

Rheinisch-Westfälische Technische Hochschule Aachen

Master Thesis

Denoising Methods for Multi-Dimensional Photoemission Spectroscopy

SUBMITTED BY

Muhammad Zain Sohail

Rheinisch-Westfälische Technische Hochschule Aachen

Christian-Albrechts-Universität zu Kiel

Deutsches Elektronen Synchrotron

SUPERVISORS

Prof. Dr. Benjamin Berkels

Rheinisch-Westfälische Technische Hochschule Aachen

Prof. Dr. Kai Rossnagel

Christian-Albrechts-Universität zu Kiel

Deutsches Elektronen Synchrotron



Aachen, October 2024

Abstract

In the realm of photoemission spectroscopy, the exploration of large multi-dimensional phase spaces necessitates time-intensive data acquisition to ensure statistical robustness. Despite the unparalleled capabilities of free-electron lasers (FELs), in peak brightness and ultra-short pulsed X-rays, the limitations of low repetition rates prolong the data acquisition process. This impedes the agility of decision making that could otherwise enhance experimental results in the limited and valuable beamtime. By employing denoising strategies to mitigate noise while preserving intrinsic information, our proposed approach aims to streamline the data acquisition process, and effectively manage the escalating size and complexity of multi-dimensional photoemission data.

Contents

Abstract	ii
Contents	iii
List of Acronyms	v
Glossary	vi
1 Introduction	1
2 Photoemission Spectroscopy: Interaction, Light Sources, and Detection	2
2.1 Light-Matter Interaction	2
2.2 Light Sources	3
2.2.1 Laser	3
2.2.2 Synchrotron	3
2.2.3 Free-Electron Laser	3
2.2.4 HEXTOF instrument at FLASH	4
2.2.5 Binning as means of Imaging	6
2.2.6 Describing the data GdW, WSe2, GrIr, and new one	6
2.2.7 Creating the dataset Corrections Calibrations etc	6
3 Denoising preliminaries and BM3D	8
3.1 Address reconstruction/denoising schemes	8
3.2 Metrics	8
3.3 BM3D: Denoising in sparse domain	9
3.3.1 Anscombe: Variance Stabilization Transform	9
4 Characterizing Photon and Photoelectron Statistics	10
4.1 Poisson distribution as a model for counting statistics	10
4.2 Correlated events in detector	11
4.2.1 Before filtering	14

CONTENTS

4.2.2	After filtering	14
4.3	Simulate Noise	14
4.4	Statistical testing	14
4.5	Chi-squared Goodness of Fit Test	16
4.6	Modeling over-dispersed count data	18
4.7	The SASE process as an explanation of over-dispersed statistics	18
5	Deep Learning	19
5.1	Statistical Learning	19
5.1.1	Optimal Loss Function	20
5.1.2	Regularization	20
5.1.3	Optimization	20
5.2	Noise2Noise: Deep Learning framework	20
5.2.1	Convolutional Neural Networks	20
5.2.2	Autoencoder	20
6	Conclusion and Outlook	21
	Bibliography	22
	Acknowledgements	23
	Contributions	24
	Appendices	25
	A Transforming Raw Data to structured format	26
	B Mathematical Background	29
B.1	Measure Space and Measures	29
B.2	Probability	30
B.3	Landau Notation	30
B.4	Statistical Inference	31
B.4.1	Testing/Confidence Intervals	31
B.4.2	Optimal parameter estimation	31
C	Deep Learning	32
C.1	Infrastructure	32

List of Acronyms

- BAM** beam arrival monitor. 24
- CDF** cumulative distribution function. 14
- DESY** Deutsches Elektronen-Synchrotron. 4
- DLD** delay line detector. 4, 8, 24
- FEL** free-electron laser. vi, 7
- GMD** gas monitor detector. 24
- HDF5** hierarchical data format version 5. 24
- HEXTOF** high energy X-ray time-of-flight. 4
- MAE** mean absolute error. 17
- MCP** micro channel plate. 4
- MSE** mean squared error. 17
- OpenCOMPES** open community of multidimensional photoemission spectroscopy. 24
- PES** photoemission/photoelectron spectroscopy. 3
- PMF** probability mass function. 7
- PSNR** peak signal-to-noise ratio. 17
- SASE** self-amplified spontaneous emission. vi
- SED** Single Event DataFrame. 24
- SSIM** structural similarity index measure. 17
- TOF** time of flight. 4

Glossary

Beamline A path leading the photons from the particle accelerator to the experimental end-station.. 24

Free-Electron Laser is an x-ray radiation source; fundamentally comprising of a linear particle accelerator and an Undulator (or a series of undulators). The accelerator produces a bunched electron beam similar to that of a synchrotron, which can be compressed to reach ultrashort pulse duration (femtosecond) with peak brightness many orders of magnitude above synchrotrons. Since the electrons move in a vacuum, it is termed Free-Electron in comparison to traditional lasers which are bound by the materials energy levels. Whereas, it is called a Laser due to there being light amplification and the shared properties with traditional optical lasers such as high pulse energy and being coherent. Taken from [1]. 1

Microbunch Microbunches are produced by the interaction between the oscillating electrons in the undulator and the radiation that they produce (due to the oscillatory acceleration) leads to periodic longitudinal density modulation known as Microbunching. The in-phase emitted radiation adds coherently, increasing intensity and enhancing microbunching. Adapted from [2]. vi

Pulse Also known as *Microbunch*. Each Train contains about 500 pulses produced from the self-amplified spontaneous emission (SASE) process (See Self-Amplified Spontaneous Emission). These are which are used as a secondary index for the data reduction process. 1, 24

Self-Amplified Spontaneous Emission is a process where the electron beam in the accelerator, when passing through an undulator, starts emitting radiation due to acceleration. The interaction between the emitted radiation and the charge distribution leads to microbunching. These microbunches emit radiation coherently, leading to the intense, coherent radiation, characteristic of a free-electron laser (FEL). For more information see [2]. vi, 1

Train Also known as *Macrobunch*. A train represents a group of closely spaced electron bunches produced and accelerated by the FEL (or more generally, any accelerator). Each train is associated with a unique identifier called `trainId`, which is used as the primary index for much of the data reduction process. vi, 1, 24

Undulator Magnets arranged periodically to produce a periodic magnetic field. It is used to produce coherent radiation by accelerating electrons through it. vi,
1

Introduction

Scientific methodology, especially in empirical fields such as physics, fundamentally rely on observations as the basis to understand the principles of nature. The act of measurement, however, is seldom free from ambiguity. For instance, due to environmental factors, instrument limitations, or the intrinsic uncertainties of quantum phenomena.

Ingeniously designed experiments, sophisticated detection schemes, and controlled environments—such as ultra-low temperatures—all contribute to minimizing external noise and improving the quality of data collected. However, as experiments push the boundaries of scale and precision, the limitations imposed by quantum mechanics, such as uncertainty and fluctuations, remain unavoidable.

In modern days, a paradigm shift has occurred where instead of trying to explicitly model the system we are trying to learn about, has also brought revolutions such as machine learning where we don't need to know the model itself and basically make the machine learn the non-linear model. This doesn't explain the process but allows us to perform for example inference. Learning is a sort of metamodel, where we don't need to know the theory itself but can still make predictions.

Expensive aspect is to buy new equipment. People don't use math methods first. Then we go into denoising

With the necessity to acquire data in so many dimensions, because we have low electron counts.

Photoemission Spectroscopy: Interaction, Light Sources, and Detection

In the seminal paper by Einstein [3], that laid foundations to Quantum Mechanics, Einstein postulated that light is made of discrete quanta of energy $E = h\nu$ to explain the observations by Hertz and J.J. Thompson; explaining the photoelectric effect. The effect can be described by the Equation 2.1 where E_e the emitted kinetic energy, h is the plank's constant, ν the frequency of the incoming photon and ϕ the material-specific work function (also known as binding energy).

$$E_e = h\nu - \phi \quad (2.1)$$

The equation describes how incident photons on a surface eject photoelectrons, provided the photon energy $h\nu$ exceeds ϕ . This also highlights that the emitted kinetic energy E_e does not depend on the photon flux (photon counts per second). However, the flux increases the total amount of electrons released from the material.

