PHY324 Data Analaysis Project:

Winter 2023

1 Project Description

This document describes a simplified process of the initial steps for data analysis, converting acquired raw data to energy spectra, and performing a fit for the converted spectrum. You will be given three data sets in pickled format, and a snippet of code to illustrate how to load the data. You are expected to analyze the data with your own Python code, and write your process and results in a report. The report needs to take the form of a self-contained scientific paper, including a brief introduction, description of the experimental setup and data collected – with the information provided by this document, what you have done and the rationales behind them, the results and conclusions. The expected audience is your peer students with no knowledge of this project. All results should be presented in forms of graphs or tables, clearly labelled with proper captions, and referenced in the main text when appropriate. Appropriate references and citations are mandatory. This applies to quoting information from this document as well – you are expected to properly rephrase the information you quote from here, and cite this document.

This is an individual assignment. While you can get help from anyone you wish, you must do the analysis yourself, make your own graphs, and write your own report.

2 Experiment and data

This experiment is conducted in a particle detection scenario. With a typical particle detector, the sensor converts each particle incident energy into measurable electrical signals, in the form of an excursion from the quiescent voltage. We recognize these excursions as "pulses". A typical pulse looks like Fig. 2. The shape of the pulse is often dictated by the characterization of the detector and its readout electronics system, as well as the type of energy deposition. In this scenario, we assume a single species of energy deposition, namely electron-recoils in the detector material caused by high energy photons, either through photoabsorption effect, or through Compton scattering. With a specific detector setup, this leads to a fixed pulse shape. The detector and its readout circuit is characterized to give a 20 μs rise time (τ_{rise}) and a 80 μs fall time (τ_{fall}) to the pulses, following the functional form of

$$y = A * C * (e^{-t/\tau_{\text{rise}}} - e^{-t/\tau_{\text{fall}}}), \text{ where} \tag{1}$$

$$C = \left(\frac{\tau_{\text{fall}}}{\tau_{\text{rise}}}\right)^{-\frac{\tau_{\text{rise}}}{\tau_{\text{fall}}-\tau_{\text{rise}}}} \cdot \left(\frac{\tau_{\text{rise}} - \tau_{\text{fall}}}{\tau_{\text{fall}}}\right) \tag{2}$$

is a normalization factor so that the term without the scale-able amplitude A has an amplitude of unity. The amplitude of the pulse, A, is varying as a function of the energy deposited in the detector. Here we have a detector with a perfectly linear response, meaning A is proportional to the energy the detector senses. An idealized pulse shape with no noise is shown

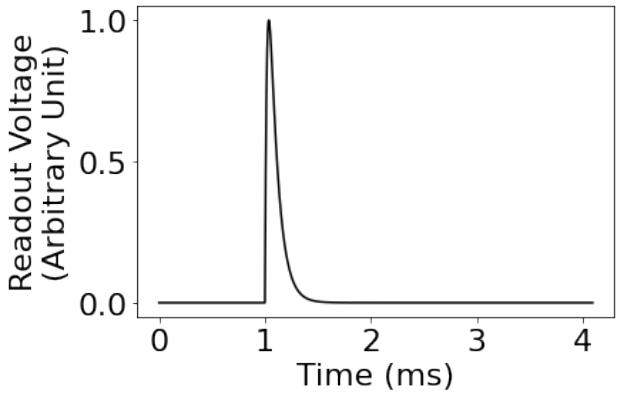

Figure 1: An idealized pulse with no noise. Each pulse is stored as 4096 voltage samples readout by a 1 MHz data acquisition system. The pulse onset is 1 ms (1000 samples) from the beginning.

in Fig. 1. Note that the pulse starts at 1 ms. We generally refer to the area before the pulse onset, from 0 ms to 1 ms in this case, as the "pre-pulse" region.

A realistic pulse, as shown in Fig. 2, is taken with an acquisition system that measures the voltage output of the detector at 1 MHz rate. With every particle incident, the system stores 4096 samples of the voltage measurements as a "trace", thanks to a trigger system. The idealized trigger system senses each pulse onset, and position the start of the pulse at the 1000 sample of the corresponding trace. The system also adjusts the quiescent voltage to around 0 V, though due to the presence of low-frequency noise, the quiescent voltage does fluctuate a bit. As usual, noise is present in the voltage readout. These noises can be attributed to sources intrinsic and extrinsic to the detector and its readout circuit. The intrinsic noise sources include noise caused by electron random motions, instability caused by temperature fluctuations, noise induced by the readout circuit, etc. The extrinsic noise sources include the instability in the power of the system and its corresponding ground, pickups from the environmental electromagnetic waves, etc. To quantify the effect of the noise, we acquired a set of data with only noise.

