Titre: Title:	Accurate Unsupervised Photon Counting from Transition Edge Sensor Signals
Auteur: Author:	Nicolas Dalbec-Constant
Date:	2024
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Dalbec-Constant, N. (2024). Accurate Unsupervised Photon Counting from Transition Edge Sensor Signals [Mémoire de maîtrise, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/61779/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/61779/
Directeurs de recherche: Advisors:	Nicolás Quesada
Programme: Program:	Génie physique

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Accurate Unsupervised Photon Counting from Transition Edge Sensor Signals

NICOLAS DALBEC-CONSTANT
Département de génie physique

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
Génie physique

Décembre 2024

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Ce mémoire intitulé :

Accurate Unsupervised Photon Counting from Transition Edge Sensor Signals

présenté par **Nicolas DALBEC-CONSTANT**
en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
a été dûment accepté par le jury d'examen constitué de :

Sean MOLESKY, président
Nicolás QUESADA, membre et directeur de recherche
Nikola STIKOV, membre

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my supervisor, Nicolás Quesada. His expertise, stubbornness, rigour, and desire to produce scientific work that is truly reproducible have played a crucial role in my personal research development. I am grateful for every opportunity he enabled.

A heartfelt thank you to my colleagues and friends, who have provided both intellectual and emotional support throughout this journey. Their camaraderie has made this experience all the more fulfilling.

Enfin, je suis profondément reconnaissant à ma famille pour leur amour et support inconditionnel, sans lesquels rien de tout cela n'aurait été possible.

RÉSUMÉ

La photonique est une plateforme prometteuse pour construire des systèmes de traitement de l'information quantique à grande échelle [1–5]. Dans beaucoup de ces systèmes, la détection du nombre de photons joue un rôle essentiel en fournissant une ressource clé pour obtenir un avantage quantique. Ces détecteurs peuvent être utilisés, par exemple, pour générer des états quantiques non-gaussiens [6–13], pour échantillonner des distributions de probabilité difficiles à calculer classiquement [14–19] ou pour améliorer la précision d'expériences interférométriques en résolvant directement plusieurs quanta [20, 21]. L'utilisation de détecteurs capables de compter le nombre de photons est intéressante, car un seul détecteur peut mesurer précisément le nombre de photons d'un état quantique [22, 23], sans avoir besoin d'un réseau complexe de détecteurs de seuil avec tous les problèmes de complexité et d'inefficacité que cela peut entraîner [16, 24, 25].

Les capteurs à effet de transition (Transition Edge Sensors ou TES) ont été utilisés pour cette tâche, offrant une résolution élevée sur une large gamme d'énergies. La résolution de 30 photons dans un pulse de lumière a été démontrée [26], bien que les performances soient souvent plus modestes, de l'ordre de 17 photons, avec des techniques plus simples [23].

Les TES exploitent la transition de phase supraconductrice de matériaux photosensibles pour obtenir un calorimètre extrêmement sensible [27]. Pendant l'utilisation des détecteurs, le matériau est refroidi en dessous de sa température critique puis soumis à un courant dans la région de transition entre l'état supraconducteur et l'état normal. L'absorption d'un photon fait alors varier de manière mesurable la résistance du matériau [28, 29]. Ce changement de résistance est ensuite détecté par un amplificateur à faible bruit comme les dispositifs SQUID (Superconducting Quantum Interference Devices), qui permettent aussi de créer des réseaux denses de détecteurs TES grâces au multiplexage [27].

La lecture des signaux de ces détecteurs n'est cependant pas triviale, car la quantité que l'on veut mesurer, l'énergie (ou le nombre de photons à fréquence fixe), se traduit de manière non linéaire dans le signal de tension [30]. Par le passé, on a utilisé l'intégrale (l'aire) du signal pour estimer le nombre de photons [23,31], mais cette méthode perd en résolution pour les grands nombres de photons. Des techniques linéaires comme l'analyse en composantes principales (PCA) ont alors été employées [32], ainsi que des méthodes d'apprentissage automatique inspirées de l'algorithme des K-moyennes pour tenir compte de la statistique de Poisson des sources laser [33].

Avec la popularité croissante de l'apprentissage automatique en traitement du signal [34]

et dans les systèmes quantiques [35], on peut se demander si des méthodes plus avancées permettraient d'améliorer encore la résolution du nombre de photons. Dans cette étude, nous évaluons les performances de plusieurs techniques pour classifier le nombre de photons à partir de signaux TES, en utilisant une mesure de confiance qui quantifie le chevauchement des distributions dans l'espace latent. Nous montrons que les méthodes précédemment utilisées comme l'aire du signal et la PCA peuvent résoudre jusqu'à 16 photons avec une confiance supérieure à 90%, tandis que les techniques non linéaires peuvent atteindre 21 photons avec le même seuil de confiance. Nous présentons aussi des réseaux de neurones pour exploiter la présence de structures locales et augmenter la confiance dans l'attribution du nombre de photons, et démontrons l'avantage de certaines méthodes non linéaires pour détecter et supprimer les signaux aberrants.

ABSTRACT

Photonics is a strong contender for building large-scale quantum information processing systems [1–5]; in many of these systems, photon number detection plays an essential role, serving as a resource for quantum advantage. Photon number resolving detectors can be used, for example, for the heralded generation of non-Gaussian states [6–13], for the sampling of classically-intractable probability distributions [14–19] or for directly resolving multiple quanta improving the Fisher information of interferometric protocols [20, 21].

Transition edge sensors (TES) have been used for this task, offering resolution over a wide energy range. Resolutions up to 30 photons have been demonstrated [26], although this quantity is typically lower, on the order of 17, if more straightforward techniques are used [23].

TESs exploit the superconducting phase transition of photosensitive materials to achieve an extremely sensitive calorimeter [27]. During operation, the material is cooled below its critical temperature and then current-biased to the transition region between its superconducting and normal state. In this region, the temperature increase following the absorption of a single photon leads to a measurable change in the material's resistance [28, 29].

The readout of these devices is non-trivial as the quantity one wants to determine, the energy (or the photon number for a fixed frequency), is reflected in a nonlinear fashion in the voltage signal produced by the detectors' electronics [30]. Historically, the integral (area) of the signals has been used to assign photon numbers [23, 31]. However, distinguishing large photon numbers becomes challenging with this technique. To address this issue, linear techniques such as Principal Component Analysis (PCA) have been used [32]. A machine learning method, adapted from the K-means algorithm to account for the Poissonian statistics of laser sources, has also been developed [33].

With the increased popularity of machine learning in the field of signal processing [34] and quantum systems [35], one might naturally ask whether employing more sophisticated methods could lead to enhanced resolution of photon numbers. In this work, we answer this question by assessing the performance of multiple techniques for photon number classification using TES signals. We do so by considering a confidence metric that quantifies the overlap of the photon number clusters inside a latent space. We demonstrate that for a test dataset, previous methods such as the signal's area and PCA can resolve up to 16 photons with confidence above 90% while nonlinear techniques can resolve up to 21 with the same confidence threshold. Furthermore, we also showcase implementations of neural networks to leverage information within local structures, aiming to increase confidence in real time pho-

ton number assignment. Finally, we demonstrate the advantage of some nonlinear methods to detect and remove outlier signals.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ACRONYMS

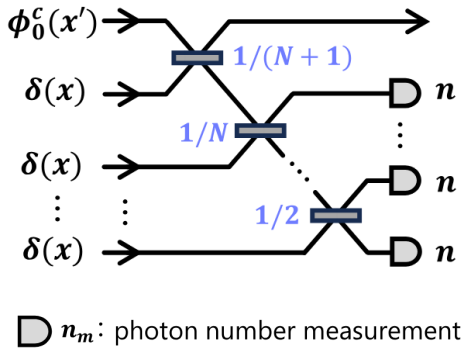
TES	Transition Edge Sensor
PNRD	Photon Number Resolving Detector
NRC	National Research Council Canada
NIST	National Institute of Standards and Technology
FNN	Feedforward Neural Network
RNN	Recurrent Neural Networks
MLP	Multi-layer Perceptrons
VAE	Variational Autoencoder
SPDC	Spontaneous Parametric Down-Conversion

LIST OF APPENDICES

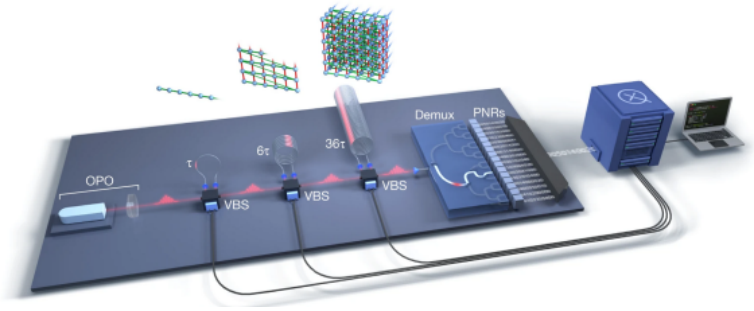
CHAPTER 1 INTRODUCTION

1.1 General Context

Photon-number-resolving detectors (PNRDs) have become a cornerstone technology for advancing quantum photonics, enabling applications across quantum information processing and quantum metrology. These detectors are particularly relevant for tasks such as generating non-Gaussian states, where precise photon statistics underpin the creation of highly entangled resources [6–13] (Fig. 1.1a). Additionally, PNRDs facilitate experiments that challenge classical computational limits, such as sampling complex probability distributions [14–19] (Fig. 1.1b). In metrology applications, these detectors enhance the precision of interferometry by providing richer statistical information, increasing the Fisher information of quantum measurements [20, 21].



(a) Optical circuit for Gaussian breeding, where photon number resolving detectors are used to make indirect measurements [6].



(b) Borealis, a quantum computer from the company Xanadu, composed of dynamically programmable loop-based interferometers. The output of the system is measured using a multiplexed array of transition edge sensors [19].

Figure 1.1 Examples of systems where photon number resolving detectors are used.

Transition edge sensors (TESs) represent one of the most advanced implementations of PNRDs. Operating at the boundary between superconducting and normal states, TESs achieve exceptional sensitivity by leveraging the steep resistance change that occurs near their critical temperature [27]. This unique property enables them to detect individual photons with high efficiency, reaching up to 98% in some configurations [39]. When combined with multiplexed readout techniques using superconducting quantum interference devices (SQUIDs), TES arrays offer a scalable solution for high-fidelity photon-number resolution [28, 29].

Despite these advantages, extracting accurate photon counts from TES signals is a non-trivial problem due to their inherently nonlinear response. Standard techniques, such as integrating the signal, are reliable for low photon counts but face significant limitations in resolving higher photon numbers or analysing light from non-classical sources [23,31]. To address these challenges, researchers have explored advanced methods, including dimensionality reduction techniques like Principal Component Analysis (PCA) [32], and machine learning approaches tailored to photon statistics [33]. By combining physical models with data analysis, this technique aims to unlock the full potential of TESs for quantum applications.

1.2 Research objectives and approach

The objective of this work is to increase the photon number resolution of transition edge sensors. The problem is labelled as an unsupervised classification task where a photon number can be assigned to every signal. The term unsupervised describes that the true number of photons inside each light pulse is unknown. Only structures inside the data provide the information that is used to make the assignment. While simulating samples or considering the statistics of a characterisation source is possible to change the problem to a supervised or semi-supervised problem, the approach is deliberately chosen to minimally include physical models in the way this problem is solved. This is done to increase the performance of existing devices, independent of their specific hardware implementation or experiment.

To find a solution to this unsupervised classification problem, an exploration of dimensionality reduction techniques is done. These techniques, through some process, transform high-dimensional TES signals (many time steps) into a low-dimensional representation while retaining a maximum amount of information about the original signals. Reducing data dimensionality makes the interpretation of large amounts of high-dimensional data easier. Inside this new space, the position of samples describe similarities with other signals, creating regions of high density.

From these low dimensional spaces, the samples are labelled by considering similarities between data inside an abstract space based on some criterion. Only a single algorithm is used in this work (gaussian mixture model) however, a wide range of techniques exist. This process is called clustering and is never done directly on the high-dimensional signals for reasons discussed in Sec. 4.4.2.

While many photon detection experiments can operate with minimal noise and background photons, this is not always guaranteed. The presence of photons hitting the detector at random times is also problematic for a number of experiments. To address these issues, we

expand the problem of photon number assignment to include the presence of noise photons inside the signals. The low-dimensional space generated by some techniques is therefore explored for exotic signal structures.

1.3 Thesis Structure

This thesis is structured as follows:

First, an introduction is provided in Chap. 2 on the context and historical foundation of the previously discussed problems. This is done by discussing photon number resolving detectors and some of their key features. From this general description, the history of transition edge sensors is described. This section includes the development and an overview of the working principle of this type of detector. An exhaustive literature review of the published signal analysis techniques for TES is done. This section aims to establish the landscape of approaches and problems that researchers have encountered while working with TESs. To further describe experimental constraints and context around the photons that are typically detected by TESs, a brief overview of the types of photon sources is done.

In Chap. 3 the description of neural networks is expanded from a computational point of view, including properties that make them useful in the context of photon detection.

In Chap. 4 the full article "Accurate Unsupervised Photon Counting From Transition Edge Sensor Signals" is presented which is at the moment of writing this thesis available online as preprint [40]. When reading the full document, it is suggested to go quickly over the abstract and introduction in Chap. 4 since the subject is already discussed in Chap. 1 and Chap. 2. This article describes the problem of photon-number discrimination in the general setting of unsupervised classification and includes the notion of dimensionality reduction. In Sec. 4.5, a brief overview is provided of the methods used to compute similarities between signals and how to distinguish events that belong to the different photon number classes. The results are presented in Sec. 4.6 using experimental data, followed by a discussion of the use cases of the described methods in Sec. 4.7.

The thesis ends in Chap. 5 by describing how this work contributes to the field of photon detection. A final summary describes the strengths and limitations of the proposed approaches for TES signal analysis and discusses ongoing challenges, highlighting avenues for future research.

CHAPTER 2 LITERATURE REVIEW

This chapter provides the necessary background for the project and article discussed in Chap. 4. First, an overview of Photon Number Resolving Detectors (PNRD) is done, highlighting the unique feature that makes them valuable for various applications. Next, Transition Edge Sensors (TES) are discussed, including their history and operating principles. Additionally, a review of the historical approaches for analysing the voltage response of TES is provided. Finally, an overview of some key concepts of photon statistics is presented to give context for Chap. 4.

2.1 Photon Number Resolving Detectors

Photodetectors function by converting incident light into a measurable electrical signal. When it comes to detecting quantum states, such as single photons, these devices utilize methods, i.e. avalanche breakdown, to achieve significant electrical gain. However, this amplification often introduces noise, leading to a loss of information regarding the precise energy absorbed by the photodetector. As a result, these detectors simply register a "click" upon the detection of at least one photon, lacking the ability to differentiate between multiple photons in a single detection event.

In contrast, more advanced photodetectors can quantify the number of photons present in a given light field. Rather than just signalling a detection, these devices generate an output that reflects the characteristics of the detected photons, allowing them to measure the number of photons in the incoming light and directly assess the statistical properties of the light source. In Fig. 2.1 an example is shown of the voltage responses of a transition edge sensor to photons number 0 to 8. In this case, the peak height allows for the distinction of photon numbers.

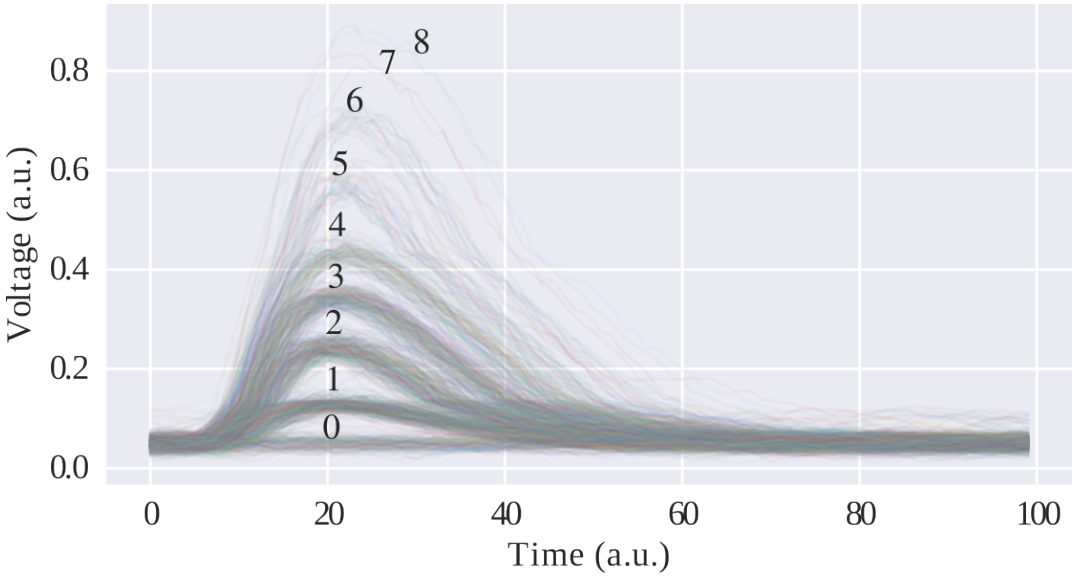


Figure 2.1 Example of 1 024 raw transition edge sensor signals with 100 time steps. The height of the voltage response gives information about the detected photon number following the labels 0 to 8.

2.1.1 Detection Efficiency

Detection efficiency describes all processes that can alter the capacity of a detector to generate a response when hit by a photon with sufficient energy. Often this process is quantified using the beam splitter model of losses which is described by imagining a beam splitter placed before the detector, this is discussed in more detail in Sec. 2.3.2 [41, 42].

Various experimental conditions can affect a detector's efficiency. For example, fibre-coupled detectors are subject to internal losses within the fibre itself and can lose photons at the coupling stage. Additionally, the absorbing surface of the detector has its own quantum efficiency, a metric that reflects how well the detector converts incoming photons into a detectable electrical output, typically in the form of electron current. Quantum efficiency can be formally expressed as the ratio of charge carriers generated (e.g. electrons) to the total incident photons that strike the detector surface [43].