It is then apparent that the binding energy of electrons can be found by irradiating light onto the material and measuring the E_e of photoelectrons. photoemission/photoelectron spectroscopy (PES), is exactly such a technique that leverages this principle to probe the electronic structure of materials. While the above equation describes at what energies electron come out, it does not explain why

A complete treatment of photoemission process needs the quantum theory of light-matter interaction, and this shall be introduced where necessary.

2.1 Light-Matter Interaction

Where a complete description is not always necessary.

2.2 Light Sources

To perform PES, it is imperative to first look at what the source of the electromagnetic radiation is necessary. The source needs to be monochromatic¹ and tunable, so that the only specific electrons are excited, allowing for precise measurement. Many materials of interest have binding energies lying in the UV to X-ray range. To measure not only the electronic states but also their dynamics, it becomes necessary to have a temporally coherent source.

2.2.1 Laser

The most ubiquitous of light sources—the laser—revolutionized experimental science as it enabled a vast range of phenomena to be precisely tested and observed, due to its high spatial and temporal coherence properties. A laser produces its light through a process known as stimulated emission, a process in which electrons, excited to higher energy states by an external energy source, emit photons as they return to lower energy states. These emitted photons, in turn, stimulate other electrons to release additional photons, leading to a cascade of coherent light. A process known as high harmonic generation (HHG) is used to generate ultrafast light of higher energies so that core electrons can be probed.

2.2.2 Synchrotron

Particle accelerators, initially used for high energy physics experiments, which produce radiation as a byproduct of particles accelerating, found their use in spectroscopy. Soon, dedicated facilities providing extremely bright and tunable source of electromagnetic radiation emerged. One class of such facilities are known as Synchrotrons, which are named due to their cyclic design. Typically, bunches of electrons are accelerated by bending magnets, known as Undulators, placed around a storage ring. These electrons are accelerated till they reach appropriate energies and provide a continuous source of radiation. Tangent to the ring, at certain points in the facility are beamlines which direct the radiation to the experimental end-stations, used to perform experiments in fields of life science, material science, and medicine to name a few.

There are several synchrotron facilities around the world, and in Germany such as BESSY II in Berlin, PETRA III at DESY campus in Hamburg.

2.2.3 Free-Electron Laser

While Synchrotrons are excellent sources of light for a multitude of experiments, they have limitations in terms of the temporal coherence of the light produced. free-electron lasers (FELs) are a class of light sources that overcome this limitation.

¹Truly monochromatic radiation is physically impossible, and generally a range is considered.

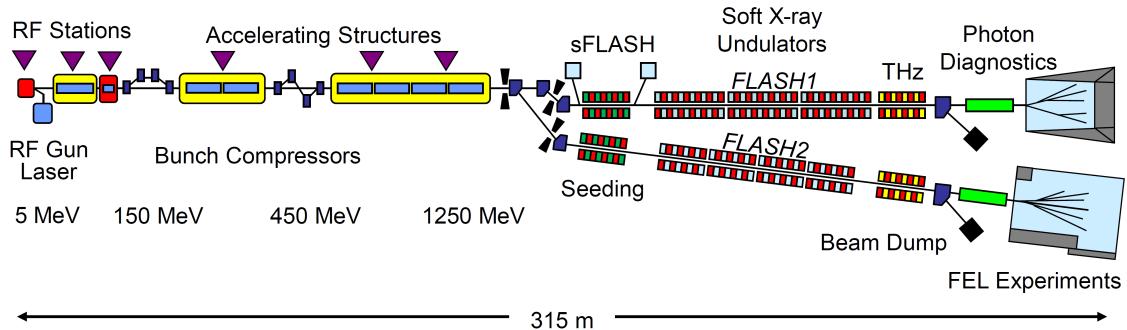


Figure 2.1: Schematic view of Free-electron LAser in Hamburg (FLASH) taken from [4].

FELs are linear particle accelerators that produce radiation through spontaneous emission, an effect explainable only through quantizing light.

Unlike traditional lasers, FELs do not require a traditional optical cavity or gain medium. Instead, they use a linear particle accelerator to produce a bunched electron beam, which is then passed through an undulator. The undulator is a series of magnets that produce a periodic magnetic field, causing the electrons to oscillate and emit radiation. The emitted radiation interacts with the electrons, leading to microbunching, which produces coherent radiation.

The light produced is of high peak brightness and can be compressed to femtosecond pulse durations, making it an ideal source for studying ultrafast phenomena.

Deutsches Elektronen-Synchrotron (DESY) is a national research center in Germany that operates the FLASH facility, which is a FEL that produces soft X-ray radiation. The facility is used for a variety of experiments, including time-resolved studies of materials, and is equipped with a variety of end-stations to perform these experiments.

2.2.4 HEXTOF instrument at FLASH

A variety of photoemission spectrometers can be devised depending on which parameters are varied and what is measured. Naturally, the most basic setup would measure the energy of the emitted electron as described in equation 2.1, but it is also common to measure the surface parallel momentum (k_x, k_y)

the whole process of excitation, transport, and emission is treated as a single coherent process using the formalism of quantum mechanics. This approach incorporates the electronic structure, electron-electron interactions, and the surface barrier in a unified way, and describes the photoemission intensity $I(E, k_x, k_y)$ as:

The high energy X-ray time-of-flight (HEXTOF) instrument is capable of performing time- and momentum-resolved photoemission studies