The noise is superpositioned with the pulses induced by energy deposited by the incident particles, making it hard to reconstruct the size of the pulse and thus infer the energy of sensed by the detector. In some scenarios, it also makes detecting tiny pulses impossible, though such a scenario is beyond the scope of this analysis, thanks to the idealized trigger system we employed. Said differently, the trigger efficiency of the system we constructed here is 100% irrelevant to the energy deposition.

To quantify the detector response to energy deposition, we also took a set of "calibration" data. This set of data is acquired by exposing the detector to a known calibration source emitting 10 keV photons, with the knowledge that the photons will interact with the detector through photoabsorption process, thus depositing all of its energy. We note that despite that this is supposed to be a calibration data, background events persist, and are indistinguishable from the calibration events on an event-by-event basis. Background events are often caused by radioactive isotopes in the environment or in the detector itself. In our data, it

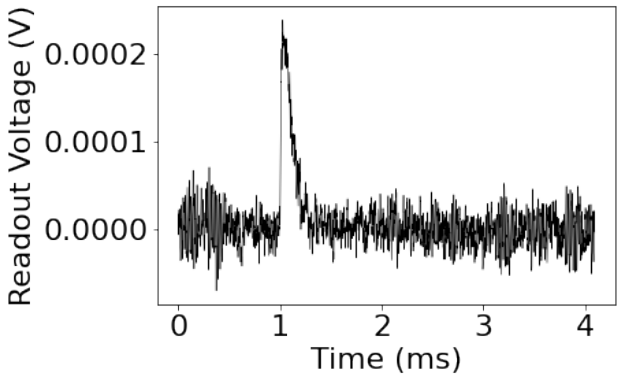

Figure 2: An example pulse from the detector used in this project. Each pulse is stored as 4096 voltage samples readout by a 1 MHz data acquisition system. The pulse onset is 1 ms (1000 samples) from the beginning. The quiescent voltage is adjusted to be around 0 V, though fluctuating due to low frequency noise present in the system. A typical pulse has a 20 μs rise time, and a 80 μs fall time. The pulse is superpositioned on top of noises caused by assorted physical phenomena, including electron random motions, pickup of environmental electromagnetic waves, etc.

presents itself as a group of events with energies distributed randomly in our energy region of interest (ROI) of 0-20 keV. Thus, the deposited energy spectrum of this calibration data is a narrow Gaussian peak at 10 keV from the calibration source on top of a uniform distribution caused by backgrounds. However, noise will broaden the Gaussian peak, and the size of the broadening depends heavily on the "energy estimator" we use to estimate the size of the pulse. Finding an optimized energy estimator is often a critical step for data analyses. We will need to explore a few energy estimators, calibrate each individually, assess the energy resolution of them, and use the best one we could find.

After calibrating the detector and the energy estimator, we can then expose the detector to the "signal source" we want to measure, and extract the energy spectrum. We did so and took another set of data. For this set of data, we also limited the ROI to 0-20 keV. We note that background is also present while we measure the signal source and remains indistinguishable on an event-by-event basis. Luckily, it remains an uniform distribution across our ROI. The ultimate goal of this measurement is to reconstruct the energy spectrum measured from this signal source, and attempt to fit it with a functional form. Typically a followup analysis would use this fitted functional form to extract physical information of the signal source. Such a followup analysis is beyond the scope of this project.

3 Tasks breakdown

While the previous section contains all the information about this project, here are a few steps we suggest you to follow to achieve the final goal of fitting the spectrum from the signal source with a functional form.

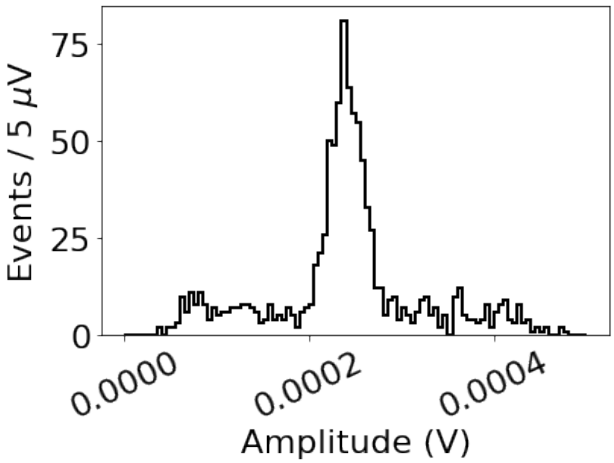

Figure 3: Example histogram of the energy estimator reconstructed from the calibration data. Note a flat background component and a Gaussian peak corresponding to the 10 keV energy depositions from the calibration source.