2.1.2 Time Jitter

The timing jitter refers to the time uncertainty at which an incident photon or event is detected. The amount and nature of jitter depend on the detector type [44]. For example, in SNSPDs (superconducting nanowire single photon detectors) the geometry of the detector can play a role in the detection time as the sensor's recorded response time is position

dependent [45]. The hardware used to generate the light and trigger the detector also has an impact on this uncertainty.

Timing jitter is important in applications requiring precise time resolution, as it affects the detector's ability to accurately resolve the timing of photon events.

2.1.3 Noise

Noise refers to unwanted phenomena that introduce uncertainty into a measurement, making the observation process more challenging. It manifests as additional random processes on the desired signal. For photon number resolving detectors, it impacts their ability to accurately discriminate between different photon number states. Noise can originate from intrinsic sources like thermal fluctuations, and environmental disturbances such as electromagnetic interference or mechanical vibrations that can affect a variety of hardware components [27]. It can also be present in terms of unexpected detection events generated inside the experiment.

In signals generated by PNRDs, the noise can appear in a variety of features. For instance, thermal and electrical noise can generate electrical uncertainty. This translates to a broadening of the observed electrical signal values, making the distinction of photon numbers harder. In the case of random detection events, the noise translates into characteristic signal structures at a random time. This kind of structure can occur in the recovery time of the detector, making the interpretation of the signal harder.

2.1.4 Dead Time

Photodetectors respond to a detection event over a period in which the device resets to its initial state, allowing it to be ready for the next detection. In Fig. 2.1, it is possible to see the dead time of a transition edge sensor. In this case, the detector outputs a voltage level that is associated with the vacuum or zero photon scenario. When a photon hits the detector, the voltage first increases then decreases to reach back to its initial state. The dead time refers to the time interval taken by the sensor to come back to its initial state after the detection event. For TESs, the area of the voltage response changes with the energy of the photon, the dead time increases with the number and energy of the photons being detected.

2.2 Transition Edge Sensors

2.2.1 History

Transition Edge Sensors (TESs) are highly sensitive calorimeters capable of detecting low-energy signals, offering exceptional precision. The concept behind TES technology dates back to the late 1930s when D.H. Andrews and A. Goetz independently thought of exploiting the rapid resistance change in a superconducting material during its phase transition to measure power and energy accurately [27, 42, 46]. Andrews' team later demonstrated this idea in 1941 with a TES-based bolometer, applying a current to a tantalum wire in its superconducting transition at 3.2 K to measure infrared signals through changes in resistance [47, 48]. By 1949, TESs were adapted to be used as calorimeters for the detection of alpha particles, extending their use beyond radiometry [49].

TES development faced two key challenges, the first being the need for amplifiers compatible with the low resistance of superconducting thin films. The second difficulty was maintaining the sensor in its transition state after the detection, since it would quickly revert to a fully superconducting or normal state [46, 50]. These issues were resolved in the 1980s with the introduction of Superconducting Quantum Interference Devices (SQUIDs), providing low-noise amplification that enhanced TES stability and efficiency [51, 52].

Originally designed for astronomy, TESs became useful for detecting faint signals from distant cosmic sources [52, 53]. To achieve the necessary sensitivity, TESs are cooled below 100 mK in a dilution refrigerator, with materials like tungsten that are kept at their superconducting transition temperature [41]. Operating in this narrow region under a voltage bias, TESs register small energy deposits from photons or particles as detectable resistance changes.

Today, TES technology plays a crucial role in fields like quantum optics, quantum information science, and photon-number-resolving applications. The journey from early bolometric sensors to advanced photon detectors highlights TESs' vital role in scientific research demanding high precision.

2.2.2 Working Principles

Transition Edge Sensors are composed of three main elements: an absorber, a thermometer, and a cold bath. The absorber intercepts incoming photons and converts their energy into heat, while the thermometer detects the resulting temperature changes to determine the energy of the photons. Operating at cryogenic temperatures, typically in the millikelvin range, the cold bath maintains a stable thermal environment, ensuring high sensitivity and

precise measurements [27, 42, 46].

TESs are voltage-biased, meaning they are maintained at a constant voltage across their superconducting layer. The superconducting state of the TES exists at the edge of the material's superconducting-to-normal transition, making it extremely sensitive to temperature changes. When a photon is absorbed, the temperature increases, causing a rise in the TES's resistance. This change in resistance serves as a feedback mechanism. The system naturally stabilizes the temperature by dissipating the extra heat, ensuring the TES quickly returns to its equilibrium temperature. This self-regulating behaviour allows for precise measurement of photon energies.

The signals generated by the TES are extremely faint, so an ultra-sensitive Superconducting Quantum Interference Device (SQUID) is typically used to amplify them. SQUID amplifiers operate at low temperatures near the TES but interface with room-temperature electronics for readout. However, the low output voltage of a SQUID presents challenges for direct coupling to room-temperature amplifiers, as noise performance can degrade significantly. To address this, specialized techniques, such as impedance matching and additional amplification stages, are used to preserve the signal quality and ensure accurate data collection.

2.2.3 Characteristics

Some features discussed for photon number resolving detectors are presented for TESs in Tab. 2.1.

Table 2.1 Technical features of transition edge sensors.

Property	Value	Reference
Working Temperature	50-100 mK	[23]
Efficiency	98%	[39]
Time Jitter	4 ns	[44]
Dead Time [1]	5-10 μs	[23]
Maximum Photon Number Resolved	33	[26]

2.2.4 TES Noise

Extending on the discussion of PNRD noise features from Sec. 2.1.3, TESs have a number of specific hardware components that generate noise. Excluding the TES itself, noise can be

[1] The reported value is given a single photon, the values changes with the number of photons measured (larger photon number results in longer delay).

associated to many phenomena such as RF pickup, ambient photon shot noise, noise in the amplifier, contact resistance fluctuations, Johnson noise (inside SQUIDs), fluctuations in the temperature bath, and many more [27].

Now, only considering the detector, noise can occur in terms of internal thermal fluctuation noise (ITFN). This noise source arises from internal thermal fluctuations between the distributed heat capacities within the TES [27]. The second type of noise is labelled as "excess electrical" noise and shares the same frequency dependence as the Johnson noise voltage in the sensor. Finally, an "excess low frequency" noise that often correlates strongly with excess electrical noise is present.

These noise features historically have been characterized, in Fig. 2.2 the frequency dependence of the noise is presented. The figure comes from the work of Irwin and Hilton presented in the chapter "Transition Edge Sensors" from the book "Cryogenic Particle Detection" [27].

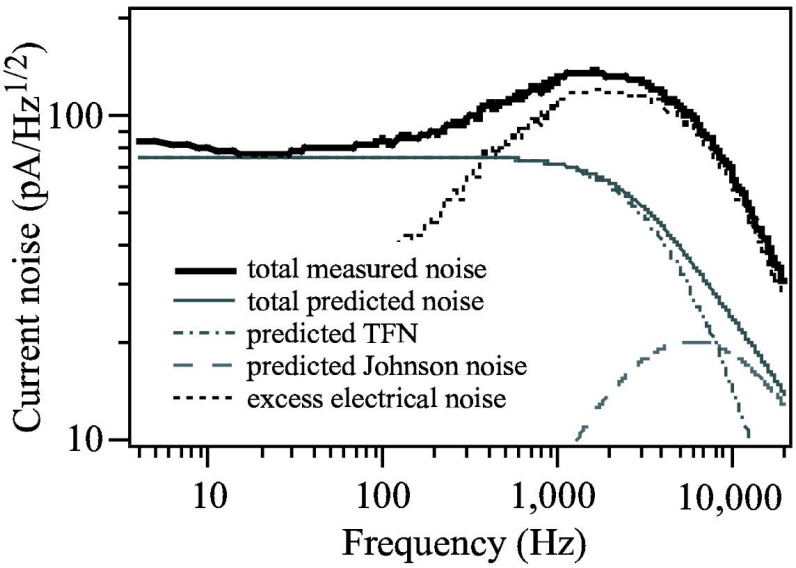
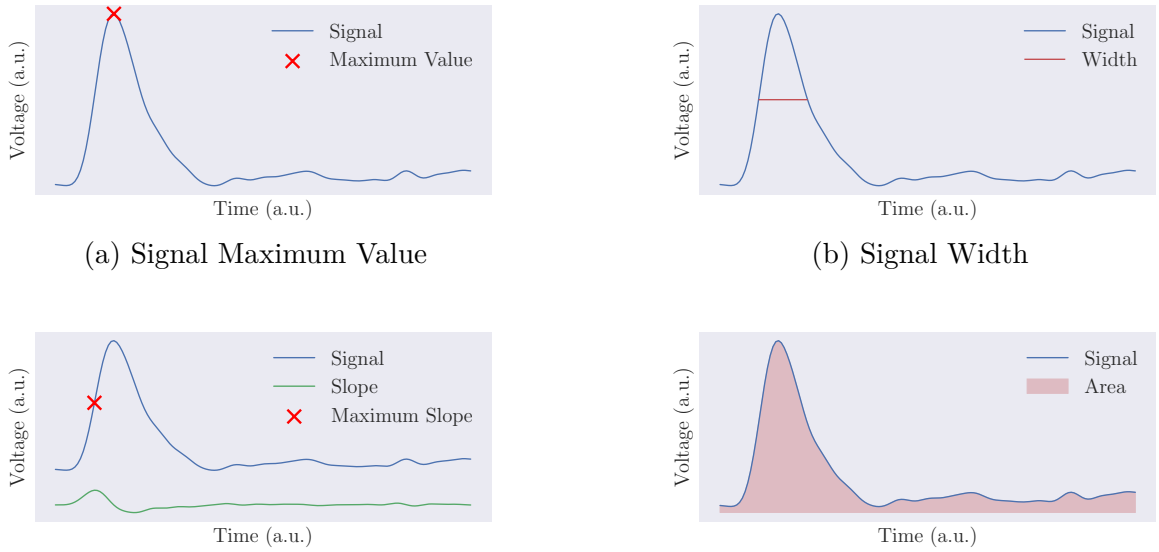


Figure 2.2 Frequency dependence of TES noise features from chapter "Transition Edge Sensors" from the book "Cryogenic Particle Detection" [27].

2.2.5 Signal Analysis of Transition Edge Sensors

To optimize TES systems, research groups initially characterized these systems empirically [54,55] and developed a theoretical understanding of the solid state physics that describe these detectors [56–61]. With strong theoretical foundations and significant advancements in solid-state devices, the photon-number resolution capability has greatly improved. To date, TESs have demonstrated the ability to resolve up to 33 photons in a single light pulse [26].

Historically people have looked at a variety of signal features to describe TES response, the most obvious features being the maximum value of the signal, the signal's pulse width, the pulse maximum slope and the signal's area [23, 26, 42, 62]. From these features, presented in Fig. 2.3, the one that is best suited for the photon number classification task is the signal's area, which relates nonlinearly to the energy of the detected particle.



(a) Signal Maximum Value

(b) Signal Width

(c) Signal Maximum Slope

(d) Signal Area

Figure 2.3 Signal features that have been used or tested to extract photon numbers from Transition Edge Sensors.

Even if this feature allows for the resolution of photon numbers, different groups have considered using more sophisticated approaches to extract the rich information contained in TES signals. For instance, in the realm of signal processing, multiple work discuss filtering techniques to remove the noise [23, 63, 64] present in these signals aiming to increase the performance of the signal's area.

TES have a recovery time on the order of a few microseconds [23] which is extremely slow when compared to other detectors like single photon avalanche photodetectors (SPADs) that can have dead times of a few nanoseconds [65]. To overcome this limitation, researchers have looked into ways to interpret the overlap of voltage response to light pulses [66]. In this work, the use of the least square fit is discussed to distinguish photon numbers of different pulses that happen within the recovery time of the detector.

In the field of machine learning, researchers at the National Institute of Standards and Tech-

nology developed a variation of the extremely popular K-means algorithm [67–69] applied for photon number resolving detectors [70]. This algorithm adds a Poisson likelihood component to the standard K-means objective function. This way the algorithm can be trained taking advantage of the Poissonian statistics of laser sources.

Later, the use of principal component analysis (PCA) to transform signals into a low dimensional representation where photon numbers appear as clusters has been introduced [32]. This work is the first to relate the photon number assignment task to the notion of dimensionality reduction, which is the main subject of the article in Chap. 4. This work also introduces a tomography routine for TES that utilizes the continuous space generated by a dimensionality reduction technique.

The application of neural networks for evaluating the quality of TES signals has shown promising results [71]. In this approach, a variational autoencoder (VAE) is trained on high-quality signals with minimal noise and no anomalies, learning the characteristic shape of TES traces. This neural network architecture compresses the signal into a low-dimensional latent space via an encoder, then reconstructs it through a decoder, attempting to recreate the original signal (with additional considerations). By training the VAE on well-behaved data, this method can serve as an anomaly filter by measuring the reconstruction error for new samples. When the VAE processes signals containing anomalies, the reconstructed output significantly deviates from the original input, effectively flagging the signal as anomalous.

2.3 Photon Statistics

Photon detection is often a tool to measure the outcome of an experiment. The properties of the detected light are influenced by various experimental factors, often tied to the characteristics of commonly used devices. In this section, the characteristics of some devices are presented.

2.3.1 Photon Sources

Lasers are an indispensable tool for modern science, this technology can come in a wide range of technical specifications depending on the specific hardware implementation. These devices are widely used as photon sources in various fields due to their precise control over wavelength, intensity, and coherence. Lasers can be used to produce entangled photons through processes like spontaneous parametric down-conversion (SPDC) [72, 73], they are the standard source type used in interferometric systems [20, 21] and used in many imaging schemes such as fluorescence microscopy [74]. Laser sources have Poisson statistics, meaning

that the probability of generating a photon number n follows

$$P(n) = \frac{e^{-\langle n \rangle} \langle n \rangle^n}{n!}, \tag{2.1}$$

where $\langle n \rangle$ is the average photon number of the light source.

Thermal sources or thermal states refer to a type of light that comes from different systems. Only single-mode thermal light is considered in this work, meaning that a single frequency is present in the system. This translates into a particular type of source which follows a Geometric distribution

$$P(n) = \frac{\langle n \rangle^n}{(1 + \langle n \rangle)^{n+1}}. \tag{2.2}$$

The thermal states are well described by the black-body radiation but can be generated using other systems. For instance, thermal light can be generated by sending a coherent light source that acts as a pump through a ring resonator. The resulting resonances are described by a series of intensity peaks in frequency space on both sides of the pump light. Each of these resonances follows thermal statistics.

Another interesting type of photon sources are single-photon sources that attempt to generate a single-photon on demand. The ideal photon number distribution would follow

$$P(n) = \delta_{n,1}, \tag{2.3}$$

where the delta function indicates that the source always generates a single photon. In practice, true single-photon sources are challenging to realize, as even the best single-photon sources may occasionally emit zero or more than one photon due to imperfect preparation, losses, or background light.

2.3.2 Optical Losses

Optical loss refers to the reduction in the intensity or energy of an optical field as it propagates through a medium or optical system. This loss is typically modelled as a beam splitter interaction, where part of the optical field couples to a vacuum mode. In the photon number picture, this model quantifies the process of losing photons. It is often convenient to describe this model using a loss matrix, where every element gives the probability that n photons are

measured considering k input photons following

$$L(\eta)_{n,k} = \eta^n (1-\eta)^{k-n} \frac{k!}{(k-n)!n!}. \tag{2.4}$$

In this equation, η describes the probability of a photon going through the beam splitter to the detector, which tunes the amount of loss in the system. For edge case $\eta = 0$, all the light is directed to the vacuum mode of the beam splitter and the probability of observing a photon at the detector becomes 0. In opposition, for $\eta = 1$ no loss is present in the system and all the photons from the system would reach the detector.

2.3.3 Second-Order Correlation Function

In quantum optics, the second-order correlation function $g^{(2)}(\tau)$ characterizes the statistical properties of light by quantifying the correlation between intensity measurements at different times. It is defined as

$$g^{(2)}(\tau) = \frac{\langle \hat{I}(t)\hat{I}(t+\tau)\rangle}{\langle \hat{I}(t)\rangle^2}, \tag{2.5}$$

where $\langle \cdot \rangle$ denotes an average, and $\hat{I}(t)$ represents the field intensity operator, which is proportional to the square of the electric field operator $\hat{E}(t)$, and τ is the time delay between the intensity measurements.

In practice, $g^{(2)}(\tau)$ is often measured using a Hanbury Brown and Twiss (HBT) setup [75]. This experiment does not require photon number resolution, the light is split into two paths, and coincident photon detection events are recorded as a function of the time delay τ.

In the context of PNRDs, it is often more convenient to describe the second-order correlation in the photon-number basis as

$$g^{(2)}(0) = \frac{\langle n^2\rangle - \langle n\rangle}{\langle n\rangle^2}, \tag{2.6}$$

Where $\langle n\rangle$ is the mean photon number of the source, and $\langle n^2\rangle$ can be expressed in terms of the photon number variance σ^2 as $\langle n^2\rangle = \sigma^2 + \langle n\rangle^2$. For classical light, $g^{(2)}(\tau) \geq 1$, with $g^{(2)}(0) = 1$ for coherent states (e.g., laser light) and $g^{(2)}(0) = 2$ for thermal light. In quantum optics, $g^{(2)}(0) < 1$ indicates non-classical behaviour, such as photon antibunching observed in single-photon sources [76].

CHAPTER 3 NEURAL NETWORKS FOR PHOTON DETECTION

In this chapter, the objective is to describe key features of neural networks to provide a general understanding. This is done by introducing how neural networks are built and trained, this section continues with a discussion on the properties of some types of networks, and concludes by highlighting the training process and considerations.

3.1 Neural Networks

In this chapter, the notation is slightly changed from Chap. 4 to follow notation from the field of neural networks. Let us consider a data matrix $\boldsymbol{X} \in \mathbb{R}^{t \times u}$ that stores u samples x_i of size t (transpose from Chap. 4). The problem that networks attempt to solve is to approximate a function $f(\boldsymbol{X})$ as closely as possible by training a function $F(\theta', \boldsymbol{X})$ with parameters θ'. To achieve this goal, a function F is trained using a procedure that automatically tunes parameters θ', resulting in a state of F that can achieve the desired operations.