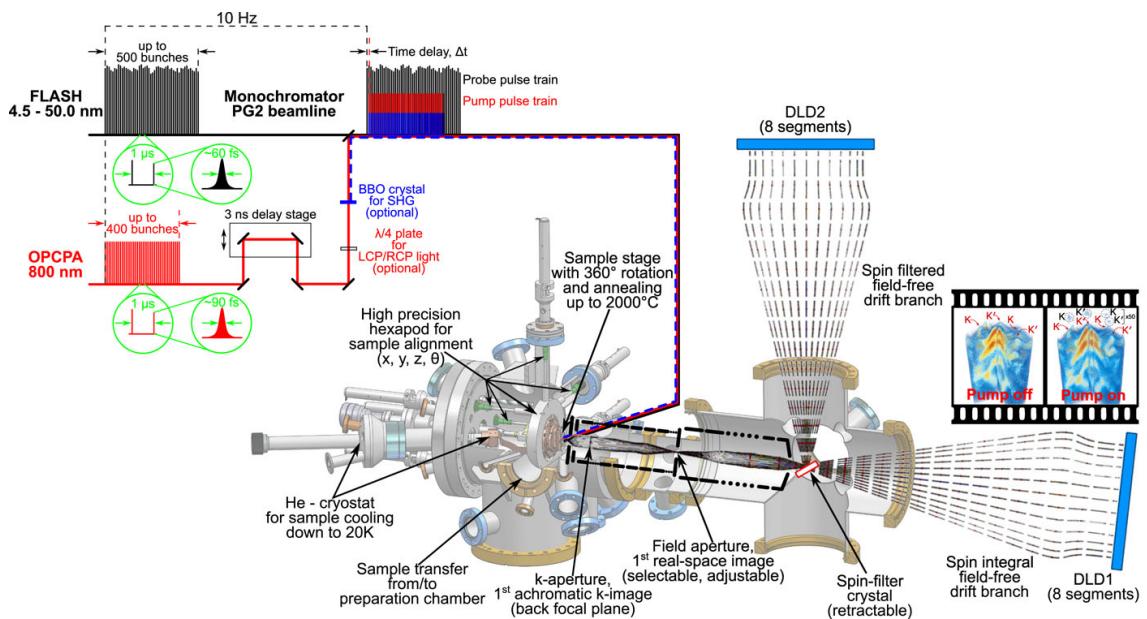
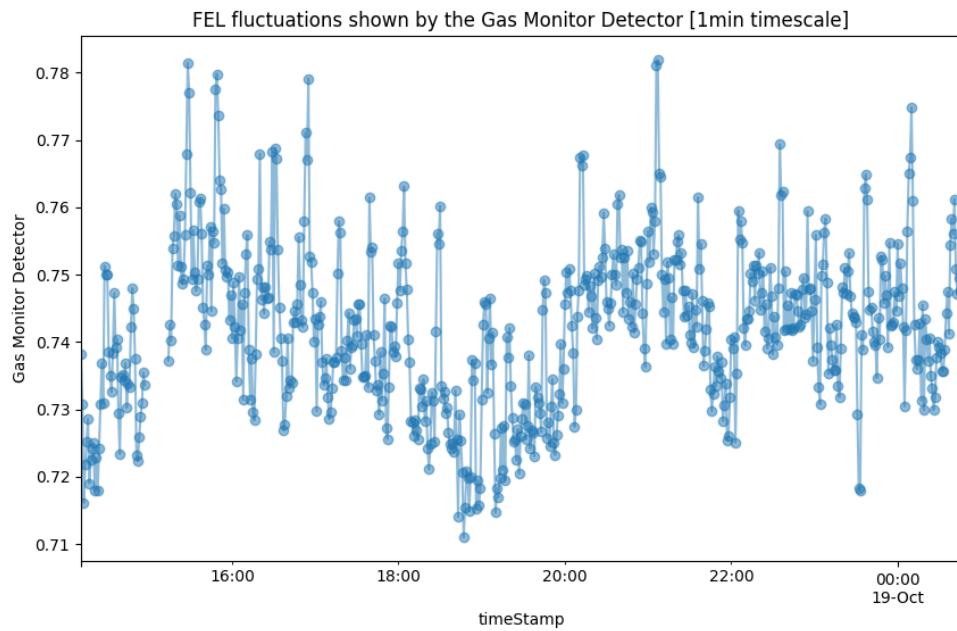


Figure 2.2: HEXTOF taken from [5]

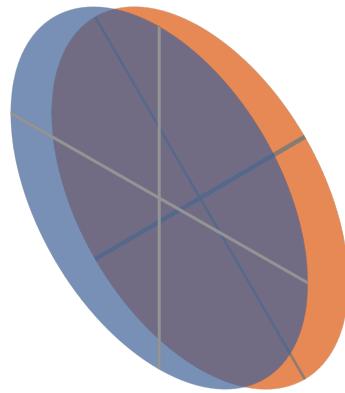


Figure 2.3: Schematic of the 2-layered DLD. The detector is divided into 2 layers, each with 4 sectors.

Delay Line Detectors

The 3D detection scheme used in the HEXTOF instrument consists of a delay line detector (DLD) and a time of flight (TOF) tube. The DLD is made of a micro channel plate (MCP) for electron amplification and meanders to

A Delayline Detector (DLD) operates by using a serpentine wire arrangement (meanders) positioned behind a micro-channel plate (MCP) for electron amplification. When particles (photons, ions, electrons) hit the MCP, they generate an electron cloud that induces electrical pulses in the meanders, allowing for precise time measurement of the hit position. This enables the reconstruction of the impact location and provides absolute time measurements relative to an external clock signal, processed by fast electronics and a time-to-digital converter (TDC).

2.2.5 Binning as means of Imaging

digital imaging from lecture sampling etc Binning is a way to find underlying distribution.

2.2.6 Describing the data GdW, WSe₂, GrIr, and new one

2.2.7 Creating the dataset Corrections Calibrations etc

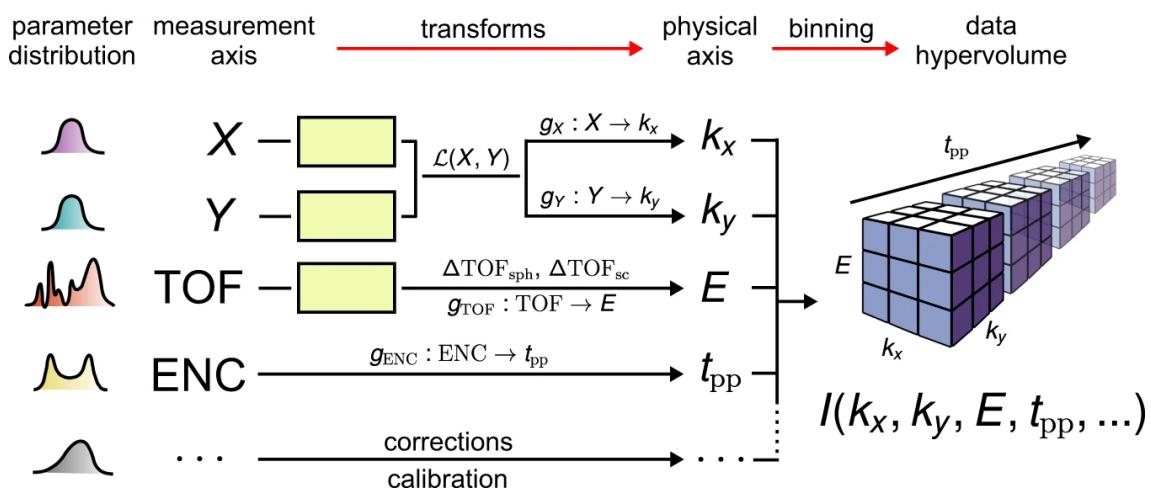
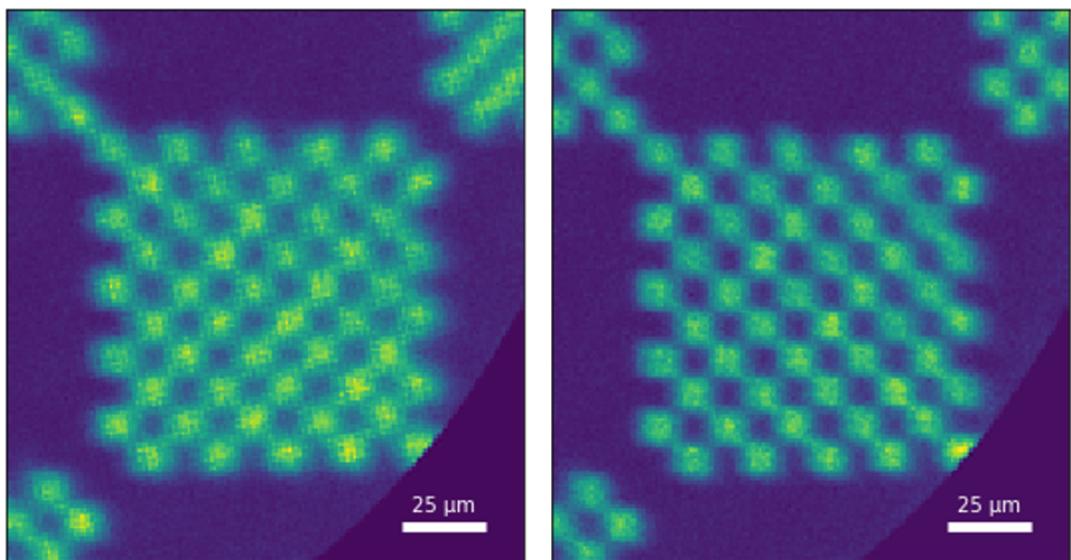


Figure 2.4: MPES taken from [6]

Denoising preliminaries and BM3D

Rather than calling it denoising, better word in image reconstruction because Image Reconstruction:

Purpose: To reconstruct an image from incomplete, noisy, or indirect measurements. This is often used in medical imaging (e.g., MRI, CT scans), computational photography, and computer vision applications.