1. It is always a good practice to have a well-established starting point of any project. As a first step, we suggest you to make sure you can successfully read the data sets provided. A snippet of code is provided in Appendix A of this document. Reproducing Fig. 2 will establish that you can successfully load the data. One hint is that Fig. 2 was produced with the event labeled as "evt_2" in the calibration data set. It might be a good idea to include a graph from each data set in the paper, with a few pulses included in each, to show that this was achieved.

2. The next step would be to establish an energy calibration of the detector.

   (a) An energy calibration is a relation between the size of the pulse and the energy deposited in the detector. This can be achieved by reconstructing an energy estimator from each pulse in the calibration data, making a histogram of them, and identifying the structure in such a histogram – in this case a Gaussian peak due to the 10 keV calibration source. Fig. 3 shows an example of this histogram. Then you can perform a fit on the histogram, with a Gaussian function on top of a flat background. The fit can be a chi-squared fit, with each bin of the histogram following a Poisson distribution, approximated by a Gaussian distribution. For each chi-squared fit, we need to establish the fit quality, by evaluating the chi-squared probability based on chi-squared calculated and its degree of freedom. The mean of the Gaussian function corresponds to 10 keV in energy. The whole spectrum can thus be converted to energy. The width of the Gaussian function, after converting to energy, is the resolution of this detector with this energy estimator.

   (b) Commonly used energy estimators are often either the amplitude of the pulse, or the integral of it. How to estimate the amplitude or the integral can be tricky though.

i. The simplest amplitude estimation can be max-min, or max-baseline, where baseline can be estimated with the average of the pre-pulse region. Give both estimations a try, and show their performances in your report.

ii. A simple integral can be the sum across the whole trace. A baseline subtraction might be able to enhance the performance a bit, just like the amplitude. Furthermore, limiting the range of the integral can also improve the performance. Give all three options a go, and show their performances in your report.

iii. A more sophisticated way to estimate the amplitude of a pulse can involve a chi-squared fit of the pulse to a known shape – in this case a pulse shape with a 20 μs rise time and a 80 μs fall time as in Eqn. 1 is well justified. The uncertainties of each voltage measurement can be estimated with the noise data with the averaged standard deviations of the traces. However, this is not ideal either – as the underlying assumption for a chi-squared fit is not strictly satisfied in this scenario, due to correlations in the noise.

(c) For this step of the project, we need to realize these energy estimators, establish energy calibrations of with each energy estimator, and estimate the energy resolution of each. For each energy estimator, include a graph of the spectrum of the calibration data before applying the calibration with a chi-squared fit overlaid with it, a calculation of the chi-squared probability to gauge whether the fit is valid, an energy spectrum after the calibration with another fit to confirm the energy calibration and resolution. A table of the information extracted from these fits could help compare the energy estimators and help reach the conclusion of the best energy estimator, thus please include that in the report. The table should contain the name of each calibration method, their derived calibration factor, its corresponding energy resolution, and the fit χ^2 probability. A set of example figures and tables for one energy estimator can be found in Appendix. B. The energy estimator in the example has fairly poor performance, so hopefully at least some of your results are significantly better. There are at least six suggested energy estimators. A set of figures and an entry to the table for each energy estimator is expected in the report.

(d) For the next step, we will use the best one concluded from the table mentioned above.

3. Apply the energy estimator of choice and its calibration on the signal data, and arrive at the energy spectrum of the signal source + background. We need to guess what is the functional form of this spectrum, and carry out another chi-squared fit to estimate the parameters in the functional form. Again, use the chi-squared probability to gauge whether the functional form you chose is valid or not. If the functional form is valid, estimate the parameters and their uncertainties, and present them in a table or a graph. A few questions to consider: 1) Are the uncertainties on the parameters independent from each other? 2) How does the uncertainty in the energy calibration factor into the final uncertainties?

   Regarding functional forms to consider, note that the events might not be from one

simple component. If one simple function doesn't capture the structure of the result, consider the possibility that there are events from a couple of different sources thus it's appropriate to sum up multiple functions. Popular guesses include using Gaussian or Lorentzian functions to model peaks; using polynomials, exponential functions or error functions to model continuous shapes, and if there seems wiggles, one can try Legendre polynomials or Chebyshev polynomials. If detector resolution is not negligible, it will "smear" the spectrum, which is functionally equivalent to convolving the underlying spectrum with a detector response function, i.e. a Gaussian function in our simplified scenario. Note that you are not required to precisely model the spectrum you get from the signal data set, only encouraged to model it to the best you can, and quantify the goodness of the fit. If a global fit becomes challenging, a piece-wise fit can sometimes be considered as a fallback option.