3.1.1 Components

Neural networks are often presented as graphs making the visualization of operations easier, following this notation a standard neural network architecture is shown in Fig. 3.1. In this figure, it is possible to see that the NN is composed of a series of layers with nodes (columns of circles), where the first describes the input and the last one the output. This graph describes a series of operations that typically goes from left to right. In these layers, each node is called a neuron, since these elements were developed to resemble the computation of biological neurons [77]. In their simplest form, neurons with a single input x compute $y = xw + b$, where w and b are trained scalar parameters and respectively named weight and bias. Typically, once y is computed, it goes through an activation function which introduces a nonlinearity in the computation. Activation functions are essential components since they significantly increase the capacity of neural networks to approximate nonlinear data structures. They can follow a number of operations that often resemble a step function, in Tab. 3.1 a few examples are presented. Let us transition from a visual discussion to a more mathematical perspective on neural networks by describing in more detail the network presented in Fig. 3.1. The discussed network belongs to an extremely common class of NN called feedforward neural network (FNN) since the flow of operations follows a single direction (left to right). The action of the network is described by following an input data matrix $\boldsymbol{X} \in \mathbb{R}^{t \times u}$ through the

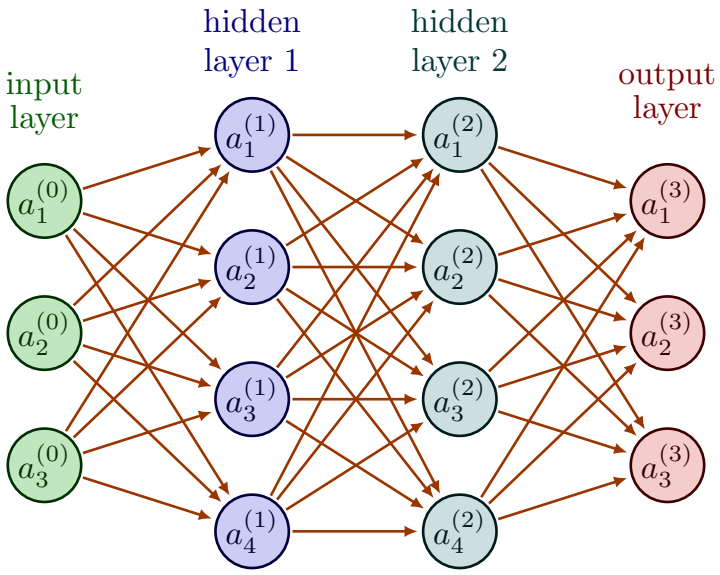


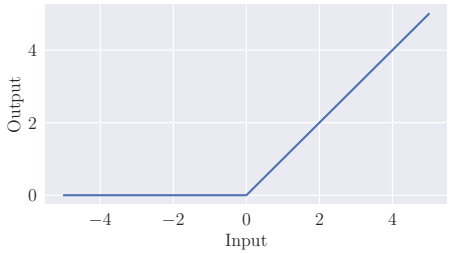
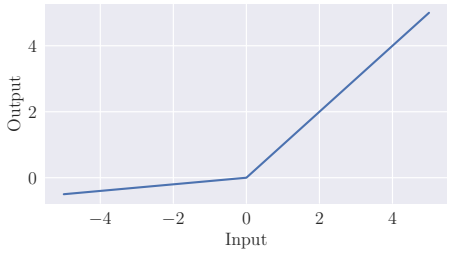
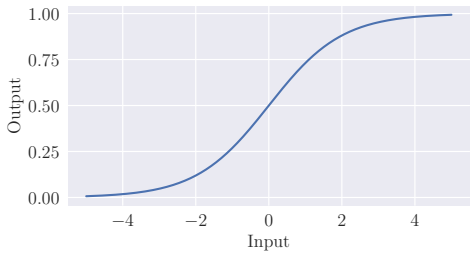
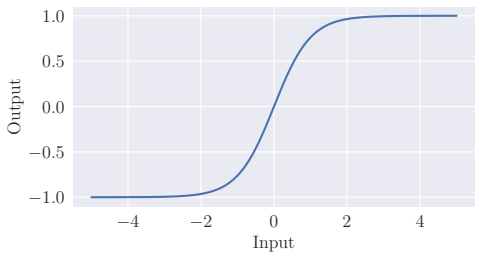
Figure 3.1 Example of a neural network graph with 2 hidden layers, each containing 4 neurons. Each orange arrow represents a weight connection between two neurons. In the representation, the first layer is the input (green), the second and third are hidden layers (blue) and the last one is the output (red)

neural network; this transformation is referred to as a forward pass. A simple architecture is considered which only contains dense layers, also called fully connected linear layers [78]. In this context, the input data $\boldsymbol{X}$ is the input layer and is connected to the first hidden layer, every connection in Fig. 3.1 describes the presence of a weight between two layers. Since fully connected layers are considered, every node of a layer is connected to every node of the previous layer. The action of the first hidden layer on $\boldsymbol{X}$ is given by a function $F_1(\cdot)$, defined (following colours from Fig. 3.1) as

$$\overset{\boldsymbol{A}^{(1)}}{\begin{bmatrix} a_{11}^{(1)} & \cdots & a_{1u}^{(1)} \\ a_{21}^{(1)} & \cdots & a_{2u}^{(1)} \\ a_{31}^{(1)} & \cdots & a_{3u}^{(1)} \\ a_{41}^{(1)} & \cdots & a_{4u}^{(1)} \end{bmatrix}} = \sigma_1 \left(\overset{\boldsymbol{W}^{(1)}}{\begin{bmatrix} w_{11}^{(1)} & w_{12}^{(1)} & w_{13}^{(1)} \\ \vdots & \vdots & \vdots \\ w_{41}^{(1)} & w_{42}^{(1)} & w_{43}^{(1)} \end{bmatrix}} \overset{\boldsymbol{X}=\boldsymbol{A}^{(0)}}{\begin{bmatrix} | & \cdots & | \\ x_1 & \cdots & x_u \\ | & \cdots & | \end{bmatrix}} + \overset{\boldsymbol{b}^{(1)}}{\begin{bmatrix} b_1 \\ \vdots \\ b_4 \end{bmatrix}} \right) = F_1(\boldsymbol{X}). \tag{3.1}$$

For every object, we use the superscript (i) to describe the ith layer. In this equation, $\boldsymbol{W}^{(1)} \in \mathbb{R}^{l_1 \times l_0}$ is a weight matrix that describes the connections between the input layer of size l_0 and the first hidden layer of size l_1. Weight matrices have elements $w_{jk}^{(i)}$ that describe the weight connecting the kth neuron in the $(i-1)$th layer to the jth neuron of the ith layer. While, $\boldsymbol{b}^{(1)} \in \mathbb{R}^{l_1 \times 1}$ is a bias vector, let us highlight that the addition considers broadcasting,

Table 3.1 Example of commonly used activation functions in neural networks. The name, plot, function, and range of the different functions are presented.

Name	Plot	Function	Range
ReLU		$\sigma(x) = \begin{cases} 0 \text{ if } x \leq 0 \\ \text{x if } x > 0 \end{cases}$	$[0, \infty)$
Leaky ReLU		$\sigma(x) = \begin{cases} \alpha \text{ x if } x \leq 0 \\ \text{x if } x > 0 \end{cases}$	$(-\infty, \infty)$
Logistic		$\sigma(x) = \dfrac{1}{1 + e^{-x}}$	$(0, 1)$
Tanh		$\sigma(x) = \dfrac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$	$(-1, 1)$

therefore the bias vector is added to every column of $\boldsymbol{W}^{(1)}\boldsymbol{X}$. Additionally, σ_1 describes the element-wise activation function for the first layer. An important property of the output of the first hidden layer $\boldsymbol{A}^{(1)} \in \mathbb{R}^{l_1 \times u}$ is that it does not have to be the same dimensions as $\boldsymbol{X}$, the input.

Following similar a structure, the second hidden layer takes $\boldsymbol{A}^{(1)}$ (first hidden layer) as an

input and the output layer takes $\boldsymbol{A}^{(2)}$ to give the following computations

$$\overset{\boldsymbol{A}^{(2)}}{\begin{bmatrix} a_{11}^{(2)} & \cdots & a_{1u}^{(2)} \\ a_{21}^{(2)} & \cdots & a_{2u}^{(2)} \\ a_{31}^{(2)} & \cdots & a_{3u}^{(2)} \\ a_{41}^{(2)} & \cdots & a_{4u}^{(2)} \end{bmatrix}} = \sigma_2 \left(\overset{\boldsymbol{W}^{(2)}}{\begin{bmatrix} w_{11}^{(2)} & w_{12}^{(2)} & w_{13}^{(2)} & w_{14}^{(2)} \\ \vdots & \vdots & \vdots & \vdots \\ w_{41}^{(2)} & w_{42}^{(2)} & w_{43}^{(2)} & w_{44}^{(2)} \end{bmatrix}} \overset{\boldsymbol{A}^{(1)}}{\begin{bmatrix} a_{11}^{(1)} & \cdots & a_{1u}^{(1)} \\ a_{21}^{(1)} & \cdots & a_{2u}^{(1)} \\ a_{31}^{(1)} & \cdots & a_{3u}^{(1)} \\ a_{41}^{(1)} & \cdots & a_{4u}^{(1)} \end{bmatrix}} + \overset{\boldsymbol{b}^{(2)}}{\begin{bmatrix} b_1^{(2)} \\ \vdots \\ b_4^{(2)} \end{bmatrix}} \right) = F_2(\boldsymbol{A}^{(1)}), \tag{3.2}$$

$$\overset{\hat{\boldsymbol{Y}}}{\begin{bmatrix} | & & | \\ y_1 & \cdots & y_u \\ | & & | \end{bmatrix}} = \sigma_3 \left(\overset{\boldsymbol{W}^{(2)}}{\begin{bmatrix} w_{11}^{(3)} & w_{12}^{(3)} & w_{13}^{(3)} & w_{14}^{(3)} \\ \vdots & \vdots & \vdots & \vdots \\ w_{31}^{(3)} & w_{32}^{(3)} & w_{33}^{(3)} & w_{34}^{(3)} \end{bmatrix}} \overset{\boldsymbol{A}^{(2)}}{\begin{bmatrix} a_{11}^{(2)} & \cdots & a_{1u}^{(2)} \\ a_{21}^{(2)} & \cdots & a_{2u}^{(2)} \\ a_{31}^{(2)} & \cdots & a_{3u}^{(2)} \\ a_{41}^{(2)} & \cdots & a_{4u}^{(2)} \end{bmatrix}} + \overset{\boldsymbol{b}^{(3)}}{\begin{bmatrix} b_1^{(3)} \\ \vdots \\ b_3^{(3)} \end{bmatrix}} \right) = F_3(\boldsymbol{A}^{(2)}). \tag{3.3}$$

The total feedforward neural network can therefore be described as nested functions

$$\hat{\boldsymbol{Y}} = F_3(F_2(F_1(\boldsymbol{X}))), \tag{3.4}$$

where every function describes a layer. In this picture, it becomes possible to see how graphs offer a visual interpretation for the flow of operations. We also see from this example how a neural network can be a single layer of trainable parameters, and how we can generalize the structure for different types of layers and activation functions. Adding on an earlier comment on dimensionality, from the matrix and graph representation it becomes possible to see how a neural network can have an input and output of arbitrary size.

3.1.2 Universal Approximation Theorem

A characteristic that makes neural networks powerful tools in a wide variety of fields is their capacity to approximate almost any function. In the previous section, we established that NNs are composed of a variety of tunable operations. The size and number of these building blocks limits the set of possible operations that the network can reproduce. Following this intuition, important results in the field of machine learning are the universal approximation theorems. These theorems typically demonstrate that for a family of neural networks and a function space there exists a neural network function F that can arbitrarily well approximate f for a given criterion [79, 80].

We do not go into the derivation of these theorems, however a key characteristic of these

proofs is that they only demonstrate approximation capabilities of the networks, never how to find these solutions. This is a key distinction, since finding these solutions is often not guaranteed and can be challenging.

3.1.3 Loss Functions

A loss function describes how close the neural network is from f (the true transformation). It is by optimizing this quantity that the network parameters are tuned, and its definition is context dependent. Most problems solved using neural networks fall into two categories: regression and classification.

Regression problems describe situations where we aim to approximate a continuous function based on input variables. Often distance based losses are used for these problems, we list a few in Tab. 3.2. In Tab. 3.2, losses are described in terms of the absolute error $|y_i - \hat{y}_i|$ between the neural network output $\hat{y}_i$ and the true value y_i. On the other hand, in classification tasks

Table 3.2 Example of commonly used loss functions in regression problems. The name and function of the different functions is presented.

Name	Function
Mean Absolute Error	$\text{MAE} = \frac{1}{u}\sum_{i=1}^{u} \lvert y_i - \hat{y}_i \rvert$
Mean Squared Error	$\text{MSE} = \frac{1}{u}\sum_{i=1}^{u} \lvert y_i - \hat{y}_i \rvert^2$

samples belong to a set of classes, the goal becomes to predict the specific class associated with any sample. In this case, networks approximate discrete solutions that describe the classes.

3.1.4 Backpropagation

The standard approach to optimize the parameters of a neural network is to use a gradient descent. For an individual parameter, this iterative process updates the parameter to converge towards a value that optimizes the selected loss function following

$$\theta'_{i+1} = \theta'_i - \alpha \frac{\partial L(\hat{y}_i, y_i)}{\partial \theta'}, \tag{3.5}$$

where $\partial L(\hat{y}_i, y_i)/\partial \theta'$ is the partial derivative of the selected loss $L(\hat{y}_i, y_i)$, and α the learning rate. The main computational cost in this optimization comes from the evaluation of the derivative for every parameter in the network. To make this process efficient, backpropagation

makes use of the chain rule to avoid having to compute multiple times the same quantities. For example, for the network in Fig. 3.1 the derivative of the loss in terms of the input can be written as

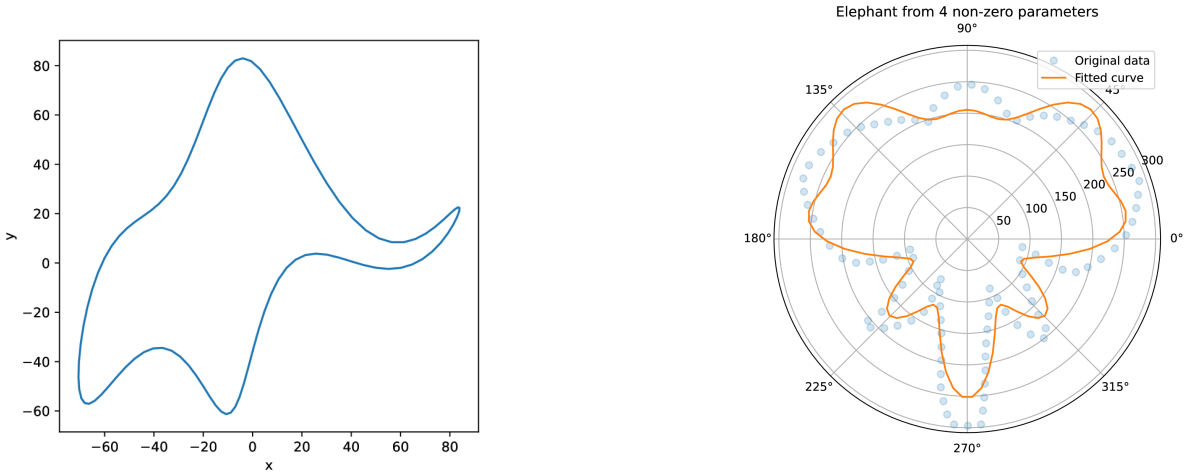
$$\frac{\partial L}{\partial \boldsymbol{X}} = \frac{\partial L}{\partial \boldsymbol{A}^{(3)}} \frac{\partial \boldsymbol{A}^{(3)}}{\partial \boldsymbol{Z}^{(3)}} \frac{\partial \boldsymbol{Z}^{(3)}}{\partial \boldsymbol{A}^{(2)}} \frac{\partial \boldsymbol{A}^{(2)}}{\partial \boldsymbol{Z}^{(2)}} \frac{\partial \boldsymbol{Z}^{(2)}}{\partial \boldsymbol{A}^{(1)}} \frac{\partial \boldsymbol{A}^{(1)}}{\partial \boldsymbol{Z}^{(1)}} \frac{\partial \boldsymbol{Z}^{(1)}}{\partial \boldsymbol{X}}, \tag{3.6}$$

where $\boldsymbol{Z}^{(i)}$ is a weighted output at layer i before the activation function step is applied. In this equation it becomes apparent that the derivative of the loss in terms of any element can be obtained by writing a chain rule that is already contained in Eq. 3.6. Therefore, by computing the intermediary elements of $\partial L/\partial \boldsymbol{X}$ it is possible to extract all the desired derivative in a single efficient pass.

3.2 Types of Neural Networks

The size and type of neural network plays a major role in determining its ability to approximate an arbitrary function and solve problems efficiently. While neural networks are often praised for their surprising capacity to approximate complex functions, there are fundamental limits to this capability that depend on the network architecture and size.

An interesting illustration of this idea is captured by the concept of Von Neumann's elephant, which originates from the famous quote "With four parameters, I can fit an elephant, and with five, I can make him wiggle his trunk" [36, 81]. This highlights the potential of mathematical models to capture seemingly complex phenomena with limited parameters. On the other hand, this idea can serve as a reminder of the inherent limitations in such approximations by looking at the actual elephants presented in Fig. 3.2 (see Ref. [82] for an interactive elephant). While these two examples are impressive, Fig. 3.2a is a crude approximation and Fig. 3.2b relies on symmetries, which is not always possible.



(a) (b)

Figure 3.2 Example of different implementations of Von Neumann's elephants [36, 37].