Reconstruction involves generating a complete image from partial or indirect data, which can include denoising and deblurring as sub-tasks.

3.1 Address reconstruction/denoising schemes

VST with BM3D: BM3D uses collaborative filtering, which is also used in recommender systems [citation need] UNET noise2noise

Why ARPES transfer learning can't work for our case.

Most methods assume a Gaussian noise model, whether in classical or DL approaches. So we need to see the noise type first.

"Owing to solve the clean image x from the Eq. (1) is an ill-posed problem, we cannot get the unique solution from the image model with noise. To obtain a good estimation image , image denoising has been well-studied in the field of image processing over the past several years. Generally, image denoising methods can be roughly classified as [3]: spatial domain methods, transform domain methods, which are introduced in more detail in the next couple of sections." from source

3.2 Metrics

peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), mean squared error (MSE), mean absolute error (MAE), Huber loss, Poisson loss are examples of metrics used to evaluate the quality of the reconstructed image. These metrics measure the similarity between the true image and the reconstructed image, and can be used to compare different reconstruction algorithms.

PSNR: Peak Signal-to-Noise Ratio can be written as:

$$\text{PSNR} = 10 \log_{10} \left(\frac{255^2}{\text{MSE}} \right) \quad (3.1)$$

SSIM: Structural Similarity Index Measure can be written as:

$$\text{SSIM} = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (3.2)$$

3.3 BM3D: Denoising in sparse domain

with and without anscombe on different datasets from flash, lab and fhi Testing s

3.3.1 Anscombe: Variance Stabilization Transform

Characterizing Photon and Photo-electron Statistics

To attempt at reconstructing the truth from a noisy observation, it is essential to understand the underlying noise statistics. Classical and quantum optics provide a comprehensive theoretical foundation to explain photon statistics. For instance, it is well understood that photons from a coherent light source follow a Poisson distribution, often referred to as shot-noise. Whereas, for chaotic (bunched) light, the variance exceeds that compared to their mean, dubbed super-Poissonian. And finally for a squeezed light source, the distribution is sub-Poissonian [7, Chapter 5].

In this chapter, we start with defining the Poisson model for counting photo-electron event data. We examine our data to identify where this model is appropriate and where its limitations become evident. To this end, we introduce some statistical apparatus necessary to estimate parameters and define goodness of model. Further, we explore why the FEL light source (see section ??), exhibits over-dispersed statistics, a characteristic necessitating moving beyond the Poisson model. Accounting for this, we transition to using the Negative Binomial distribution as a more suitable model for the count data. Lastly, we acknowledge that theoretically these models primarily address photon statistics and a more sophisticated treatment building on light-matter interaction might be necessary to connect photon statistics with the statistics of the emitted electrons.

4.1 Poisson distribution as a model for counting statistics

To model counting of independently occurring events, the Poisson distribution is often used. The discrete probability distribution expresses the probability of a given number of events occurring in a fixed interval of time, space or other domains[8]:

Poisson distribution. The probability mass function (PMF) for Poisson distribution $\text{Poi}(\lambda)$ with $\lambda > 0$ is defined as

$$P(n; \lambda) = \frac{\lambda^n e^{-\lambda}}{n!}, \quad n \in \mathbb{N}_0 \quad (4.1)$$

1. $\lambda = E(X) = \text{Var}(X)$

2. Additivity: If $X_1 \sim \text{Poi}(\lambda_1)$ and $X_2 \sim \text{Poi}(\lambda_2)$ are independent Poisson random variables, then the sum $X_1 + X_2$ also follows a Poisson distribution with parameter $\lambda_1 + \lambda_2$

$$X_1 + X_2 \sim \text{Poi}(\lambda_1 + \lambda_2). \quad (4.2)$$

This property is useful when considering counts from multiple independent sources.

4.2 Correlated events in detector

Owing to the 2-layered DLD in our experimental setup, measuring two electrons impinging on locations nearby to one another can be detected. In a single layer DLD, due to the dead-time experienced by each sector after recording, only a single event can be recorded. However, the multi-layer detection scheme comes with a caveat: the layers must be calibrated appropriately otherwise the electron counting routines can report two events for the same electron.

There's a lot of parameters that need to be tested to determine what sort of counting statistics the dataset has.

Controls for the test: lets see

- Total time being looked at (like 1000 s or 20 hours)
 - distribution might change due to overtime FEL intensity changes
- Time bins being used (like 2 s vs 20 s and so on).
 - Seems like distribution changes based on that too
- Looking at individual pixels on X and Y
- Looking at Energy axis as it behaves weird
- Looking at a larger region in X and Y
 - should follow same statistics as single pixels
- Check after removing correlated electrons within each pulse
 - It is possible that electrons are correlated between different pulses because the time delay is long enough. But seems highly unlikely!

