letter of intent for KM3NeT 2.0. J. Phys., G43(8):084001, 2016.
[5] Kyle S. Cranmer. Kernel estimation in high-energy physics. Comput. Phys. Commun., 136:198–207, 2001.
Figure 5: Average template generation time during a typical analysis for input data sets of varying size, shown for the direct histogramming, the direct KDE, and the staged approach. Solid lines represent timings based on (partial) GPU acceleration, whereas the dashed ones are for CPU-only calculations.