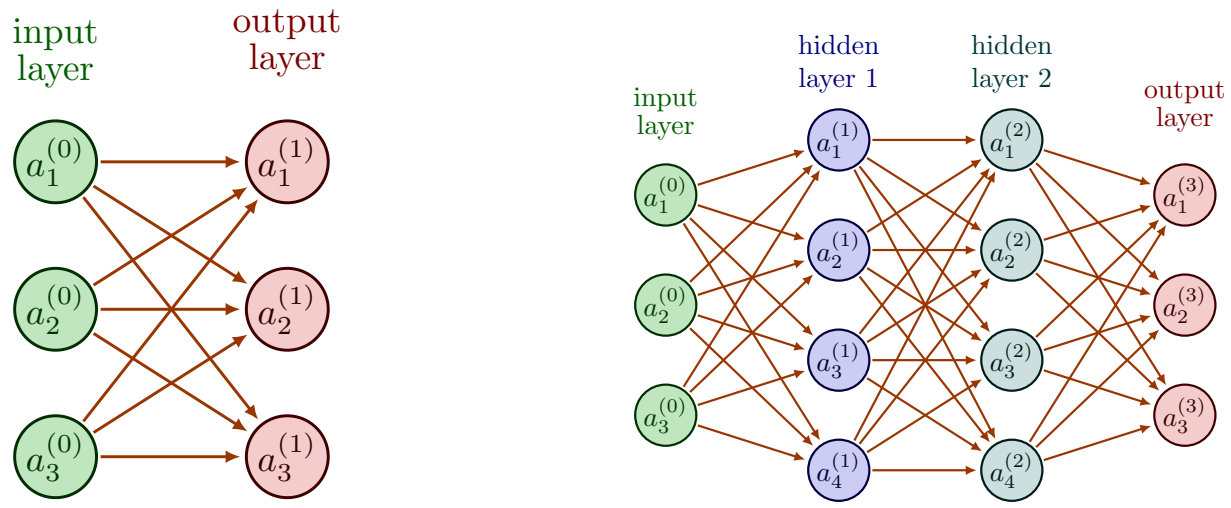
With this in mind, the choice of neural network architecture is highly context dependent. To approximate a complex function, a network can utilize situational properties that will reduce the number of required parameters. In opposition, a small network offers a smaller set of possible transformations, reducing its potential to consider subtle features.

3.2.1 Feedforward Neural Networks (FNN)

Feedforward neural networks (FNN) are a broad class of networks that process information in a single direction, in opposition to recurrent neural networks (RNN) that can compute operations in cycles, using multiple times the same set of transformations. One main advantage of FNNs is the simplicity of the information flow, making them easier to interpret. FNN can come in a wide variety of architectures, making them extremely versatile.

3.2.2 Perceptrons (MLP)

Perceptrons are one of the most common types of networks, first described in the mid 19th [83]. Furthermore, Perceptrons are one of the simplest form of neural network, this is why one was used in Sec. 3.1 to introduce the concept of NNs. Like their name suggests, single-layer Perceptrons are composed of a single layer, transforming directly the input into an output. The architecture is limited to solving linearly separable problems due to the reduced number of layers; a graph structure is presented in Fig. 3.3a.



(a) Single-Layer Perceptron

(b) Multi-Layer Perceptron

Figure 3.3 Comparison of single and multi-layer perceptrons.

Multi-Layer Perceptrons (MLPs) on the other hand are a class of feedforward artificial neural networks composed of multiple layers of nodes, where each node (neuron) is connected to every node in the subsequent layer. These networks have at least one hidden layer between the input and output layers, enabling them to model complex nonlinear relationships. An example of MLP is presented with two hidden layers in Fig. 3.3b.

3.2.3 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are deep learning models specifically designed for processing data with a grid-like structure, such as images and time series. CNNs exploit the spatial and temporal dependencies in data by applying convolutional operations, making them highly effective for tasks in image recognition, computer vision, and other domains with structured data.

The first main characteristic of this family of networks is the use of convolutional layers. Naturally, this type of layer applies a convolution operation between the input and a kernel $\boldsymbol{K}$ composed of tunable parameters. A visual representation of the convolution operation on a two-dimensional input matrix is shown in Fig. 3.4. The elements of the convolution describe the sum of the element-wise multiplication between the kernel and a sub-matrix of the input. The operation is flexible in the way it is implemented, for example the size of the kernel is a user defined parameter. Additionally, padding around the input can be added,

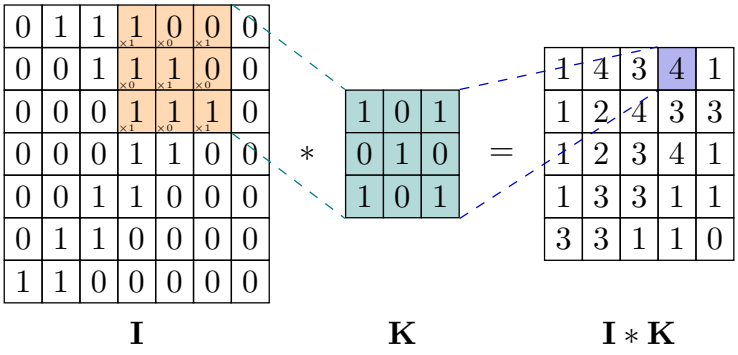


Figure 3.4 Convolution operation of a 3×3 kernel applied to a 7×7 matrix.

which describes increasing the size of the input to include user defined values around the input (often zeros). The step size of the convolution is also sometimes changed to give a coarser description of the operation. Finally, multiple kernels or filters can be applied on the same input, this is often described using the term channels [84, 85]. In this case, parallel operations can be applied to the input.

The convolution operation is interesting since it reduces the number of connections between two layers by only considering local features, in contrast to fully connected layers [84]. This property is incredibly powerful in the context of image recognition, since features are often local. Furthermore, CNNs have shift invariance properties which describe the capacity to capture features independently of the position of the feature inside the input [86]. Intuition for this property is that a CNN can detect an object inside an image independently of the position of the object inside the image. It is important to mention that CNNs are not perfectly shift invariant, but that existing literature on the subject has mainly solved this issue [86].

The second important building block of CNNs is the pooling layer. Often a maximum pooling layer is used in combination with convolution layers. From a computation point of view, it describes the maximum value inside a sub-matrix of some input matrix I. This computation is shown in Fig. 3.5 for a 6×6 input matrix and a kernel of size 2×2. The main property of pooling layers is to provide a down sampling process for the convolution layer. This down sampling describes a reduction of the input size that the convolution cannot achieve directly. To increase the dimensionality of the data, it is also possible to achieve up sampling [87]. It is through a combination of convolutional layers, each followed by a pooling step, that CNN computes process data.

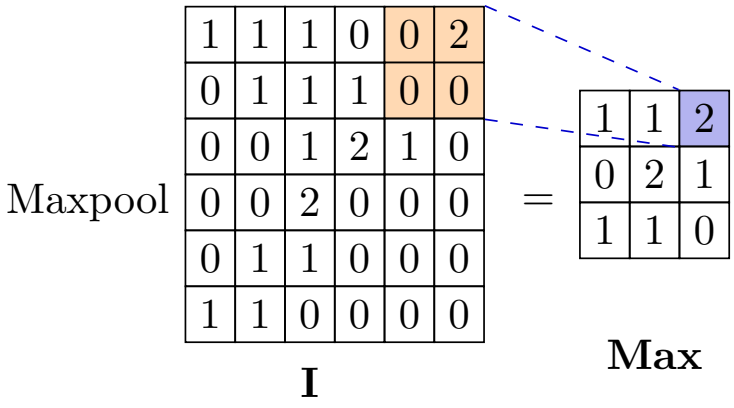


Figure 3.5 Max pooling operation on a random input matrix.

3.2.4 Autoencoders

Autoencoders describe a family of networks that contain an encoder and decoder layer. The encoder typically transforms high dimensional inputs into a lower-dimensional representation, this new representation is then transformed back to the original dimension. In Fig. 3.6 an example of an autoencoder is described where a size 6 input is encoded into a size 3 intermediary output, this output is then processed by a decoder layer that transforms back the information into a size 6 output. A powerful property of autoencoders is that the network

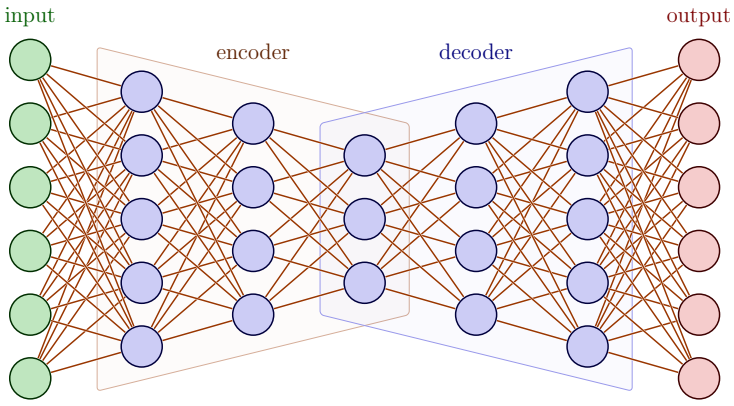


Figure 3.6 Example of an autoencoder neural network, where the first section describes the encoder layer and the second the decoder.

does not require the problem to be supervised. In fact, so far, the discussion around neural networks considers that the user has access to the ground truth and can feed this knowledge through the optimization of the loss function. However, this is not always possible, an example is the previously discussed photon number classification problem.

Autoencoders can be trained to reproduce their input at the output. While this objective might initially seem unproductive, it serves a critical purpose: the network learns to transform the data into a compact, low-dimensional representation. The accuracy of this compression is ensured because the decoder must reconstruct the original data from the compressed representation, effectively capturing the essential features of the input.

The size of the bottleneck (size of encoder output) sets the complexity of the features the network can describe and limits the accuracy of the reconstruction (output of the decoder compared to the autoencoder input). To enhance the capabilities of the network to create a low-dimensional that describes a maximum number of features, the loss function is often defined in terms of a combination of multiple criteria. For example, variational autoencoders (VAE) typically optimize a combination of MSE between the input and output of the autoencoder and a KL divergence between the output of the encoder and a normal distribution [88].

CHAPTER 4 ARTICLE 1: ACCURATE UNSUPERVISED PHOTON COUNTING FROM TRANSITION EDGE SENSOR SIGNALS

4.1 Submission Information

Authors : Nicolas Dalbec-Constant, Guillaume Thekkadath, Duncan England, Benjamin Sussman, Thomas Gerrits, and Nicolás Quesada

Journal Name : PRX Quantum

Date of Submission : November 19 2024

4.2 Abstract

We compare methods for signal classification applied to voltage traces from transition edge sensors (TES) which are photon-number resolving detectors fundamental for accessing quantum advantages in information processing, communication and metrology. We quantify the impact of numerical analysis on the distinction of such signals. Furthermore, we explore dimensionality reduction techniques to create interpretable and precise photon number embeddings. We demonstrate that the preservation of local data structures of some nonlinear methods is an accurate way to achieve unsupervised classification of TES traces. We do so by considering a confidence metric that quantifies the overlap of the photon number clusters inside a latent space. Furthermore, we demonstrate that for our dataset previous methods such as the signal's area and principal component analysis can resolve up to 16 photons with confidence above 90% while nonlinear techniques can resolve up to 21 with the same confidence threshold. Also, we showcase implementations of neural networks to leverage information within local structures, aiming to increase confidence in assigning photon numbers. Finally, we demonstrate the advantage of some nonlinear methods to detect and remove outlier signals.

4.3 Introduction

Photonics is a strong contender for building large-scale quantum information processing systems [1–5]; in many of these systems, photon number detection plays an essential role, serving as a resource for quantum advantage. These detectors can be used, for example, for the heralded generation of non-Gaussian states [6–13, 89], for the sampling of classically-intractable

probability distributions [14–19] or for directly resolving multiple quanta improving the Fisher information of interferometric protocols [20,21,90]. The use of photon number resolving detectors provides a significant advantage as a single detector can determine the number of photons associated with a quantum state accurately [22, 23], without requiring a multiplexed network of threshold detectors with its concomitant complexity and potential inefficiency [16, 24, 25]. Transition edge sensors (TES) have been used for this task, offering resolution over a wide energy range. Resolutions up to 30 photons have been demonstrated [26], although this quantity is typically lower, on the order of 17, if more straightforward techniques are used [23].

TESs exploit the superconducting phase transition of photosensitive materials to achieve an extremely sensitive calorimeter [27]. During operation, the material is cooled below its critical temperature and then current-biased to the transition region between its superconducting and normal state. In this region, the temperature increase following the absorption of a single photon leads to a measurable change in the material's resistance [28, 29]. The resistance change is read-out using a low noise amplifier such as superconducting quantum interference devices (SQUIDs), which also enable the creation of large arrays of TES detectors via read-out multiplexing [27]. Optimized materials and coupling techniques have demonstrated efficiencies of up to 98% [39].

The readout of these devices is non-trivial as the quantity one wants to determine, the energy (or the photon number for a fixed frequency), is reflected in a nonlinear fashion in the voltage signal produced by the detectors' electronics [30]. Historically, the integral (area) of the signals has been used to assign photon numbers [23, 31]. However, distinguishing large photon numbers becomes challenging with this technique. To address this issue, linear techniques such as Principal Component Analysis (PCA) have been used [32]. A machine learning method, adapted from the K-means algorithm to account for the Poissonian statistics of laser sources, has also been developed [33]. However, these methods' simplicity or assumptions can limit their performance or usability for model-free photon number detection and when measuring non-classical sources, which typically do not have Poisson photon-number statistics.

With the increased popularity of machine learning in the field of signal processing [34] and quantum systems [35], one might naturally ask whether employing more sophisticated methods could lead to enhanced resolution of photon numbers. In this work, we answer this question by assessing the performance of multiple techniques for photon number classification using TES signals. We do so by considering a confidence metric that quantifies the overlap of the photon number clusters inside a latent space. We demonstrate that for our dataset previous methods such as the signal's area and PCA can resolve up to 16 photons

with confidence above 90% while nonlinear techniques can resolve up to 21 with the same confidence threshold. Furthermore, we also showcase implementations of neural networks to leverage information within local structures, aiming to increase confidence in assigning photon numbers. Finally, we demonstrate the advantage of some nonlinear methods to detect and remove outlier signals.

Our manuscript is structured as follows: in the next section, Sec. 4.4, we formulate the problem of photon-number discrimination in the general setting of unsupervised classification and dimensionality reduction. Next, in Sec. 4.5, we offer a brief overview of the methods used to compute similarities between signals and how we distinguish signals that belong to the different photon number classes. We present our results in Sec. 4.6 using experimental data, followed by a discussion of the use cases of the described methods in Sec. 4.7

4.4 Methodology

4.4.1 Problem Formulation

Consider a data matrix $\boldsymbol{X} \in \mathbb{R}^{u\times t}$ that stores u signals x_i of size t. We assume there exists an operation $f(\boldsymbol{X})$ that can transform $\boldsymbol{X}$ into a vector $\boldsymbol{n} \in \mathbb{R}^{u\times 1}$ that contains the photon number associated with every signal. The goal of the classification becomes finding a parametric transformation $F(\theta', \boldsymbol{X})$ with user-defined parameters θ' that approximates as closely as possible the true transformation $f(\boldsymbol{X})$.

The problem is defined as an unsupervised classification, meaning the true elements of $\boldsymbol{n}$ are unknown. Additionally, given an experiment, the method needs to accept arbitrarily high photon numbers within the visibility limit of the detector.

4.4.2 Dimensionality Reduction

To solve this unsupervised classification problem, dimensionality reduction techniques are used. This process describes the transformation of $\boldsymbol{X}$ into a lower-dimensional output $\boldsymbol{Y} \in \mathbb{R}^{u\times r}$ that retains a meaningful amount of the input information. The new space of dimension $r < t$ is referred to as a latent space and is limited to one and two dimensions in this study. The proposed approach could be used for an arbitrarily large latent space, although these higher dimensional spaces are harder to interpret.

We use dimensionality reduction since it is a natural extension of previous work that uses PCA [32]. Moreover, this framework is used to make the current work compatible with existing tomography routines [32]. It also enables the visualization and interpretation of an

entire dataset, a task difficult by directly observing the TES signals. Supposing an accurate transformation exists and is faster to process than the acquisition rate of the detector, the low-dimensional representation reduces the memory requirements of experiments by acting as a compression step.