—

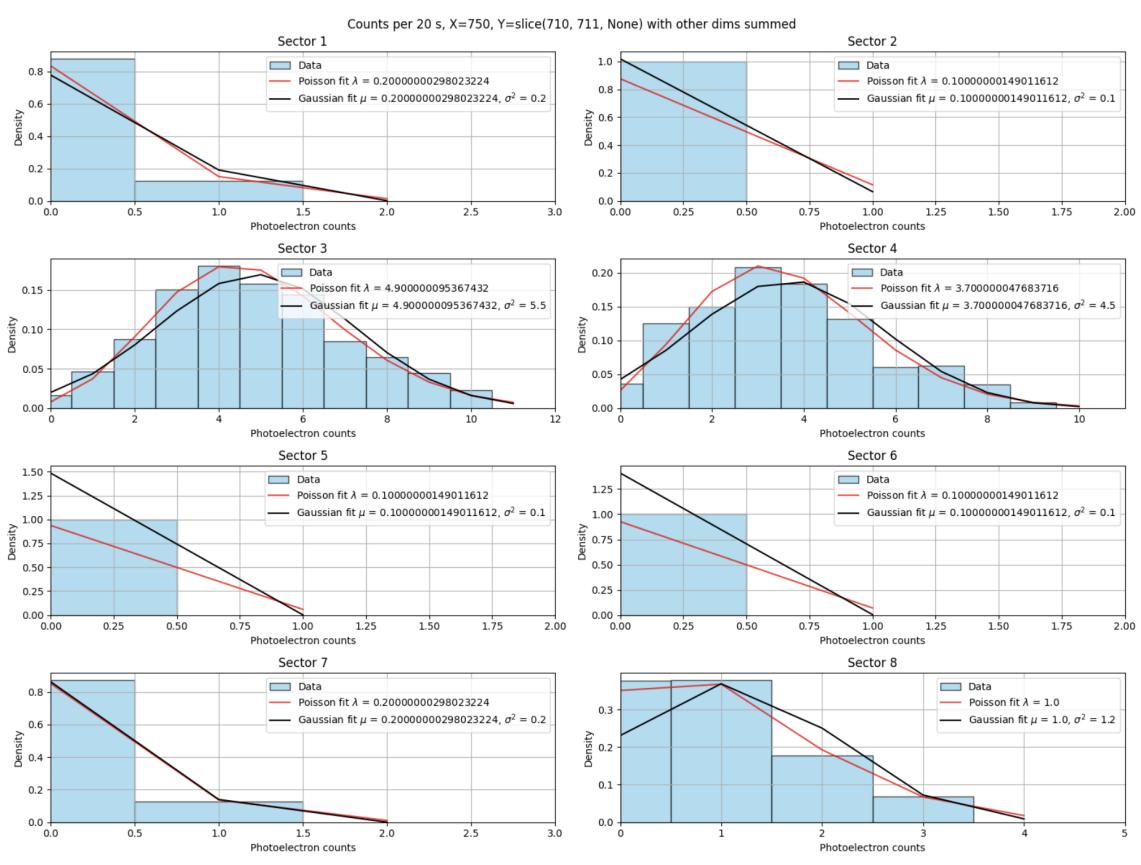


Figure 4.1: Enter Caption

4.2. Correlated events in detector

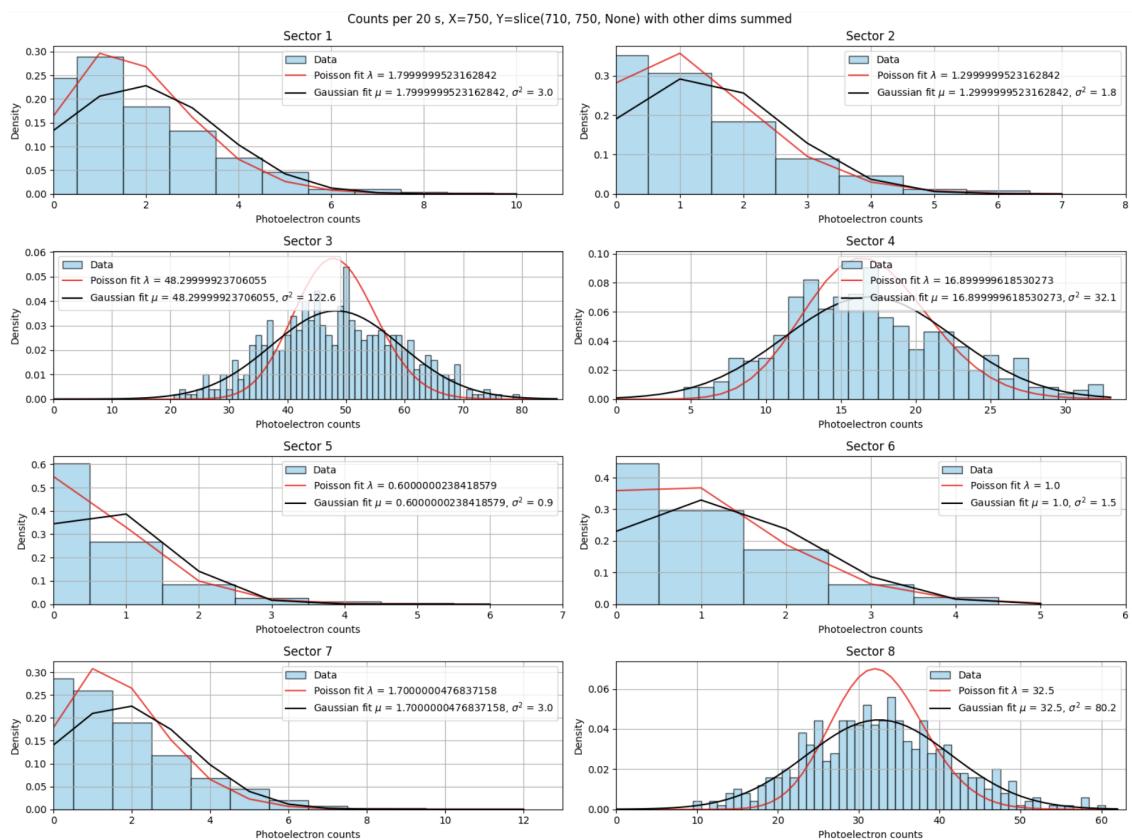


Figure 4.2: Image over 40 pixels and summed over energy

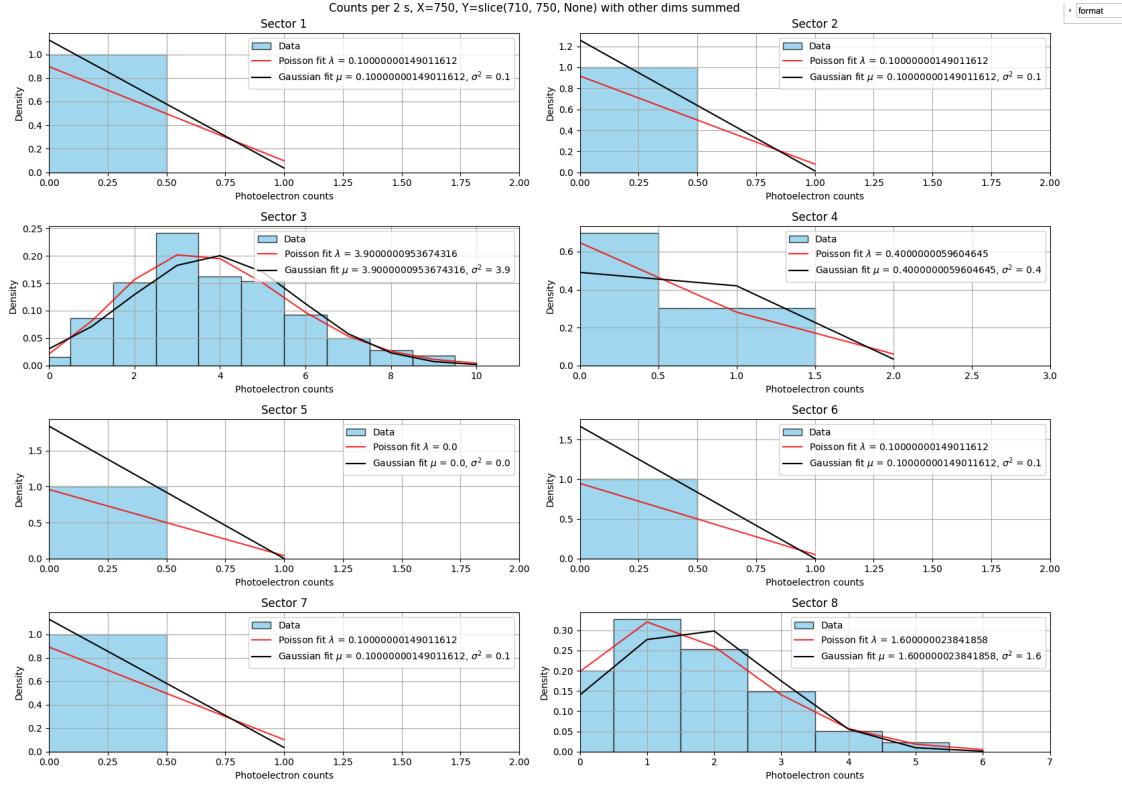


Figure 4.3: Data is filtered here with the KNN and looking at small region

4.2.1 Before filtering

4.2.2 After filtering

Taken from [9] which cites [8]

4.3 Simulate Noise

We simulate the data with Poissonian data.

4.4 Statistical testing

The **Central Limit Theorem** states that for a sequence of i.i.d. random variables X_1, X_2, \dots, X_n with mean μ and variance σ^2 , the normalized sum of these variables approaches a standard normal distribution as n tends to infinity:

$$\frac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0, 1) \quad \text{as } n \rightarrow \infty.$$

This convergence is in distribution.