You have two primary tasks: first, find the best energy estimator you can; second, apply it to the data and reconstruct what happened. This last step takes some creativity on your part. Science is a creative process!

Final note: when you are doing any calculations or fittings for any of the steps of this project, you can make any choices you wish such as, but not limited to, ignoring/deleting some of the data, and changing the number of bins used in a histogram (which can have a major impact on the results). Document your choices, especially if they improve the resolution of your energy estimators. These choices are what we want to see you do! You do not have to use just the six suggested energy estimators as your final choice (you do have to try them all and document them). If you find a better one, document it as a seventh option and use it.

Appendix A Data format

As stated in the document, we took three sets of data for this project, and saved them into three pickled files. The file names and their contents can be found in Table 1. In each pickled file, there is a dictionary, with 1,000 keys. Each key is named as "evt_x" where x iterates from 0 to 999. The content of each key is a numpy array of the length of 4096, where each entry of the array is a voltage measurement done by a 1 MHz data acquisition system. The following snippet of code can load and plot a pulse from one of the data file.

```
import pickle
import matplotlib.pyplot as plt
import numpy as np
with open('calibration.pkl','rb') as file:
    data_from_file=pickle.load(file)
xx = np.linspace(0, 4095/1e3, 4096)
plt.plot(xx, data_from_file['evt_0'], color='k', lw=0.5)
plt.xlabel('Time (ms)')
plt.ylabel('Readout Voltage (V)')
```

Data	Description	Filename
Noise	Noise only	`noise.pkl`
Calibration	Background + 10 keV calibration source	`calibration.pkl`
Signal	Background + signal source	`signal.pkl`

Table 1: List of data sets taken, with descriptions and file names.

Appendix B Example figures and tables for one energy estimator

Figure 4 shows a set of example figures for a not optimized energy estimator (the amplitude of the whole trace), before and after the energy calibration. Note that after energy calibration, the center of the peak correctly lands at 10 keV, and the σ of the Gaussian peak estimates the energy resolution of this data set with this energy estimator. Also note that the χ^2 probability shows that this is not a good fit at all. This often happens when a model, flat background + Gaussian calibration peak, does not properly describe the data. (Can you figure out why that is the case?)

Whether such a non-ideal fit is still usable is a different question. Here we are mostly interested in the location of the peak. As shown in the right panel of Fig. 4, after the calibration, the mean of the Gaussian calibration peak is consistent with the 10.0 keV energy we aimed for. This calibration fit can be used, with the caveat that it is not a good fit, thus during all future steps of the data analysis we need to raise the question of "would a non-ideal calibration fit potentially affect the analysis".

The summaries of this energy estimator is included in Table 2. As more energy estimators are explored, additional rows with more optimized energy estimators should be included in this table.

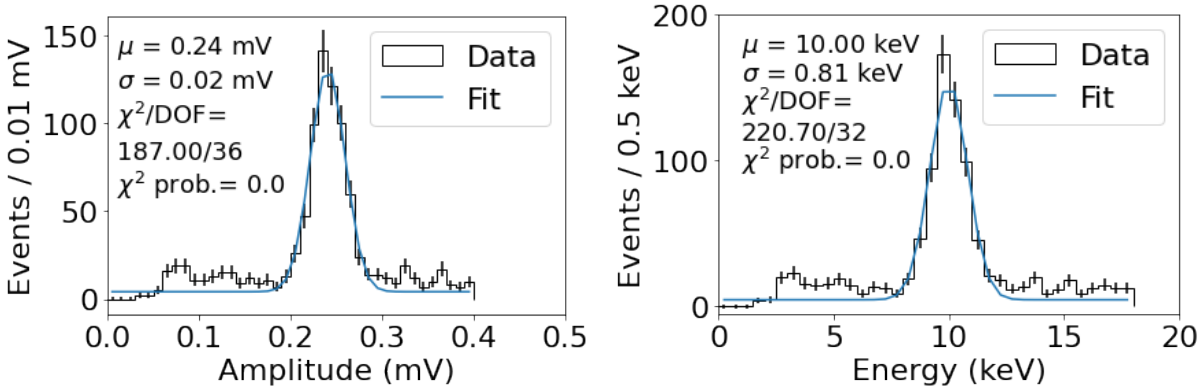

Figure 4: Example histograms of a not-optimized energy estimator from the calibration data set, before (left) and after (right) energy calibration.

Table 2: Example table of energy estimator evaluation. Additional rows with more optimized energy estimators should be included.

Energy estimator	Calibration factor	Energy resolution	Fit χ^2 probability
Amplitude of trace	41.7 keV/mV	0.81 keV	0