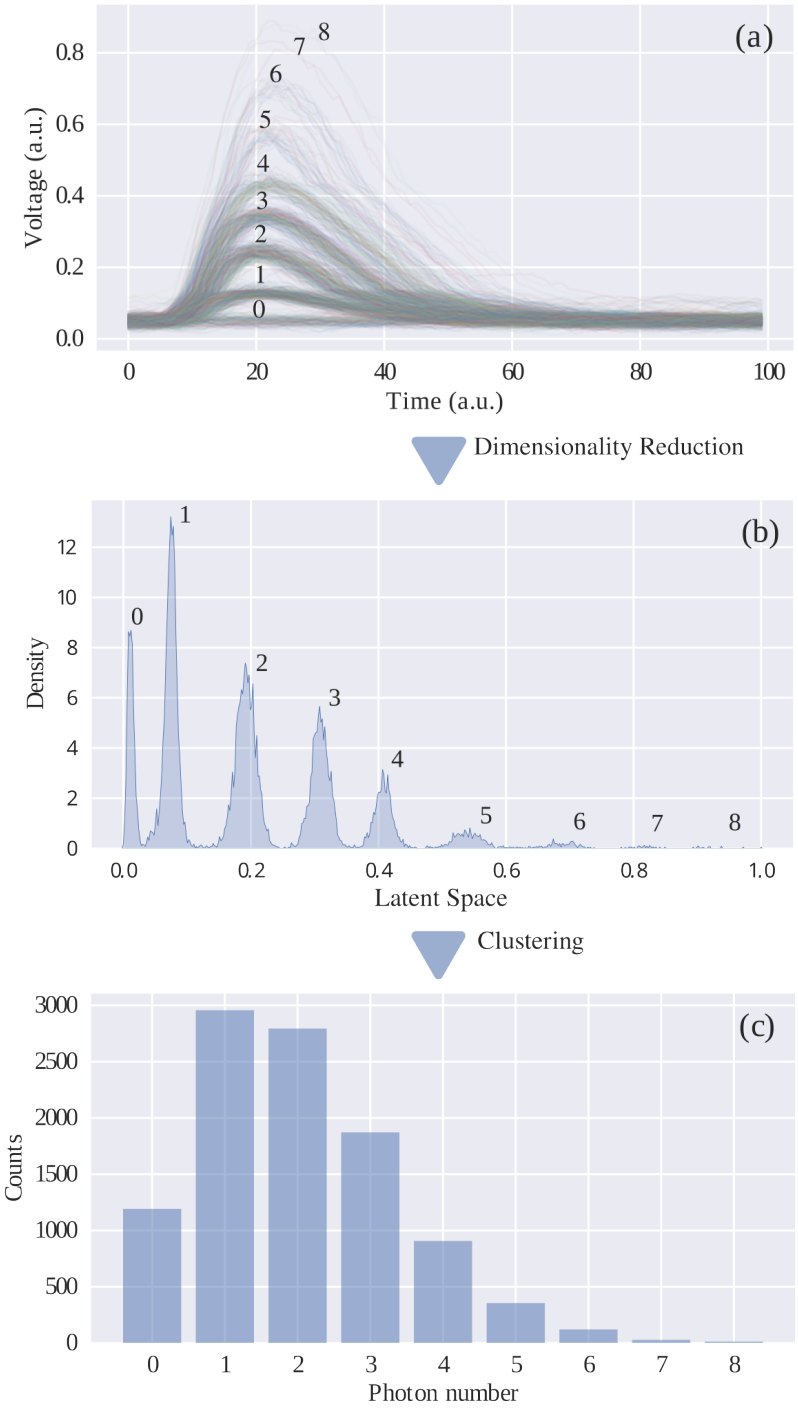


Figure 4.1 **(a)** Example of a dataset $\boldsymbol{X}$ with $u = 1\,024$ raw TES traces with $t = 100$. **(b)** The dataset $\boldsymbol{X}$ is transformed into $\boldsymbol{Y}$ which has a single dimension ($r = 1$), here plotted using a kernel density estimation [38]. The dimensionality reduction technique (maximum value of the signals in this case) creates a low-dimensional space where signal features become apparent. Each peak is a cluster that represents the underlying dominant feature of the signals: the photon numbers. **(c)** In this case, clusters in the latent space are assigned a photon number $n \in \{0, 1, ..., 8\}$. To assign samples, the space is divided in regions most likely to be associated with a specific photon number (see Sec. 4.5.6). From labelled samples, a photon number distribution can be generated.

Considering every signal in $\boldsymbol{X}$ can be associated with a photon number $n \in \{0, 1, ..., c\}$, where c is the photon-number cutoff, i.e., the largest distinguishable photon number. We assume that effective dimensionality reduction organizes similar samples near each other, forming regions of high density.

We illustrate the process in Fig. 4.1 by transforming the TES signals (Fig. 4.1a) into one-dimensional samples presented in Fig. 4.1b. This low dimensional space is visualized using a kernel density estimation of the latent space (Gaussian kernel) [38]. From the position of the samples in the latent space (never considering the density estimation in the computation) it is possible to find regions most likely to describe a photon number $n \in \{0, 1, ..., 8\}$, we discuss this step in Sec. 4.5.4. Finally, from this interpretation of the low-dimensional space, a photon number can be assigned to every sample (Fig. 4.1c). The regions of high density in Fig. 4.1b are called clusters and are associated with photon numbers. We note that clusters can be defined using other heuristics like neighbour distances.

An additional justification for the use of dimensionality reduction in combination with clustering instead of directly clustering over high dimensional data is that existing work has empirically demonstrated that creating a low dimensionality embedding increases the clustering capabilities in unsupervised settings [91].

4.5 Methods

We test a wide range of methods to showcase different approaches to the dimensionality reduction task. Due to the range of published solutions to the dimensionality reduction task, we limit our tests to the methods described in this section.

With experimental motivations, we consider the properties and use cases of dimensionality reduction techniques. To do so, the methods are divided into three categories based on their characteristics: basic feature, non-predictive, and predictive.

4.5.1 Basic features

The methods in this category rely on some feature with physical significance, and their latent space represents the value of this feature. These methods are fast to compute due to their simplicity and can be combined with noise filtering to increase resolution [23].

Maximum Value

The maximum value of the signals has been used in some cases for photon number resolution [23]. For experiments that only require the measurement of low photon numbers, sufficient information is found in the maximum value. For high enough photon numbers, the traces reach a plateau and the maximum value no longer gives information [30].

Area

TES pulse area relates non-linearly to the energy absorbed by the sensor and therefore can be used for dimensionality reduction [23]. The area is sensitive to noise outside the pulse, hence filtering and background rejection are used in some cases to increase the performance of this method. To offer a fair representation of this technique, a Butterworth filter is applied to the signals and a threshold is introduced to reduce the influence of noise. Following existing work, the threshold is defined above the noise distribution in the flat region of the TES signals (where only vacuum is detected) [23].

4.5.2 Non-predictive methods

The methods in this category organize data within a latent space by considering the entire dataset. However, once computed, these methods do not provide a transformation that can be directly applied to new data. To predict the position of a new sample in the latent space, the entire dataset must be recomputed. As a result, these methods are less scalable and are better suited for post-processing data.

t-Distributed Stochastic Neighbour Embedding (t-SNE)

The method t-SNE is non-predictive and attempts to create a low-dimensional representation of the data by organizing all the samples in a low-dimensional space. The position of the samples is assigned using a gradient descent by minimizing the Kullback-Leibler divergence (KL)

$$\mathrm{KL}(P\|Q) = \sum_{i=1}^{u} \sum_{\substack{j=1 \\ j\neq i}}^{u} p_{ij} \log \frac{p_{ij}}{q_{ij}}. \tag{4.1}$$

In the KL divergence, p_{ij} represents joint probabilities that describe the similarities between high-dimensional samples x_i and x_j and is the q_{ij} joint probabilities for low-dimensional samples y_i and y_j [92]. The high-dimensional joint probabilities are set to be symmetric

conditional probabilities defined as

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2u}, \tag{4.2}$$

with conditional probabilities defined using Gaussian functions

$$p_{j|i} = \frac{\exp\left[-\frac{1}{2}||x_i - x_j||^2/\sigma_i^2\right]}{\sum_{\substack{k=1 \\ k\neq i}}^{u} \exp\left[-\frac{1}{2}||x_i - x_k||^2/\sigma_i^2\right]}, \tag{4.3}$$

where $||x|| = (\sum_i x_i^2)^{1/2}$ represents the Euclidean norm. In low-dimensional space, the joint probabilities are given by Student t-distribution

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k=1}^{u} \sum_{\substack{l=1 \\ l\neq k}}^{u} (1 + ||y_k - y_l||^2)^{-1}}. \tag{4.4}$$

To offer high resolution over local structures in the data the variance σ_i^2 of each high dimensional Gaussian is tuned using an information parameter called the Perplexity. Perplexity is defined as

$$\text{Perp}(P_i) = 2^{H(P_i)}, \tag{4.5}$$

where $H(P_i)$ is the Shannon entropy

$$H(P_i) = -\sum_{j=1}^{u} p_{j|i} \log_2 p_{j|i}. \tag{4.6}$$

This parameter, initially introduced in speech recognition, is user-defined and is often described as an effective number of neighbours [93]. The intuition behind this value is that the variance of each Gaussian in the high dimensional space is tuned to have a tail with a limited number of relevant neighbours. This means neighbours outside the effective range of the Gaussian will have similarity values considerably smaller.

Uniform Manifold Approximation and Projection (UMAP)

We describe UMAP by emphasizing its similarities with t-SNE. UMAP makes use of stochastic approximate nearest neighbour search and stochastic gradient descent to optimize a cross-

entropy cost function [94] defined as

$$C = \sum_{i=1}^{u} \sum_{\substack{j=1 \\ j \neq i}}^{u} v_{ij} \log \left(\frac{v_{ij}}{w_{ij}} \right) + (1 - v_{ij}) \log \left(\frac{1 - v_{ij}}{1 - w_{ij}} \right), \tag{4.7}$$

where v_{ij} and w_{ij} are similarities respectively in high and low-dimensional space. UMAP's high-dimensional conditional probabilities $v_{i|j}$ are defined as local fuzzy simplicial set memberships

$$v_{i|j} = \exp\left[\left(-d(x_i, x_j) - \rho_i \right) / \sigma_i \right]. \tag{4.8}$$

In $v_{i|j}$, a user-selected smooth nearest neighbours distance $d(x_i, x_j)$ is defined (only Euclidean distance is used in this work), ρ_i is the nearest neighbour distance [95] and σ_i is an approximation for the k-nearest neighbour distance.

Like t-SNE the high dimensional similarities v_{ij} are defined to be symmetric and follow

$$v_{ij} = (v_{j|i} - v_{i|j}) - v_{j|i} v_{i|j}. \tag{4.9}$$

As for the low-dimensional similarities w_{ij} they follow

$$w_{ij} = \left(1 + a || y_i - y_j ||^{2b} \right)^{-1}, \tag{4.10}$$

where a and b are user-defined parameters found through a fitting algorithm. If a and b are 1, we have the t-student function of t-SNE.

Isometric Mapping (Isomap)

Isometric mapping finds the nearest neighbours of every sample and creates a graph representation where every point is connected to its neighbour [96]. The algorithm attempts to compute the shortest distance between every connected point. Finally, a multidimensional scaling step computes a low-dimensional graph representation.

4.5.3 Predictive methods

Predictive methods need to be trained using data, once trained these methods offer a transformation that can be used to label new signals. This generally translates into fast computation but requires an initialization step to train the model.

Principal Component Analysis (PCA)

Principal component analysis is a linear method previously used for TES and superconducting nanowire single-photon detector (SNSPD) signal classification [32,97]. For a data matrix $\boldsymbol{X}$, PCA transforms $\boldsymbol{X}$ to a new coordinate system to minimize the total distance between the samples and the principal components (columns of $\boldsymbol{W}$). By minimizing this distance, the variance of the projected points is maximized [98]. For a data matrix $\boldsymbol{X}$ and a principal component matrix $\boldsymbol{W} \in \mathbb{R}^{u\times r}$, the matrix multiplication

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}, \tag{4.11}$$

transforms every signal into a low-dimensional representation $\boldsymbol{Y} \in \mathbb{R}^{u\times r}$ of size r equal to the number of principal components considered. It can be shown that optimal vectors of $\boldsymbol{W}$ are given by the singular value decomposition (SVD) of the covariance matrix $\boldsymbol{X}^\top\boldsymbol{X}$. This is further simplified to SVD elements of $\boldsymbol{X}$ where $\boldsymbol{W}$ is taken directly from $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{W}^\top$. In this decomposition, $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal and $\boldsymbol{\Sigma}$ is a rectangular diagonal matrix. Once $\boldsymbol{W}$ is defined, prediction is done by replacing $\boldsymbol{X}$ by new data $\boldsymbol{X}_{\text{pred}}$ in equation (4.11).

Kernel Principal Component Analysis (Kernel-PCA)

Kernel principal component analysis uses a mapping to project data onto a feature space of size Q (typically $Q \gg t$) where the data has the potential of being linearly separable [99]. It can be shown that the projection of the data points inside the feature map $\phi(x)$ onto the principal components in the feature space can be computed without explicitly computing the mapping $\phi(x)$. This is done through the introduction of a kernel function that follows some restrictions in its construction [100].

We benchmark a Polynomial (Poly), Radial Basis Function (RBF), Sigmoid and Cosine kernel defined as:

$$\text{Poly}: \quad k(x_n, x_m) = (\gamma x_n^\top x_m + c)^d, \tag{4.12}$$

$$\text{RBF}: \quad k(x_n, x_m) = \exp\left(-\gamma||x_n - x_m||^2\right), \tag{4.13}$$

$$\text{Sigmoid}: \quad k(x_n, x_m) = \tanh(\gamma x_n^\top x_m + c), \tag{4.14}$$

$$\text{Cosine}: \quad k(x_n, x_m) = (x_n x_m^\top)(||x_n||\,||x_m||)^{-1}. \tag{4.15}$$

Non-Negative Matrix Factorization (NMF)

Non-negative matrix factorization is an iterative process that attempts to find a decomposition without negative elements to minimize some objective function. The method gives an approximate decomposition of the data matrix $\boldsymbol{X}$ described by

$$\boldsymbol{X} \approx \boldsymbol{Y}\boldsymbol{H}, \tag{4.16}$$

where $\boldsymbol{Y}$ represents the transformed data matrix and $\boldsymbol{H}$ the transformation matrix, which are both smaller matrices than X. The general process behind NMF offers a framework to compute adequate decompositions for specific applications. In other words, the loss function is chosen given an application. In this paper, we use a loss defined as

$$L(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{H}) = ||\boldsymbol{X} - \boldsymbol{Y}\boldsymbol{H}||^2_{\text{Frob}}. \tag{4.17}$$

The Frobenius norm is a matrix norm defined for a matrix $\boldsymbol{A}$ with elements a_{ij} as $||\boldsymbol{A}||_{\text{Frob}} = (\sum_{ij} |a_{ij}|^2)^{1/2}$. It can be shown that the optimization of the Frobenius norm is equivalent to the maximum likelihood estimate of $\boldsymbol{X}$ without Gaussian noise [101]. Additionally, we test NMF optimization using the KL divergence, where the loss function becomes

$$L(\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{H}) = \text{KL}(\boldsymbol{X} \| \boldsymbol{Y}\boldsymbol{H}). \tag{4.18}$$

Similarly to the Frobenius norm, the use of the KL divergence is equivalent to the maximum likelihood estimate of $\boldsymbol{X}$ without Poissonian noise [101].

To make a prediction using NMF, a new approximate decomposition is optimized based on a close-to-optimal initial guess defined in the training step.

Neural Networks

Neural networks have the potential to reproduce a wide variety of operations in a numerical structure that can be used efficiently to process large amounts of data. To quickly apply UMAP and t-SNE on new data, we use parametric implementations of these methods using neural networks. The main principle behind these parametric implementations is to constrain the embedding to transformations done through a neural network. In other words, a neural network is trained to optimize the KL divergence in the case of t-SNE and the cross-entropy in UMAP. By applying this constraint during training, we create a neural network that considers local structures and behaves similarly to t-SNE and UMAP. At this stage, new data can be embedded at an efficiency restricted by the complexity of the neural network

architecture.

More details about the neural network architecture and the training process are provided in Appendix A.

4.5.4 Clustering

Clustering refers to identifying groups of similar samples inside a latent space. For this task we use a Gaussian mixture model, given a user-defined number of clusters, this method finds the parameters of a mixture of Gaussians to describe the sample's distribution.

The choice is highly inspired by a similar model previously used in the tomography of TESs in combination with PCA [32]. Mixture models offer a statistical interpretation of latent spaces convenient for metrology and performance evaluation (confidence metric in Sec. 4.5.6).

The mixture model gives a continuous probability density function for the position s of samples given optimal parameters $\theta = \{(\omega_k, \mu_k, \Sigma_k) : k = 1, \cdots, K\}$. In the model, every cluster k is weighted by a value ω_k (where $\sum_{k=1}^{K} \omega_k = 1$), and modelled by a Gaussian with mean μ_k and covariance matrices Σ_k. The individual Gaussians $\mathcal{N}$ give the cluster probability density function and the probability of observing samples in position s given parameters θ are defined by

$$p(s|\theta) = \sum_{k=1}^{K} \omega_k \mathcal{N}(s|\mu_k, \Sigma_k). \tag{4.19}$$

The probability density function is found through an expectation maximization algorithm (EM algorithm) that attempts to find the maximum likelihood estimate of samples following a likelihood of

$$\mathcal{L}(\theta) = \prod_{i=1}^{p} \sum_{k=1}^{K} \omega_k \mathcal{N}(s_i|\mu_k, \Sigma_k). \tag{4.20}$$

Numerically it is more convenient to express this problem in terms of the log-likelihood given by

$$\ell(\theta) = \log(\mathcal{L}(\theta)) = \sum_{i=1}^{p} \log\left(\sum_{k=1}^{K} \omega_k \mathcal{N}(s_i|\mu_k, \Sigma_k)\right), \tag{4.21}$$

where the problem can be computed in terms of sum instead of products.

4.5.5 Number of clusters

The Gaussian mixture model offers different advantages for quality assessment but cannot directly determine the number of clusters in a latent space. The problem is solved using an elbow method considering the Akaike information criterion (AIC)

$$\mathrm{AIC} = 2K - 2\ln(\mathcal{L}(\theta)), \tag{4.22}$$

or the Bayesian information criterion (BIC)

$$\mathrm{BIC} = K\ln(u) - 2\ln(\mathcal{L}(\theta)). \tag{4.23}$$

The criteria assign a score given a number of clusters K, a likelihood function $\mathcal{L}(\theta)$, and a total number of data points u. By sweeping the number of clusters used in some models, these criteria give a way to find a balance between the number of clusters and the likelihood. In our case, the likelihood of the Gaussian mixture model is used to evaluate the information scores. The general idea of these criteria is to negatively score the number of clusters, considering it is always possible to overfit the data with more clusters. In other words, a model with more clusters can always achieve a higher or equal likelihood than a model with fewer clusters. The point of diminishing return is given by the "elbow" of the AIC and BIC when evaluating the criteria as a function of the number of clusters. After this point, the additional clusters mostly overfit the data.

The Silhouette score is also used with the information criteria to evaluate the number of clusters [102]. Since similar results are found with this method, the details are not described here.

4.5.6 Quality Assessment

Assessing the performance of dimensionality reduction techniques in an unsupervised setting is difficult since the ground truth is unknown. To tackle this task, we quantify cluster separation. To improve the performance evaluation it is also important to understand that the problem is not completely unsupervised considering photon sources used to generate samples follow known distributions. We include this knowledge of photon number distributions as an additional validation to cluster separation evaluation in the confidence metric (Sec. 4.5.6).

Confidence

We consider the probability density of photon events can be approximated from the sample's distribution in the latent space following the Gaussian mixture model. Following previous work [32], the confidence C_n is used as a performance metric for the resolution of photon numbers in a latent space following,

$$C_n = \int_{-\infty}^{\infty} \frac{p(s|n)^2 P(n)}{\sum_k p(s|k)P(k)} \mathrm{d}s. \tag{4.24}$$

In this equation, $p(s|n)$ is the probability density of observing a sample in position s in the latent space given it is labelled as n photons. Additionally, $P(n)$ is the probability of assigning a photon number n. In this model, we consider that the true clusters follow a Gaussian structure inside the latent space.