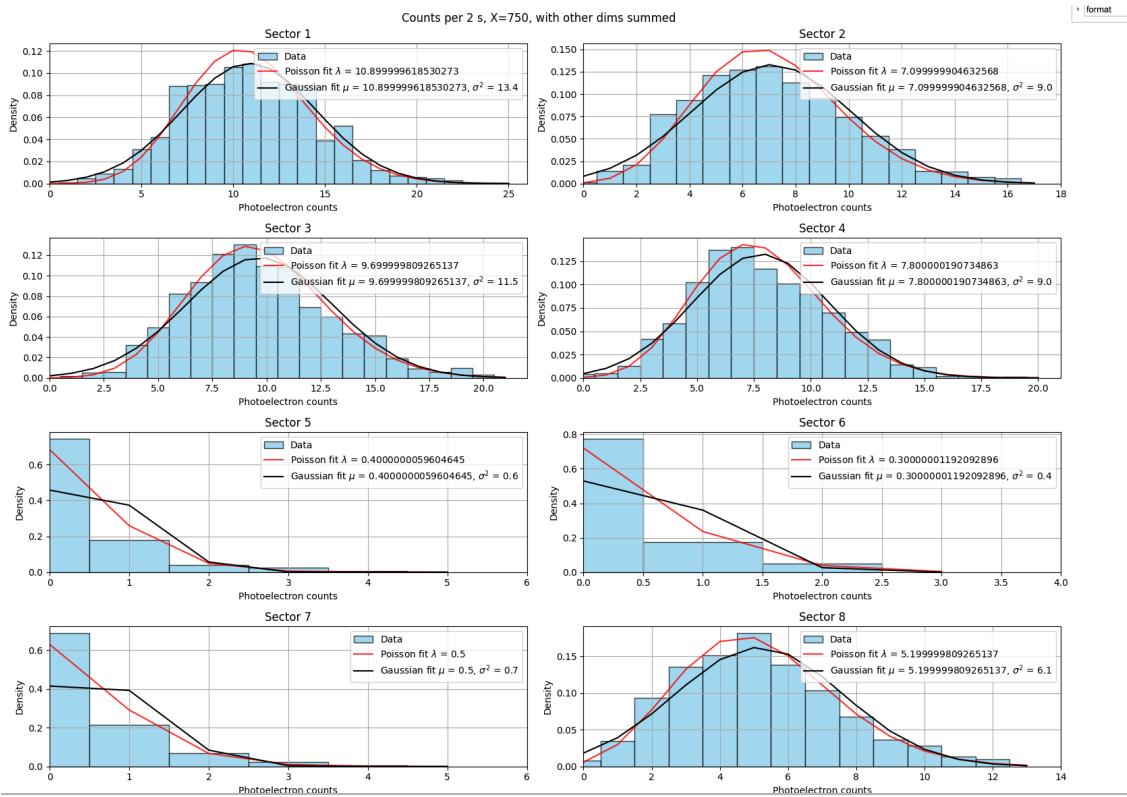


Figure 4.4: Enter Caption

From the law of large numbers, one can show that the relative fluctuations reduce as the reciprocal square root of the number of throws, a result valid for all statistical fluctuations, including shot noise. From Wikipedia

Shot noise exists because phenomena such as light and electric current consist of the movement of discrete (also called "quantized") 'packets'.

Can't correlate FEL intensity with electron counts per pulse as the GMD is before the monochromator.

The physical assumptions which we want to express mathematically are that the **conditions of the experiment remain constant** in time*, and that non-overlapping time intervals are stochastically independent in the sense that information concerning the number of events in one interval reveals nothing about the other. The theory of probabilities in a continuum makes it possible to express these statements directly, but being restricted to discrete probabilities, we have to use an approximate finite model and pass to the limit.

Imagine a unit time interval partitioned into n subintervals of length $1/n$. A given collection of finitely many points in the interval may be regarded as the result of a chance process such that each subinterval has the same probability P_n to contain one or more points of the collection. A subinterval is then either occupied or empty, and the assumed independence of non-overlapping time intervals implies that we are dealing

Example S3.18 (Two-sample Kolmogorov-Smirnov test)

Let $X_1, \dots, X_{n_1} \stackrel{iid}{\sim} F$, and $Y_1, \dots, Y_{n_2} \stackrel{iid}{\sim} G$, for strictly increasing, continuous CDFs F, G .

- ▶ Hypotheses:

$$H_0 : F = G \leftrightarrow H_1 : F \neq G$$

- ▶ Estimate F and G by

$$\begin{aligned}\hat{F}(x) &= \frac{1}{n_1} \sum_{i=1}^{n_1} \mathbb{1}_{(-\infty, x]}(X_i), & x \in \mathbb{R}, \\ \hat{G}(x) &= \frac{1}{n_2} \sum_{i=1}^{n_2} \mathbb{1}_{(-\infty, x]}(Y_i), & x \in \mathbb{R}.\end{aligned}$$

- ▶ Example P5.18: $\|\hat{F} - F\|_\infty \xrightarrow{P} 0$ and $\|\hat{G} - G\|_\infty \xrightarrow{P} 0$ as $n_1, n_2 \rightarrow \infty$

- ▶ Test statistic: $T_n = \|\hat{F} - \hat{G}\|_\infty$.

Figure 4.5: Enter Caption

Poisson data occur in all imaging processes where images are obtained by means of the count of particles, in general photons, arriving in the image domain. In the case of radioactive decay, fluorescence emission or similar phenomena, the arrival of particles is described by a **Poisson process**, i.e. a continuous stochastic process that is a collection of independent random variables $\{N(t); t \geq 0\}$, where $N(t)$ is the number of particles arrived up to time t . The number of particles arriving within a given time interval T is a random variable (r.v.) with a **Poisson distribution** [52, 118], i.e. the probability of receiving n particles is given by

$$p(n) = \frac{e^{-\lambda} \lambda^n}{n!}, \quad n = 0, 1, 2, \dots, \quad (1)$$

where λ , proportional to T , is the expected value of the counts. This statistical model is appropriate to describe data acquired in fluorescence microscopy, emission tomography, optical/infrared astronomy, etc. Even if the energy of the photons (hence their wavelength) is different in the different applications, the statistics of the data is the same. The imaging system, however, significantly varies from an application to another. It is described by a sparse matrix in tomography, and by a convolution matrix in microscopy and astronomy.

Figure 4.6: Enter Caption

with Bernoulli trials: We assume that the probability for exactly k occupied subintervals is given by $b(k; n, P_n)$. We now refine this discrete model indefinitely by letting $n \rightarrow \infty$. The probability that the whole interval contains no point of the collection must tend to a finite limit. But this is the event that no cell is occupied, and its probability is $(1 - p_n)^n$. Passing to logarithms it is seen that this quantity approaches a limit only if np_n from book [8]

4.5 Chi-squared Goodness of Fit Test

We hypothesize that the data follows a certain distribution e.g. Poisson, Normal, Negative Binomial. The Chi-squared Goodness of Fit Test is used to determine if the observed data is consistent with the expected distribution.

4.5. Chi-squared Goodness of Fit Test

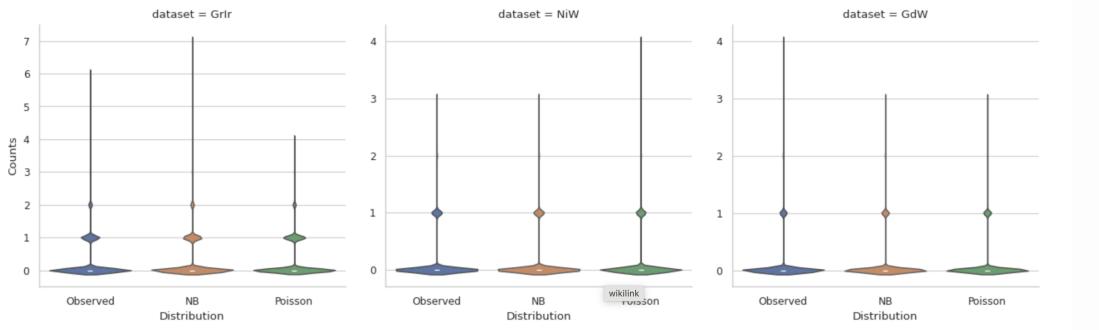


Figure 4.7: Enter Caption

Let $X_1, X_2, \dots, X_n \sim$ i.i.d. F and $Y_1, Y_2, \dots, Y_m \sim$ i.i.d. G , where F and G are strictly increasing continuous CDFs.

The hypotheses for the chi-square test are:

$$H_0 : F = G \leftrightarrow H_1 : F \neq G \quad (4.3)$$

The hypothesis test for the Poisson distribution is the Chi-squared Goodness of Fit Test. The test statistic is given by:

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \quad (4.4)$$

where O_i is the observed frequency and E_i is the expected frequency. The degrees of freedom are given by $k - 1$, where k is the number of bins.