The confidence represents the probability of correctly labelling a sample in a given cluster in the mixture model. We note that equation 4.24 describes the confidence for a one-dimensional space but can be generalized to an arbitrarily high-dimensional latent space.

In practice $p(s|n)$ can be measured using a trusted source of n photons (i.e. using detector tomography), in which case C_n is equal to the probability of the detector measuring and assigning the correct number of photons [32].

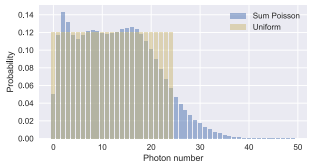
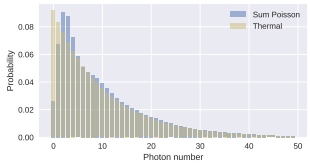
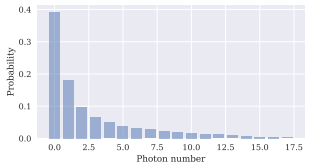
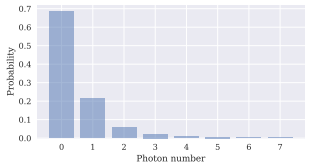
It is important to mention that the distances in the latent space do not necessarily have a physical meaning. The separation must only be interpreted as our capacity to distinguish clusters, and the confidence translates this concept into a probabilistic framework.

4.5.7 Datasets

Experimental data from previous work at the National Institute of Standards and Technology (NIST) is used to benchmark the different techniques in this work [30]. The original dataset was generated by progressively attenuating a coherent source from 29dB to 7dB, leading to 24 datasets each containing $u = 20\,480$ signals and $t = 8\,192$ time steps. This results in datasets that each have Poisson photon number distributions and mean photon number $\langle n_1 \rangle = 2.26$ to $\langle n_{24} \rangle = 7.08 \times 10^6$. These values were independently measured using a calibrated photodetector.

Instead of directly using these distributions, we construct two synthetic datasets (made of real traces) that follow a close-to-uniform and close-to-geometric distributions $P(n)$. These datasets are labelled as Synthetic Uniform and Synthetic Geometric in Table 4.1. Furthermore, for both of these datasets, a training and testing set were generated. Considering

Table 4.1 Number of samples u, number of time steps t and photon number distribution for both the training and testing portion of all the datasets used in this work. For cases where the photon number distribution is engineered to resemble a goal distribution, the **blue bars** represent the expected photon number distribution for a mixture of Poisson distribution and the **yellow bars** are the goal distributions used to fit the weights $w_{\langle n \rangle}$.

Name	Number (u)	Size (t)	Distribution	Reference
Synthetic Uniform	Train : 30 550 Test : 30 550	350		[30, 103]
Synthetic Geometric	Train : 57 020 Test : 57 020	350		[30, 103]
Synthetic Large	Train : 550 000 Test : 550 000	200		[104]
Noise	Test : 200 000	50		[104]

randomly selecting a portion of the samples in each experiment is equivalent to varying the weight $w_{\langle n \rangle}$ of a given Poisson distribution $P_{\langle n \rangle}(n)$ inside a mixture of Poisson distributions. The total expected distribution $P(n)$ can be described by

$$P(n) = \frac{1}{\xi} \sum_{\langle n \rangle \in \bar{N}} w_{\langle n \rangle} P_{\langle n \rangle}(n), \tag{4.25}$$

with

$$\xi = \sum_{\langle n \rangle \in \bar{N}} w_{\langle n \rangle}, \tag{4.26}$$

and where $\bar{N}$ is the set of available mean photon numbers $\langle n \rangle$. With this construction, the expected photon number distribution is a mixture of Poisson distributions shown in Table. 4.1. The choice of a uniform distribution is motivated by the desire to make the labelling task difficult by maximizing the distribution's entropy. In other words, for every sample

in a perfectly uniform distribution, the method would have equal chances of guessing every class. The choice of testing a geometric distribution comes from the desire to precisely measure thermal optical sources that follow a geometric photon number distribution. Also, distributions with a long tail can be difficult to process for certain methods since fewer examples are present in some classes (imbalanced dataset).

We add that these expected distributions are used as $P(n)$ in the computation of the confidence. The predictive methods are trained with the training set, and the analysis of performance metrics is done by feeding the test set to the trained methods. In the case of non-predictive and basic feature methods, the test set is directly used. The training and test datasets contain a total of $u = 30\,550$ traces of size $t = 350$ (first 350 values of the 8192 available time steps). We note that most of the weights $w_{\langle n \rangle}$ are set to zero because of the number of available Poisson distributions in the desired photon number range is small, making the synthetic distribution not perfectly uniform (see top row in Table 4.1).

To validate a hypothesis discussed in Sec. 4.7.2 we also use a larger dataset named Synthetic Large that was created using signals generated by TESs at the National Research Council Canada (NRC) in Ottawa. The data was generated by tuning the attenuation of a laser and measuring $u = 100\,000$ signals for each of these coherent sources.

Finally, we also make use of a dataset labelled Noise, in Sec. 4.7.5, for this dataset, $u = 200000$ TES signals were produced by detecting light generated by an integrated optical parametric oscillator (OPO) pumped below threshold using a pulsed-carved continuous wave laser, as in Ref. [105]. The OPO generated signal photons following a quasi-thermal distribution. In addition, noise photons from the pump leaked into the detected mode due to imperfect pulse carving and filtering. These noise photons were generated at random times relative to the signal photons. All datasets are summarized in Table 4.1.

4.6 Results

4.6.1 Validation

Before looking at performance metrics, a sanity check is done to validate the basic characteristics of the Synthetic Uniform dataset. This is done for the data from the different coherent sources (all with different mean photon numbers). Since coherent sources are used to generate the samples, a $g^{(2)}$ (second-order correlation) of 1 is expected. This quantity is

defined in terms of the first two moments of the photon number distribution, as

$$g^{(2)} = \frac{\langle n^2 \rangle - \langle n \rangle}{\langle n \rangle^2}. \tag{4.27}$$

We use the $g^{(2)}$ as a validation metric both ways by making sure the statistics of the light are correct and that the generated statistics using the numerical methods follow the physics of the system. In Fig. 4.2 we can see that every method has a $g^{(2)}$ close to 1 for most datasets. All methods consistently get farther from one as the mean photon number increases, the lack of resolution for high photon numbers explains this behaviour. Additionally, the number of signals associated with the high mean is limited compared to the low mean cases. The lack of resolution is especially present for the method based on the maximum value of the signals, since it cannot resolve photon numbers higher than 10 in our dataset.

4.6.2 Confidence

Considering the different dimensionality reduction techniques and following Gaussian mixture clustering, the confidence associated with every method is compiled in Fig. 4.3 for the Synthetic Uniform dataset. In this plot, the Kernel PCA techniques and NMF are not presented to facilitate readability, since they do not offer significant differences with PCA or are significantly worse. The number of clusters considered in the confidence plots is defined using the AIC and BIC information criteria and other considerations. First, the last cluster is always removed since it often offers an artificially high confidence considering there is no other cluster to overlap with farther in the latent space. Additionally, regions associated with multiple photons described by a uniform density are ignored. This is done since regions of uniform density can be described by an arbitrarily large number of Gaussians.

We found a significant increase in performance can be achieved using nonlinear methods. In Fig. 4.3 and Fig. 4.4 we show the confidence metric for the different methods considered for both the Synthetic Uniform and Synthetic Geometric datasets. We see that for both datasets previous methods like the signal's area and PCA can resolve up to 16 photons with confidence above 90% while t-SNE and UMAP can resolve up to 21 with the same confidence threshold. Parametric implementations of t-SNE and UMAP did not give satisfying results for these datasets however, in Sec. 4.7.2 we show that these implementations can outperform PCA if the dataset is sufficiently large.

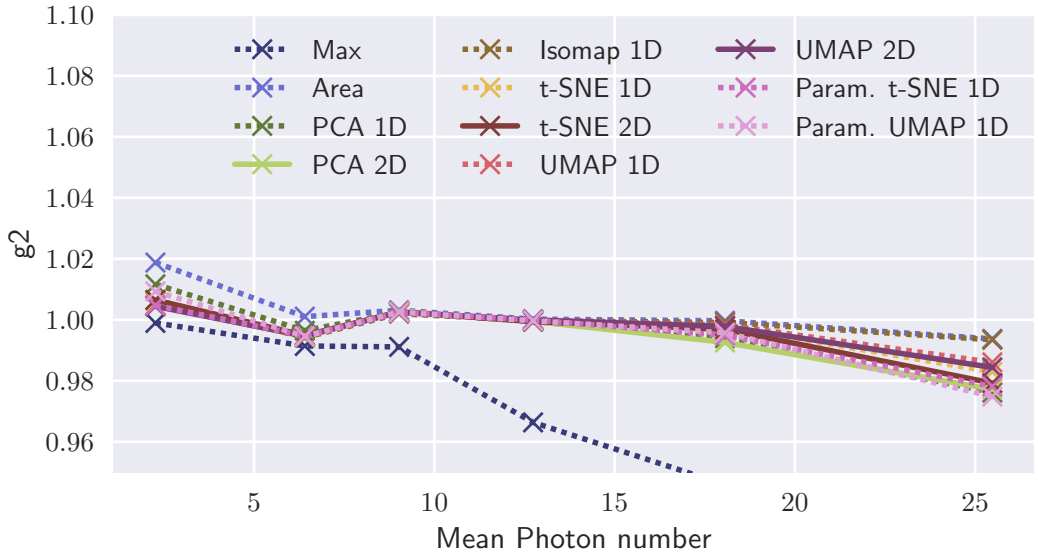


Figure 4.2 Computed second-order correlation for the different datasets (where markers are the mean photon number of the available coherent sources) and methods. In this figure, and the ones that follow, methods using a 1D latent space are represented by dotted lines, while those with 2D latent spaces are shown with solid lines.

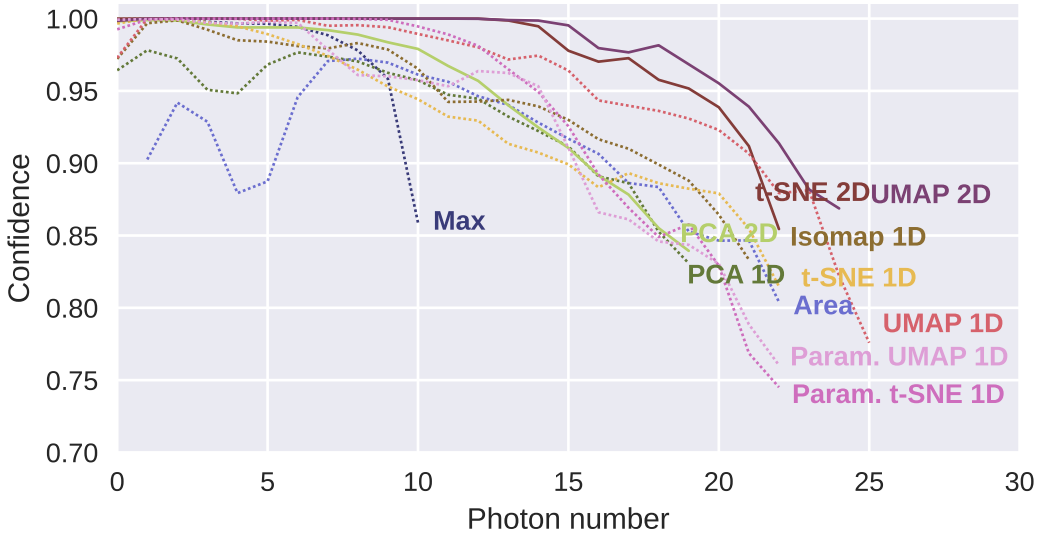


Figure 4.3 Confidence of photon number clusters for the different methods using the Synthetic Uniform dataset. In this figure, and the ones that follow, methods using a 1D latent space are represented by dotted lines, while those with 2D latent spaces are shown with solid lines.

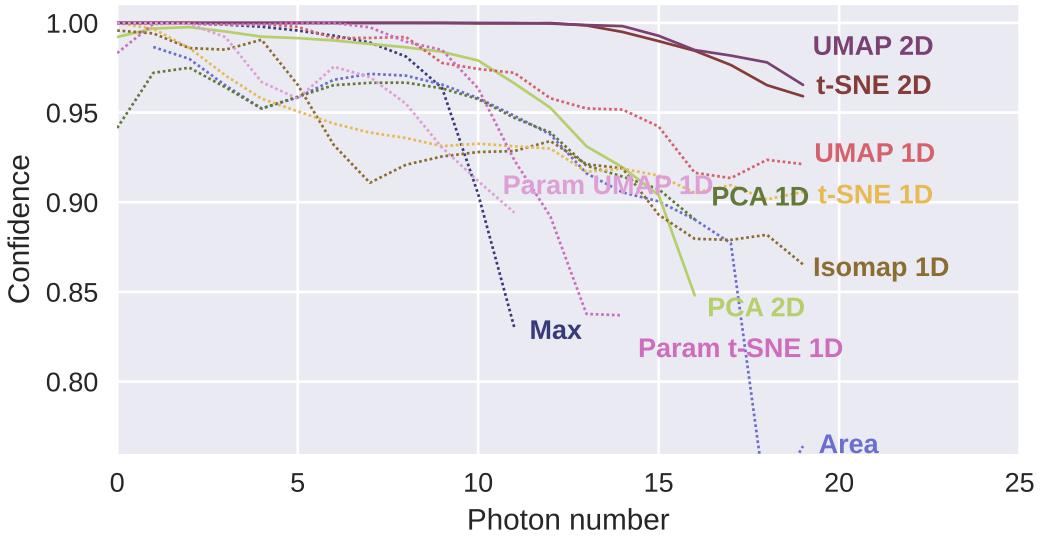


Figure 4.4 Confidence of photon number clusters for the different methods using the Synthetic Geometric dataset

4.7 Discussion

4.7.1 Qualitative Analysis

Through a visual analysis of the sample's distributions in latent spaces, it is possible to identify methods that show potential for unsupervised classification. In other words, methods that visually offer clear cluster separation have the potential to better perform at the classification task. To visualize the data in these different spaces, we use kernel density estimation, which involves summing a kernel function (Gaussian in this case) over all the samples to provide a smooth representation of the data distribution.

PCA is the first interesting method, since it was previously used for this task. We observe clear clusters, and the samples followed the expected arc-like structure presented in Fig. 4.5a and observed in previous work [32].

We also notice the promising separation of clusters using both t-SNE and UMAP. The sample distributions generated by these methods in two dimensions are presented in Fig. 4.5b and Fig. 4.5c.

The other methods tested in this work generate sample distributions with no special properties and, for this reason, are not further discussed. However, all methods and their results are available online [106].

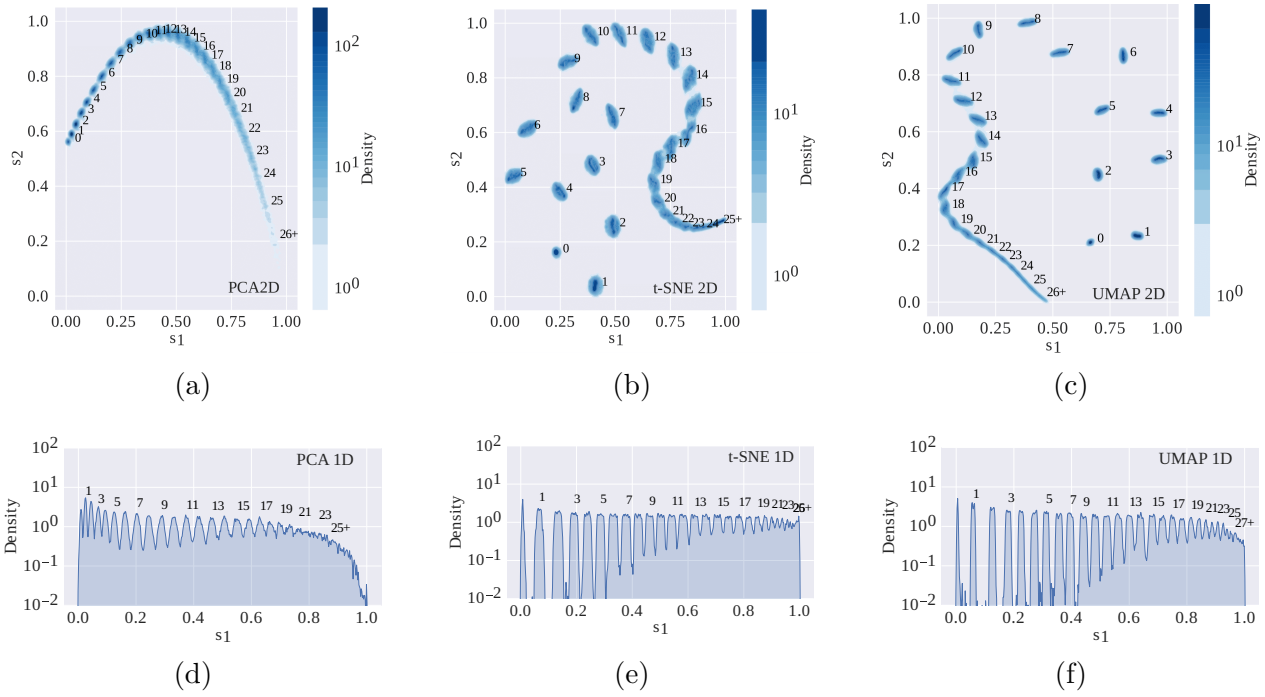


Figure 4.5 Kernel density estimation of the low dimensional embedding of TES signals generated by **(4.5a)** PCA 2D, **(4.5b)** t-SNE 2D, **(4.5c)** UMAP 2D, **(4.5d)** PCA 1D, **(4.5e)** t-SNE 1D, **(4.5f)** UMAP 1D.

4.7.2 Limits for Parametric Implementations

We consider t-SNE and UMAP to offer some approximate upper bound on the confidence of their parametric implementation. This is justified by the fact that both methods follow the same optimization scheme. However, non-parametric methods are not limited by the set of possible transformations in the neural network architecture. We therefore hypothesize that given a large enough neural network and adequate hyperparameters, the performance of Parametric t-SNE and UMAP has the potential to resemble their non-parametric equivalent.

The training process to generate a network with the reported performance for the Synthetic Uniform and Geometric datasets required a fair amount of tuning to give satisfying results, which is not ideal for experimental setups. We mainly attribute this problem to the limited amount of training data, which makes it easy to overfit the model to the training data. More precisely, by learning local data structures the neural network learns less generalized features which limits its capacity to make predictions. This family of neural networks is therefore more reliant on having access to a large training dataset, since it needs examples for a wider range of fine signal features. This limits the performance capabilities demonstrated

in this work, however, with a larger training set the neural networks can have prediction capabilities similar to the transformation of their non-parametric implementation. To verify this intuition, we used the Synthetic Large dataset previously mentioned in section 4.5.7. Using the $u = 300000$ signals, we trained a small feedforward neural network (5 linear layers of size 300). We present in Fig. 4.6 that with sufficient data, this network offers advantageous confidence values compared to previously used techniques in one-dimension.

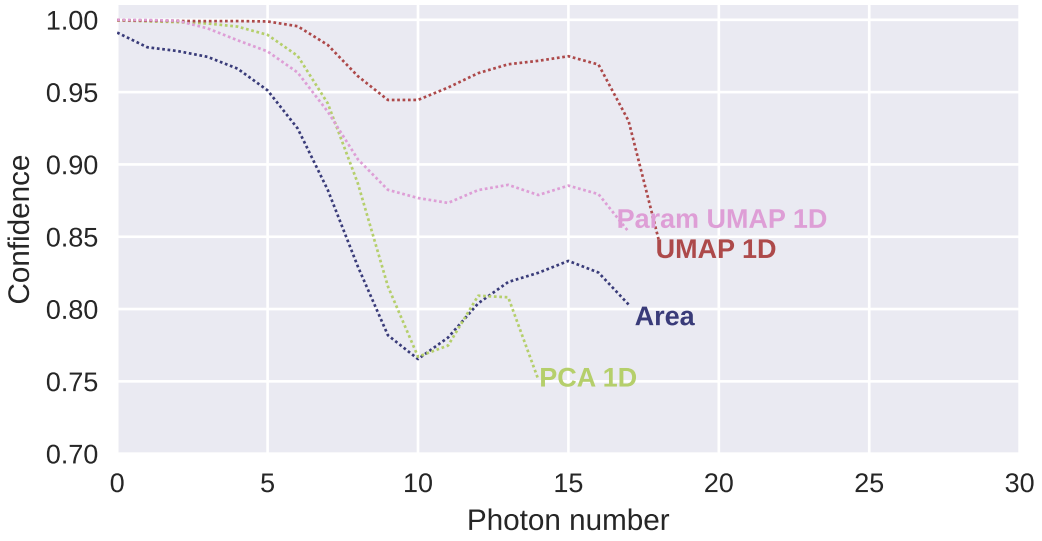


Figure 4.6 Confidence of Parametric UMAP compared with the non-parametric implementation and 1D PCA, for the Synthetic Large dataset taken at the National Research Council in Ottawa.

4.7.3 Impact of Embedding Dimension

The analysis of the dimensionality reduction techniques in this work assumes that the underlying true classes are associated with the photon numbers. This gives satisfying results because the traces for each photon number follow a clear pattern that different methods can easily capture. However, additional considerations are needed to solve the photon number classification problem. First, cluster distinguishability inside the low-dimensional representations is possible because the underlying structures of photon numbers are dominant in comparison to other characteristics like noise. Additionally, the dimensionality reduction techniques are only aware of data structure at different scales and never explicitly have a grasp of the physical system. We emphasize this property since it makes the method almost completely independent of the statistics of the measured light and does not require prior knowledge of the light source. To come back to the data structures, when methods encode

data in a low dimensional space they need to find a representation that describes the entire complexity of the signals. This means that noise and photon number structures are equally preserved in the embedding. If enough noise structures exist, the method will not have enough space in a single dimension to represent this variety, and the resulting embedding can show excessive broadening of clusters. The constraint of preserving structures in the data limits the potential of finding well separated clusters in lower dimensional embeddings. This is a reason why it is easier to find an embedding with well-separated clusters in two-dimensional spaces, even if the underlying classes we wish to identify are contained in a single dimension: the photon number.

4.7.4 Global vs Local data structures

In unsupervised classification tasks, it is often suggested to use dimensionality reduction techniques that preserve global structures rather than local structures [107]. This is because preserving the local structure may alter the distances and density of the data from the original space to the generated embedding. This characteristic makes it harder to guarantee that generated clusters are real or associated with the desired classes. Depending on the data, noise structures can also be grouped, creating artificial clusters. While this can be true, in the case of TES traces we argue that data does not contain electrical noise important enough to create artificial clusters. Additionally, noise from temporally uncorrelated photons is described by well-defined signal signatures. Looking at local structures gives the capacity to cluster these structures, arguably making it a positive rather than a negative feature, as we explain in the next section.

4.7.5 Outlier Detection

A one-dimensional embedding is efficient from a computational point of view, since the clustering problem can be translated into a sorted array search. However, depending on the use case, we argue that two dimensions may offer deeper insight due to their capacity to capture a wider range of structures. For example, if temporally uncorrelated light overlaps with the light modes one seeks to analyse, then a single dimension is likely not enough space to correctly capture the photon-number statistics of the modes under analysis. Adding to what is mentioned in the previous section, the noise becomes an additional structure to represent, and effectively the proportion of information that the method can allocate to the photon number structure is reduced. This is shown in Fig. 4.7a where we use the Noise dataset (section 4.5.7) and observe cluster broadening due to the presence of temporally uncorrelated photons. In this case, the two-dimensional representation becomes more useful,

cf. Fig. 4.7b, to describe the complexity of the dataset. Using a second dimension, the

(a) Density estimation of PCA embedding using the first principal component.

(b) Scatter plot of embedding of TES traces using PCA in two dimensions.

Figure 4.7 Low dimensional representation using PCA of the Noise dataset containing signals from a system with temporally uncorrelated photons.

uncorrelated light becomes distinguishable, as shown in Fig. 4.8. In this space, it is not only easier to interpret the proportion of uncorrelated light, but it is also possible to remove these outliers by carefully selecting the latent space regions associated with correlated light.

We also noticed that methods that preserve local structures tend to create clearer clusters for noise structures, facilitating the clustering task. This effect is seen in Fig. 4.8 where the uncorrelated noise is found on curve structures and photon numbers in tear-like shapes. If we look closely at the content of these clusters, we see that it is possible to identify signals of uncorrelated single photons before the trigger time (cluster 4.8c) and after the trigger time (cluster 4.8b). Similarly, we find uncorrelated single photons combined with correlated single photons in clusters 4.8e and 4.8f. In clusters 4.8a, 4.8d, 4.8g, and 4.8h we find the standard photon numbers 0 to 3 without uncorrelated light. Similar analysis could be done using more traditional methods like PCA, however the clustering becomes significantly harder. This lack of cluster structure is visually demonstrated in Fig. 4.7 where the uncorrelated light becomes a broadening of the temporally correlated photon numbers.

We note that the Gaussian Mixture Model is not as effective in clustering noise features, especially considering a photon number embedding from UMAP. We found that methods like HDBSCAN, which is a hierarchical density-based clustering technique, are well-suited for UMAP embedding [108]. This technique has the main advantage of working on clusters that do not follow a Gaussian structure, which is adequate for noise clusters that can have a variety of shapes.

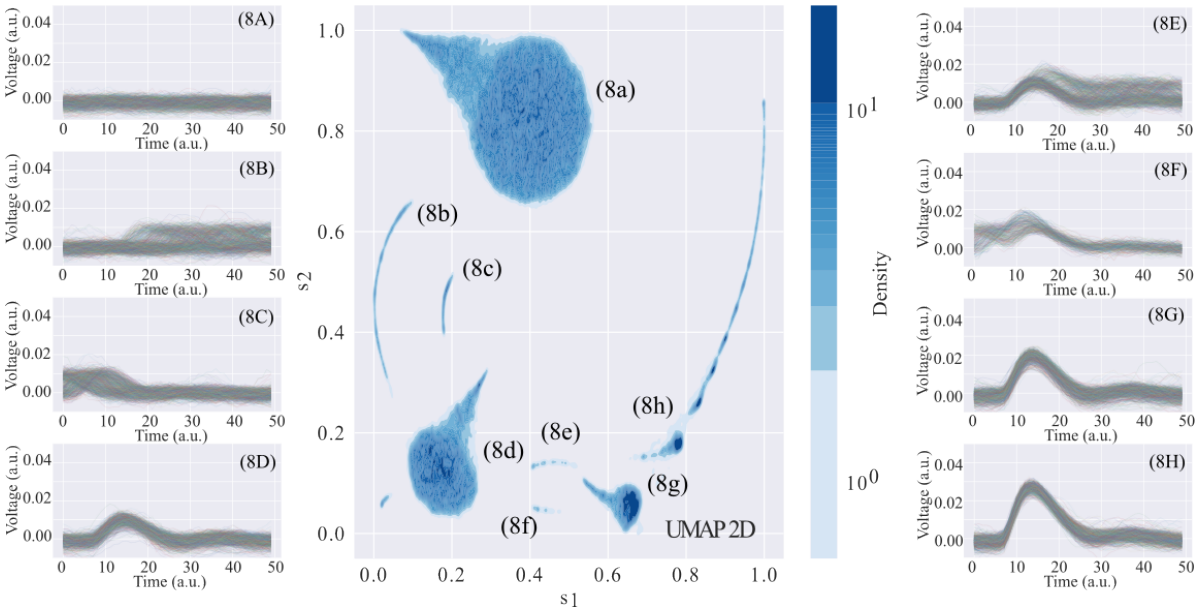


Figure 4.8 In the centre, we present a low dimensional representation using UMAP of a dataset containing signals from a system with temporally uncorrelated noise. Each cluster in the kernel density estimation is identified using lower case letters, and each graph, identified using the associated upper case letter, represents the signals in each labelled clusters. **(4.8a)**, **(4.8d)**, **(4.8g)**, and **(4.8h)** give the temporally correlated photon numbers 0 to 3. **(4.8b)** and **(4.8c)** are associated to uncorrelated signals, with zero photons correlated before and after the trigger time. **(4.8e)** and **(4.8f)** are single photons at the trigger time and uncorrelated signals before and after the trigger.

4.7.6 Impact of Gaussian Mixture Model

We note that one-dimensional results for t-SNE offer clusters that follow top-hat-like distributions, cf. Fig. 4.5. This feature decreases the confidence results, but not the actual potential clustering over this embedding. For a more accurate representation of t-SNE clusters, we use a generalized Gaussian distribution to represent the probability density of each cluster defined as

$$p(s|n) = \frac{\beta}{2\zeta_n \Gamma(1/\beta)} \exp\left[-\left|\frac{s-\mu_n}{\zeta_n}\right|^{\beta}\right], \tag{4.28}$$

with

$$\zeta_n^2 = \frac{\sigma_n^2 \Gamma(1/\beta)}{\Gamma(3/\beta)}. \tag{4.29}$$

In these equations, μ_n, σ_n^2, and Γ are respectively the mean and variance of a given photon number cluster and the Gamma function. In Fig. 4.9a we present a qualitative representation of the fit quality of t-SNE embedding using the standard and generalized Gaussians functions.

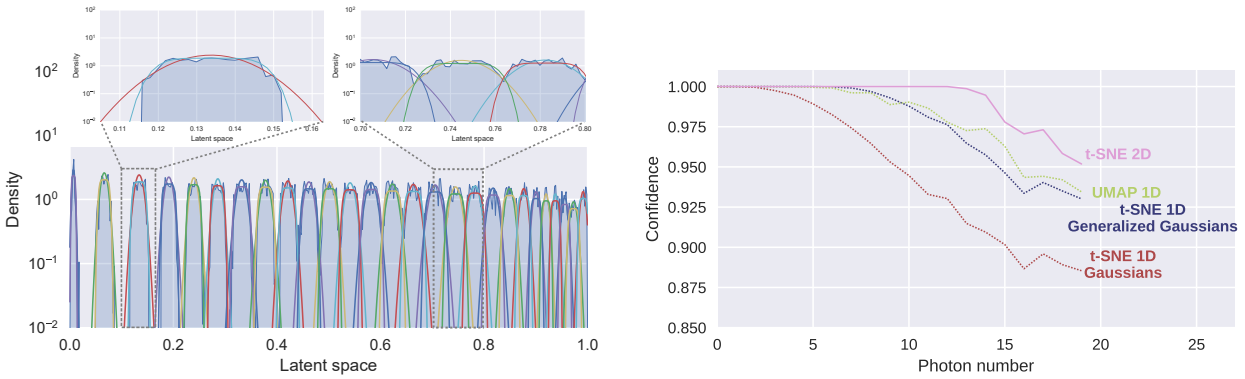
In Fig. 4.9a we see that the generalized Gaussian distribution is a better estimation of the density inside the latent space. The small tail reduces greatly the overlap of probability density functions, which increases the computed confidence. The new values of confidence are plotted in Fig. 4.9b where we observe a significant increase in the confidence, reaching values similar to one-dimensional UMAP.

4.7.7 Potential Applications

Based on the benchmarks, the dimensionality reduction techniques that focus on local structure preservation offer the best low-dimensional representation of the transition edge sensor signals.

These methods provide high cluster separation and follow the expected distribution to a degree unmatched by other techniques. For these reasons, t-SNE and UMAP are effective methods for applications that do not involve frequently adding new samples to their dataset and require high accuracy. The existence of open-source platforms like UMAP-learn [107] and Scikit-Learn [109] that offer complete and optimized implementations of these methods facilitates its usability. The number of user parameters necessary to use these methods is also very small, which makes them ideal for experts and non-experts. We note that the operation complexity scaling of UMAP is much more advantageous when compared to t-SNE, it is therefore more efficient to use UMAP since they both have similar performances.

Considering the previous performance results, neural networks (5 linear layers of 300 neurons) can offer a trustworthy and interpretable low-dimensional representation of the TES traces. The condition necessary for this method to be accurate is to provide a balanced dataset containing the range of photon numbers we want to detect. It is essential to understand that the network cannot predict photon number outside the trained range, since it never learned an embedding for these signals. The training data restricts the learned transformation. Moreover, our results suggest that a small neural network implemented in a Field Programmable Gate Array (FPGA) [110] could replace currently used methods like trace area and PCA [23] to process the TES traces directly. With this type of hardware, we believe real-time processing can be achieved considering TESs have a dead time of a few microseconds and knowing our CPU implementation can process a TES signal of 200 points in $4.9\mu s$. This value is obtained using a laptop with a clockspeed of 3.2 GHz, 8 cores and 16



(a) Gaussian and generalized Gaussian fit over the kernel density estimation of the t-SNE 1D embedding, where the broader distributions are the standard Gaussians.

(b) Confidence associated with the one-dimensional embedding from t-SNE using standard and generalized Gaussian functions to describe the clusters.

Figure 4.9 Impact of using a generalized Gaussian function to estimate the clusters generated by t-SNE.

threads.

We emphasized using a close-to-uniform distribution to train the network, since it becomes equally optimized for every class (photon number). Following the example used to benchmark the different methods, the use of a coherent source with tunable mean photon number is more than sufficient to create a balanced dataset. It is therefore possible to create suitable conditions only using a laser and tunable attenuation. We could also imagine using a high mean photon number thermal source, depending on the available equipment.

Coming back to the use of methods that preserve local structures, we believe that using methods like UMAP can enable the use of TESs on temporally uncorrelated light, making it a useful tool to remove noise in a variety of cases. Also, this feature can be exploited to characterize photon statistics of continuous-wave sources where no time trigger can be used. Existing work on the topic [111] uses a different approach to this problem, making it difficult to compare. However, the methods we describe make this task simple to implement for a wide variety of cases, since it is invariant to the combinations of photon events inside a single signal. In other words, traces associated with exotic scenarios, for example a single photon trace slightly overlapped by a two-photon trace, should have its position inside the latent space making it distinguishable. This task is also well suited for neural networks since they can be designed to be shift-invariant, meaning that similar structures, independent of their position, could be clustered.

4.7.8 Future work

A one-dimensional embedding is optimal for experimental systems where the number of possible outliers is limited since the clustering task becomes simplified. To improve on this work, we hypothesize that there is a solution in one dimension that can reach the confidence values of two-dimensional UMAP and t-SNE. To address this problem, we could enhance our understanding by examining the relationship between the dimensionality reduction process and clustering. Additionally, it may be possible to strengthen the representation of photon numbers while minimizing the space allocated to noise features.

While testing the different methods, the clustering step (Gaussian Mixture Model) was particularly sensitive to the initialization process. Often some manual adjustments had to be done to guarantee the quality of the results. To further improve the quality and robustness of photon number classification, future work could explore clustering techniques that may be better aligned with the novel methods introduced in this study. This way it could be possible to completely automate the photon number classification process even for low visibility clusters.

Accessing the ground truth in the case of photon number classification remains difficult, and further validation would be desirable to guarantee the performance of the proposed methods. We propose using the joint probability distribution of photon pairs to benchmark dimensionality reduction methods. In more detail, in a perfect system, photon pairs should have the same photon number in both modes. Experimentally, loss and misassigned photons broaden the joint probability distribution, which would be otherwise perfectly diagonal [112]. With this in mind, we expect broadening effects associated with the numerical analysis. This way, the width of the joint probability distribution becomes an experimental tool to quantify the performance of numerical techniques.

4.8 Conclusion

Nonlinear methods like t-SNE and UMAP that aim to preserve local data structures offer better resolution over photon numbers in the case of transition edge sensor signals compared to previously used techniques like signal area and PCA. These methods can be used directly to replace currently used methods, with the caveat that they cannot predict new samples without computing the entire dataset.