For goodness-of-fit tests, small p-values indicate that you can reject the null hypothesis and conclude that your data were not drawn from a population with the specified distribution. Consequently, goodness-of-fit tests are a rare case where you look for high p-values to identify candidate distributions. (From This website)

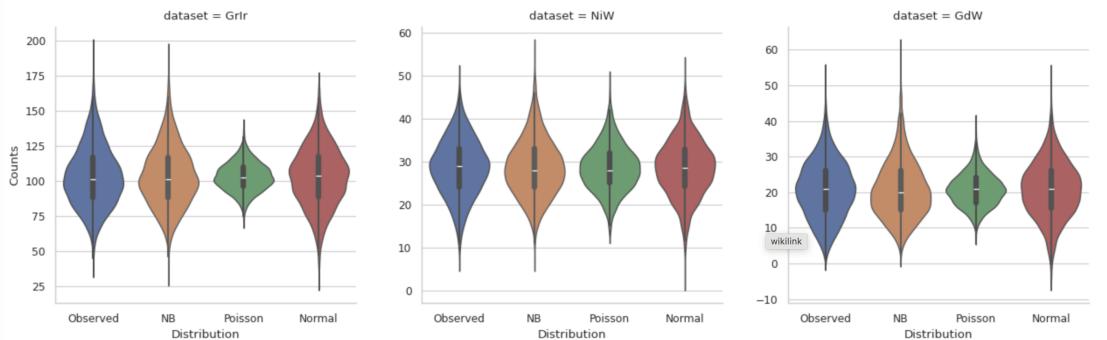


Figure 4.8: ss

4.6 Modeling over-dispersed count data

4.7 The SASE process as an explanation of over-dispersed statistics

Deep Learning

5.1 Statistical Learning

In the section prior, we saw that the photoelectron emission process follow quite complex statistics,

One can do a lot of different learning approaches. We will build towards the deep learning approach. Networks which can deal with image data are generally Convolutional Neural Networks. Looking at research, UNET has been used for image segmentation and denoising, which combines the concept of autoencoders and convolutional neural networks, along with skip connections. We use this approach with the noise2noise framework, which is a deep learning framework for image reconstruction.

For classical learning algorithms, the learning problem is not always realizable, meaning that not always is the ERM Deep learning is just linear sepeartor problem with a non linear function applied to it like RELU

Our aim is to reduce the error between the true image and the reconstructed image. This is a regression problem, where we are trying to learn a function that maps the noisy image to the true image. Mathematically, this can be written as:

ERM: Empirical Risk Minimization

$$\hat{f} = \arg \min_f \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2 \quad (5.1)$$

Training error and empirical risk

Popular loss functions include the mean squared error (MSE), mean absolute error (MAE), and Huber loss. The choice of loss function depends on the noise model and the desired properties of the reconstruction. For example, the MSE is commonly used for Gaussian noise, while the MAE is more robust to outliers.

The MSE is defined as:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2 \quad (5.2)$$

MSE is the squared L2 norm of the difference between the predicted and true values. It is sensitive to outliers and can be dominated by large errors. The MAE is defined as:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |f(x_i) - y_i| \quad (5.3)$$

5.1.1 Optimal Loss Function

Optimal loss can be derived from the distribution model. The Method of Moments and Maximum Likelihood Estimation are two common methods for deriving the optimal loss function.

For Poisson noise, the optimal loss function is the negative log-likelihood of the Poisson distribution. This is because the Poisson distribution is the maximum entropy distribution for count data, and the negative log-likelihood is the maximum likelihood estimator for the Poisson distribution. This can be written as the following optimization problem:

$$\hat{f} = \arg \min_f - \sum_{i=1}^n \log \left(\frac{e^{-f(x_i)} f(x_i)^{y_i}}{y_i!} \right) \quad (5.4)$$

5.1.2 Regularization

There are different ways to regularize the loss function to prevent overfitting. This can be done by adding a penalty term to the loss function. Common regularization techniques include L1 and L2 regularization, which add the absolute value of the weights and the square of the weights to the loss function, respectively. This can be written as:

$$\hat{f} = \arg \min_f \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2 + \lambda \sum_{j=1}^p |\beta_j| \quad (5.5)$$

5.1.3 Optimization

5.2 Noise2Noise: Deep Learning framework

Important to take care not to train with empty data.

Here we take care of data generation. Finite Capture Budget. We use the Graphene on Iridium dataset. The Noisy realizations are just less counts binned. E.g. 96M counts as Noisy and 186M counts as Target. Or 8M counts as Noisy and 96M counts as Target.

5.2.1 Convolutional Neural Networks

5.2.2 Autoencoder

Conclusion and Outlook

We use SASE currently but with seeded FEL, might be more robust poisson statistics. Maybe even sub-poissonian.

Bibliography

- [1] M. Z. Sohail, “Ultrafast dynamic studies in spin-filter material Europium Oxide with Resonant Inelastic X-ray Scattering,” Bachelor’s Thesis, Carl von Ossietzky Universität Oldenburg, 2021.
- [2] W. Ackermann, G. Asova, V. Ayvazyan, *et al.*, “Operation of a free-electron laser from the extreme ultraviolet to the water window,” *Nature Photonics*, vol. 1, no. 6, pp. 336–342, Jun. 2007, ISSN: 1749-4893. DOI: 10.1038/nphoton.2007.76.
- [3] A. Einstein, “Über einen die Erzeugung und Verwandlung des Lichtes betreffenden heuristischen Gesichtspunkt,” *Annalen der Physik*, vol. 322, no. 6, pp. 132–148, 1905. DOI: 10.1002/andp.19053220607.
- [4] B. Faatz, E. Plönjes, S. Ackermann, *et al.*, “Simultaneous operation of two soft x-ray free-electron lasers driven by one linear accelerator,” *New Journal of Physics*, vol. 18, no. 6, p. 062 002, Jun. 2016, ISSN: 1367-2630. DOI: 10.1088/1367-2630/18/6/062002.
- [5] D. Kutnyakhov, R. P. Xian, M. Dendzik, *et al.*, “Time- and momentum-resolved photoemission studies using time-of-flight momentum microscopy at a free-electron laser,” *Review of Scientific Instruments*, vol. 91, no. 1, p. 013 109, Jan. 2020, ISSN: 0034-6748, 1089-7623. DOI: 10.1063/1.5118777.
- [6] R. P. Xian, Y. Acremann, S. Y. Agustsson, *et al.*, “An open-source, end-to-end workflow for multidimensional photoemission spectroscopy,” *Scientific Data*, vol. 7, no. 1, p. 442, Dec. 2020, ISSN: 2052-4463. DOI: 10.1038/s41597-020-00769-8.
- [7] M. Fox, *Quantum Optics: An Introduction* (Oxford Master Series in Physics 15). Oxford: Oxford University Press, 2006, ISBN: 978-0-19-856672-4.
- [8] W. Feller, *An Introduction to Probability Theory and Its Applications* (Wiley Series in Probability and Mathematical Statistics), Third ed. rev. New York Chichester Brisbane [etc.]: J. Wiley, 1968, ISBN: 978-0-471-25708-0.
- [9] M Bertero, P Boccacci, G Desiderà, and G Vicidomini, “Image deblurring with Poisson data: From cells to galaxies,” *Inverse Problems*, vol. 25, no. 12, p. 123 006, Dec. 2009, ISSN: 0266-5611, 1361-6420. DOI: 10.1088/0266-5611/25/12/123006.

Acknowledgements

Contributions

A preliminary analysis during the early works of this thesis titled *Efficient Data Acquisition in Multi-Dimensional Photoemission Spectroscopy using Denoising*, was presented, in the form a of poster, at the DPG Spring Meetings 2024, as well as NanoMat Science Day 2024 at DESY.

Appendices

Transforming Raw Data to structured format

Extract, Transform, Load

Raw data from the experiment is stored in hierarchical data format version 5 (HDF5) files. This includes many Beamline diagnostic information such as beam arrival monitor (BAM), gas monitor detector (GMD), the delay stage readings, the monochromator energy, sample specific information such as extractor voltage, and the three-dimensional electron counting by the DLD. This information is resolved at each bunch of electrons coming from the accelerator called a Train, which are further microbunched into pulses.

The open community of multidimensional photoemission spectroscopy (Open-COMPES) was established to develop tools and infrastructure to make analysis easier. To this end, a modular Python library called **Single Event DataFrame** (SED) was created that provides the entire pipeline from easy data ingestion to common calibration and corrections, Multidimensional binning to create images, and saving the images standard formats, with data provenance.

The data pipeline is designed to extract raw data from HDF5 files, transform the data into a structured format for analysis, and load the transformed data into buffer files. These buffer files are subsequently used for downstream processing, analysis, and visualization. The following sections provide a detailed description of each stage in the ETL process.

Pipeline Overview

The ETL pipeline is divided into several key stages, each critical to the preparation and validation of the data. These stages include:

- **Data Extraction:** Retrieving raw H5 files from experimental runs.
- **Buffer File Creation:** Generating interim buffer files that facilitate further processing.

-
- **Data Transformation and Validation:** Applying domain-specific transformations, such as forward-filling non-electron channels and splitting sector IDs, while validating data integrity against predefined schemas.
 - **Data Structuring:** Organizing the processed data into structured Parquet files suitable for downstream analysis.

Detailed Pipeline Stages

Data Extraction

The initial stage of the pipeline involves the extraction of raw data from H5 files. These files contain experimental data with various channels, such as electron and timed information, stored in a hierarchical structure. The paths to these H5 files are provided as input to the pipeline, along with configuration settings that dictate the subsequent processing steps.

Buffer File Creation

To streamline the data processing, the pipeline first creates buffer files for each type of data (electron and timed dataframes). This is achieved through the `BufferFilePaths` class, which initializes paths for the raw H5 files and corresponding buffer files. The class checks for existing buffer files and determines whether new files need to be generated or existing ones should be reused, based on the `force_recreate` flag.

For each H5 file, the pipeline generates two buffer files:

- **Electron Buffer File:** Contains data relevant to individual electron events.
- **Timed Buffer File:** Aggregates data at pulse and train levels, resolving timing information without individual electron data.

Data Transformation and Validation

Once the buffer files are established, the pipeline proceeds to transform the data. The key transformation steps include:

- **Forward-Filling Non-Electron Channels:** Missing values in non-electron channels are filled using a forward-fill strategy to ensure data continuity.
- **Schema Validation:** The schema of the generated Parquet files is validated against the expected schema derived from configuration files. This ensures that all required channels are present and correctly formatted. The schema check involves:
 - Reading the schema from existing Parquet files.
 - Comparing the actual schema with the expected schema.

- Raising errors if discrepancies are found, prompting a review of the configuration or a forced recreation of buffer files.
- **Splitting Sector ID from DLD Time:** A custom transformation is applied to separate the sector ID from the DLD (Delay Line Detector) time within the electron dataframe. This operation is essential for accurately resolving electron events.

Data Structuring and Finalization

After transforming and validating the data, the pipeline structures it into Parquet files, which are columnar storage formats optimized for analytical queries. The steps involved include:

- **Electron Dataframe Structuring:** The electron-resolved dataframe is processed to drop non-electron data and reset the index. The processed data is then saved as a Parquet file.
- **Timed Dataframe Structuring:** The timed dataframe is derived by aggregating data at pulse and train levels, excluding electron-specific data. This structured data is also saved as a Parquet file.
- **Metadata Generation:** Metadata related to file statistics, filling operations, and schema checks is compiled and saved, providing crucial information for downstream analysis and data auditing.

Mathematical Background

B.1 Measure Space and Measures

Measurable Space.

Let $\Omega = \emptyset$, $P(\Omega)$ the power set of Ω and $\mathcal{A} \subset P(\Omega)$. If the following conditions are met, \mathcal{A} is called a σ -algebra, and the (Ω, \mathcal{A}) pair make up a measurable space.

1. $\Omega \in \mathcal{A}$.
2. If $A \in \mathcal{A}$, then $\Omega \setminus A \in \mathcal{A}$.
3. If $(A_n)_{n \in \mathbb{N}}$ is a sequence of sets where $A_n \in \mathcal{A}$ for all $n \in \mathbb{N}$, then

$$\bigcup_{n \in \mathbb{N}} A_n \in \mathcal{A}.$$

Positive Measure. Let (Ω, \mathcal{A}) be a measurable space as defined in above. A function $\mu : \mathcal{A} \rightarrow [0, \infty]$ is called a *positive measure* if it satisfies the following conditions:

1. $\mu(\emptyset) = 0$, (the measure of the empty set is zero)
2. μ is *countably additive*: For any countable collection of disjoint sets $(A_i)_{i \in \mathbb{N}} \subset \mathcal{A}$, we have

$$\mu \left(\bigcup_{i=1}^{\infty} A_i \right) = \sum_{i=1}^{\infty} \mu(A_i).$$

If these conditions are met, then μ is called a *measure* on the measurable space (Ω, \mathcal{A}) .

B.2 Probability

The theory of probability is necessary to quantify stochastic and uncertain quantities. Throughout this text, it can be seen used the in context of inherently stochastic processes, such as quantum effects, and to quantify uncertainty in measurements.

Probability Measure. A *probability measure* $P : \mathcal{A} \rightarrow [0, 1]$ satisfies all the properties of a positive measure (see Section B.1), with the additional property known as the normalization condition:

$$P(\Omega) = 1$$

This leads to the measure space for probability (or probability space) being (Ω, \mathcal{A}, P) . The probability of an event $A \in \mathcal{A}$ is $P(A)$. The probability measure is hence a real number between 0 and 1 that defines the likelihood of an event to occur.

In the *frequentist* interpretation, probability is defined as the long-run relative frequency of an event occurring in repeated independent trials. It assumes that probabilities are objective and intrinsic properties of the physical world. For example, the probability of getting heads in a fair coin toss is 0.5, meaning that if we were to toss the coin an infinite number of times, half of the outcomes would be heads.

The *Bayesian* interpretation, on the other hand, views probability as a measure of belief or certainty about an event, given the available information. It is inherently subjective and updates as new evidence is introduced. For instance, if we initially believe that a coin is fair, but after observing several tosses we notice a bias, we update our belief (and hence the probability) to reflect the new evidence. This process of updating beliefs is formalized through *Bayes' rule*, which is a cornerstone of Bayesian inference.

Law of Total Probability. The law of total probability provides a way to compute the probability of an event based on a partition of the sample space. If $\{B_i\}_{i=1}^n$ is a partition of the sample space Ω (i.e., $B_i \cap B_j = \emptyset$ for $i \neq j$ and $\bigcup_{i=1}^n B_i = \Omega$), then for any event A :

$$P(A) = \sum_{i=1}^n P(A | B_i)P(B_i). \quad (\text{B.1})$$

This law is particularly useful when dealing with complex events that can be decomposed into simpler, mutually exclusive cases.

B.3 Landau Notation

Landau notation, commonly referred to as Big O notation and its relatives (Big Omega, Big Theta, etc.), is used to describe the asymptotic behavior of functions.

This is particularly important in the analysis of algorithms and in expressing the growth rates of functions.

Big O Notation (\mathcal{O})

B.4 Statistical Inference

B.4.1 Testing/Confidence Intervals

Testing if distribution is poissonian

B.4.2 Optimal parameter estimation

Used in estimating distribution parameters

Deep Learning

C.1 Infrastructure

Model Architecture: UNET 2D and 3D

Trained on A100 GPU 80 GB using Maxwell Cluster at DESY.