With a large dataset ($u = 550\,000$ samples), we demonstrate the potential of neural network that recreate the embedding of t-SNE and UMAP. These models remain simple and could be further explored, offering a promising direction for future research. Enhancing the

generalization capabilities of these models could enable their application in real-time photon number resolution, advancing the field of quantum optics.

Beyond TES devices, the techniques explored in this work hold promise for enhancing the performance of other single-photon detectors, such as SNSPDs. For instance, principal component analysis (PCA) has shown potential in processing SNSPD signals [22,97], highlighting the versatility of these approaches across photon-detection technologies.

All the numerical methods and data discussed in this document are available in Ref. [103, 104, 106].

Acknowledgements

N.D.-C. and N.Q. acknowledge support from the Ministère de l'Économie et de l'Innovation du Québec, the Natural Sciences and Engineering Research Council Canada, Photonique Quantique Québec, and thank S. Montes-Valencia, J. Martínez-Cifuentes and A. Boon for valuable discussions. We also thank Z. Levine and S. Glancy for their careful feedback on our manuscript.

CHAPTER 5 CONCLUSION

5.1 Computational Considerations

While the performance of the techniques is discussed in terms of the confidence in previous sections, it is interesting to have a notion of the time required to execute the different algorithms. In Fig. 5.1 this time is presented for the different techniques, the computer used to execute the code has a CPU with up to 6 GHz of clock speed, 24 cores, and 32 threads. All algorithms are executed on a CPU in this test, even if techniques utilizing neural networks can be computed much faster on a GPU.

Table 5.1 Comparison of the time required to process the Synthetic Uniform and Synthetic Geometric datasets for all the dimensionality reduction techniques (once trained).

Name	Time (s) Synthetic Uniform	Time (s) Synthetic Geometric
Maximum value	0.191	0.335
Area	1.209	2.259
PCA 1D	0.055	0.074
PCA 2D	0.050	0.087
KPCA RBF 2D	20.210	71.617
KPCA Sigmoid 1D	20.456	71.923
KPCA Cosine 2D	12.288	42.817
tSNE 1D (perplexity of 450)	122.950	262.758
tSNE 2D (perplexity of 450)	129.805	260.479
UMAP 1D (1000 neighbors)	251.482	468.574
UMAP 2D (700 neighbors)	165.661	321.298
NMF 1D	2.090	5.477
Isomap 1D	1160.342	1072.104
PTSNE 1D	0.050	0.130
PUMAP 1D	0.061	0.152

The scripts are executed using a random seed to assure reproducibility, however this reduces the speed of the UMAP implementation since the script becomes single threaded when using this mode. Additionally, the number of neighbours considered has a significant impact on the run time. When removing the requirement for a random seed and selecting 450 neighbours, UMAP 1D runs in 78.155 s compared to 127.015 s for tSNE 1D with the same considerations. As a general rule UMAP should be used instead of tSNE since it has better complexity scaling.

5.2 Application of Neural networks for TES signals

In Chap. 4 it is demonstrated that parametric implementations of UMAP and t-SNE offer a latent space where photon number clusters are easier to distinguish. While this result is interesting, neural networks offer a platform with more potential than simply creating interpretable latent spaces. For instance, in Chap. 4 only the use of an encoder is discussed, however a decoder can easily be added to resemble the architecture of a VAE. In this structure the encoder provides a space to assign standard photon numbers and the decoder can be used for anomaly rejection following the same strategy as Ref. [71]. Precisely, this configuration can be trained on "clean" data, learning an accurate transformation to encode and decode the desired signals. With this trained network, clustering can be done on the encoder output to assign photon numbers and by defining a threshold on the reconstruction loss discard signals that the network does not recognize (cannot reconstruct correctly).

Another configuration that is compatible with the previously described autoencoder is to use shift invariant layers in the construction of the network. In this configuration, the network can interpret photon numbers independently of the time of arrival of the photons. This network could be relevant for instance in the detection of light coming from a continuous light source where the time of arrival is unknown.

In the opposite case, the deliberate decision to make the network sensitive to shifts can help discard unwanted light that hits the detector at random times. This can be explained by considering that the network learns to encode and decode signals at the trigger time and never outside this time frame. The network being unable to recreate such a signal will interpret the signal as unwanted since the reconstructed signal does not resemble the input.

5.3 Summary of Works

This work discusses the problem of photon number assignment in the case of transition edge sensors. To increase the performance of such detectors, this work explores the use of dimensionality reduction techniques that transform the high-dimensional data (signals) into a low-dimensional representation where the features of interest appear from data structures. The comparison of techniques demonstrated that some nonlinear techniques can offer an embedding for TES signals where photon numbers are easier to distinguish. The techniques that show potential have never been used in this context and show a new approach for researchers to analyse data generated from TESs.

The comparison of techniques is done in a quantitative manner, following a confidence metric (Sec. 4.5.6) that gives the probability of correctly assigning a photon number based on a latent

space.

While t-SNE and UMAP are powerful tools, they do not offer a platform to predict the class of new samples. This work demonstrates that neural networks can learn an embedding of TES signals that closely approximate the latent space of t-SNE and UMAP. From these transformations, the prediction of new samples is possible with greater confidence than previously used techniques.

The entire work was designed with reproducibility in mind. This translates into the creation of a public repository where all the code necessary to reproduce the results is available [106]. Also, a major part of this work has been to access data from TESs since this type of detector is expensive and only used by a few research groups in the world. From collaboration with NIST and NRC the creation of two public data repositories was achieved [103, 104] making the comparison of analysis techniques accessible to the full scientific community.

5.4 Limitations

While interesting results came out of this exploration of dimensionality reduction techniques, some problems remain. First, the use of a one-dimensional embedding simplifies clustering, but does so at the cost of potentially losing critical information necessary for resolving photon numbers in more complex scenarios.

Additionally, the Gaussian Mixture Model (GMM) clustering step exhibited sensitivity to initialization, occasionally requiring manual intervention to achieve optimal results. This is inconvenient considering the desire to create a fully automated photon-number classification pipeline.

Furthermore, the entire work is done on data without accessing the ground truth, which is an intrinsic weakness of the methodology. Indeed, the only physical validation in this work is done via the evaluation of the $g^{(2)}$ which does not give a direct confirmation that an individual signal is correctly classified. Moreover, the performance evaluation only considers the distinguishability of clusters inside an abstract space. The methodology is therefore sensitive to techniques that can create artificial clusters.

Further experimental validation would be helpful to guarantee the reported performance of the discussed dimensionality reduction techniques. One approach that is previously described in Chap. 4 is the use of photon pairs. Indeed, in a perfect system, the joint probability distribution of photons from both modes should be perfectly diagonal. The presence of loss in the real implementation and misassigned photons in the post-processing step result in a broadening of this distribution. Considering the average loss is constant, the width of

this joint probability distribution can become an experimental validation for the numerical analysis.

5.5 Future Research

To overcome the limitations of one-dimensional representations, future work could explore whether more robust encoding schemes can achieve the performance of two-dimensional t-SNE and UMAP. This involves investigating the relation between dimensionality reduction and clustering to find solutions that better represent photon numbers while minimizing noise contributions.

To enhance automation and robustness, research into alternative clustering techniques, such as density-based or deep-learning-based clustering, is possible. These methods could reduce sensitivity to initialization and handle low-visibility clusters more effectively.

The width of the joint probability distribution in photon pair experiments offers an experimental playground for benchmarking numerical techniques. Future work could focus on systematically analysing broadening effects to establish quantitative metrics for method performance, enabling more rigorous validation.

Although this work demonstrates the feasibility of neural networks to emulate t-SNE and UMAP embeddings, future efforts could aim to improve the generalization capabilities of these models. This includes exploring architecture designs or transfer learning approaches to enable real-time photon number resolution, which would have significant implications for quantum optics and related fields.

Finally, the techniques proposed in this work could potentially benefit other single photon detection technologies, such as SNSPDs. By using methods like t-SNE, UMAP, and PCA to these devices, it may be possible to improve their performance.

Continuing to expand and refine the publicly available datasets and codebases will facilitate further development and adoption of the proposed methods. Future work should prioritize creating comprehensive benchmarks and ensuring the reproducibility of numerical techniques across the community.

By addressing these limitations and pursuing the outlined research directions, this work lays the foundation for more robust, automated, and generalizable photon number classification methods, advancing the state-of-the-art in single-photon detection technologies.

REFERENCES

[1] J. M. Arrazola, V. Bergholm, K. Brádler, T. R. Bromley, M. J. Collins, I. Dhand, A. Fumagalli, T. Gerrits, A. Goussev, L. G. Helt, J. Hundal, T. Isacsson, R. B. Israel, J. Izaac, S. Jahangiri, R. Janik, N. Killoran, S. P. Kumar, J. Lavoie, A. E. Lita, D. H. Mahler, M. Menotti, B. Morrison, S. W. Nam, L. Neuhaus, H. Y. Qi, N. Quesada, A. Repingon, K. K. Sabapathy, M. Schuld, D. Su, J. Swinarton, A. Száva, K. Tan, P. Tan, V. D. Vaidya, Z. Vernon, Z. Zabaneh, and Y. Zhang, "Quantum circuits with many photons on a programmable nanophotonic chip," *Nature*, vol. 591, no. 7848, pp. 54–60, Mar 2021. [Online]. Available: https://doi.org/10.1038/s41586-021-03202-1

[2] S. Slussarenko and G. J. Pryde, "Photonic quantum information processing: A concise review," *Appl. Phys. Rev.*, vol. 6, no. 4, Oct. 2019. [Online]. Available: http://dx.doi.org/10.1063/1.5115814

[3] T. Rudolph, "Why I am optimistic about the silicon-photonic route to quantum computing," *APL photonics*, vol. 2, no. 3, 2017.

[4] J. E. Bourassa, R. N. Alexander, M. Vasmer, A. Patil, I. Tzitrin, T. Matsuura, D. Su, B. Q. Baragiola, S. Guha, G. Dauphinais *et al.*, "Blueprint for a scalable photonic fault-tolerant quantum computer," *Quantum*, vol. 5, p. 392, 2021.

[5] N. Maring, A. Fyrillas, M. Pont, E. Ivanov, P. Stepanov, N. Margaria, W. Hease, A. Pishchagin, A. Lemaître, I. Sagnes *et al.*, "A versatile single-photon-based quantum computing platform," *Nat. Photonics*, vol. 18, no. 6, pp. 603–609, 2024.

[6] K. Takase, F. Hanamura, H. Nagayoshi, J. E. Bourassa, R. N. Alexander, A. Kawasaki, W. Asavanant, M. Endo, and A. Furusawa, "Generation of flying logical qubits using generalized photon subtraction with adaptive gaussian operations," *Phys. Rev. A*, vol. 110, no. 1, p. 012436, 2024.

[7] K. Alexander, A. Bahgat, A. Benyamini, D. Black, D. Bonneau, S. Burgos, B. Burridge, G. Campbell, G. Catalano, A. Ceballos *et al.*, "A manufacturable platform for photonic quantum computing," *arXiv preprint arXiv:2404.17570*, 2024.

[8] Y. Yao, F. Miatto, and N. Quesada, "Riemannian optimization of photonic quantum circuits in phase and Fock space," *Scipost Phys.*, vol. 17, no. 3, p. 082, Sep. 2024.

[9] Y.-R. Chen, H.-Y. Hsieh, J. Ning, H.-C. Wu, H. L. Chen, Z.-H. Shi, P. Yang, O. Steuernagel, C.-M. Wu, and R.-K. Lee, "Generation of heralded optical cat states by photon addition," *Phys. Rev. A*, vol. 110, no. 2, p. 023703, 2024.

[10] M. Melalkia, J. Huynh, S. Tanzilli, V. d'Auria, and J. Etesse, "A multiplexed synthesizer for non-gaussian photonic quantum state generation," *Quantum Sci. Technol.*, vol. 8, no. 2, p. 025007, 2023.

[11] J. Tiedau, T. J. Bartley, G. Harder, A. E. Lita, S. W. Nam, T. Gerrits, and C. Silberhorn, "Scalability of parametric down-conversion for generating higher-order fock states," *Phys. Rev. A*, vol. 100, p. 041802, Oct 2019. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevA.100.041802

[12] T. Sonoyama, K. Takahashi, T. Sano, T. Suzuki, T. Nomura, M. Yabuno, S. Miki, H. Terai, K. Takase, W. Asavanant, M. Endo, and A. Furusawa, "Generation of multiphoton Fock states at telecommunication wavelength using picosecond pulsed light," *Opt. Express*, vol. 32, no. 18, pp. 32 387–32 395, Aug. 2024.

[13] M. Endo, K. Takahashi, T. Nomura, T. Sonoyama, M. Yabuno, S. Miki, H. Terai, T. Kashiwazaki, A. Inoue, T. Umeki *et al.*, "Optically-sampled superconducting-nanostrip photon-number resolving detector for non-classical quantum state generation," *arXiv preprint arXiv:2405.06901*, 2024.

[14] S. Aaronson and A. Arkhipov, "The computational complexity of linear optics," in *Proceedings of the forty-third annual ACM symposium on Theory of computing*, 2011, pp. 333–342.

[15] C. S. Hamilton, R. Kruse, L. Sansoni, S. Barkhofen, C. Silberhorn, and I. Jex, "Gaussian boson sampling," *Phys. Rev. Lett.*, vol. 119, no. 17, p. 170501, 2017.

[16] R. Kruse, J. Tiedau, T. J. Bartley, S. Barkhofen, and C. Silberhorn, "Limits of the time-multiplexed photon-counting method," *Phys. Rev. A*, vol. 95, no. 2, p. 023815, 2017.

[17] A. Deshpande, A. Mehta, T. Vincent, N. Quesada, M. Hinsche, M. Ioannou, L. Madsen, J. Lavoie, H. Qi, J. Eisert, D. Hangleiter, B. Fefferman, and I. Dhand, "Quantum computational advantage via high-dimensional Gaussian boson sampling," *Sci. Adv.*, vol. 8, no. 1, p. eabi7894, Jan. 2022.

[18] D. Grier, D. J. Brod, J. M. Arrazola, M. B. de Andrade Alonso, and N. Quesada, "The complexity of bipartite gaussian boson sampling," *Quantum*, vol. 6, p. 863, 2022.

[19] L. S. Madsen, F. Laudenbach, M. F. Askarani, F. Rortais, T. Vincent, J. F. Bulmer, F. M. Miatto, L. Neuhaus, L. G. Helt, M. J. Collins *et al.*, "Quantum computational advantage with a programmable photonic processor," *Nature*, vol. 606, no. 7912, pp. 75–81, 2022.

[20] G. Thekkadath, M. Mycroft, B. Bell, C. Wade, A. Eckstein, D. Phillips, R. Patel, A. Buraczewski, A. Lita, T. Gerrits *et al.*, "Quantum-enhanced interferometry with large heralded photon-number states," *npj Quantum Inf.*, vol. 6, no. 1, p. 89, 2020.

[21] C. F. Wildfeuer, A. J. Pearlman, J. Chen, J. Fan, A. Migdall, and J. P. Dowling, "Interferometry with a photon-number resolving detector*," in *Conference on Lasers and Electro-Optics/International Quantum Electronics Conference (2009), Paper IWF1.* Optica Publishing Group, May 2009, p. IWF1.

[22] A. Divochiy, F. Marsili, D. Bitauld, A. Gaggero, R. Leoni, F. Mattioli, A. Korneev, V. Seleznev, N. Kaurova, O. Minaeva, G. Gol'tsman, K. G. Lagoudakis, M. Benkhaoul, F. Lévy, and A. Fiore, "Superconducting nanowire photon-number-resolving detector at telecommunication wavelengths," *Nat. Photonics*, vol. 2, no. 5, pp. 302–306, May 2008. [Online]. Available: https://doi.org/10.1038/nphoton.2008.51

[23] L. A. Morais, T. Weinhold, M. P. de Almeida, J. Combes, M. Rambach, A. Lita, T. Gerrits, S. W. Nam, A. G. White, and G. Gillett, "Precisely determining photon-number in real time," *Quantum*, vol. 8, p. 1355, May 2024.

[24] M. Jönsson and G. Björk, "Evaluating the performance of photon-number-resolving detectors," *Phys. Rev. A*, vol. 99, no. 4, p. 043822, 2019.

[25] M. Jönsson, M. Swillo, S. Gyger, V. Zwiller, and G. Björk, "Temporal array with superconducting nanowire single-photon detectors for photon-number resolution," *Phys. Rev. A*, vol. 102, no. 5, p. 052616, 2020.

[26] M. Eaton, A. Hossameldin, R. J. Birrittella, P. M. Alsing, C. C. Gerry, H. Dong, C. Cuevas, and O. Pfister, "Resolution of 100 photons and quantum generation of unbiased random numbers," *Nat. Photonics*, vol. 17, no. 1, pp. 106–111, Jan. 2023.

[27] K. Irwin and G. Hilton, "Transition-edge sensors," in *Cryogenic Particle Detection*, C. Enss, Ed. Springer, 2005, pp. 63–150. [Online]. Available: https://doi.org/10.1007/10933596_3

[28] D. S. Phillips, "Advanced measurements for quantum photonics and quantum technologies," Ph.D. dissertation, University of Oxford, 2020.

[29] R. H. Hadfield, "Single-photon detectors for optical quantum information applications," *Nat. Photonics*, vol. 3, no. 12, pp. 696–705, 2009.

[30] T. Gerrits, B. Calkins, N. Tomlin, A. E. Lita, A. Migdall, R. Mirin, and S. W. Nam, "Extending single-photon optimized superconducting transition edge sensors beyond the single-photon counting regime," *Opt. Express*, vol. 20, no. 21, pp. 23 798–23 810, Oct 2012. [Online]. Available: https://opg.optica.org/oe/abstract.cfm?URI=oe-20-21-23798

[31] M. Schmidt, I. H. Grothe, S. Neumeier, L. Bremer, M. von Helversen, W. Zent, B. Melcher, J. Beyer, C. Schneider, S. Höfling, J. Wiersig, and S. Reitzenstein, "Bimodal behavior of microlasers investigated with a two-channel photon-number-resolving transition-edge sensor system," *Phys. Rev. Res.*, vol. 3, p. 013263, Mar 2021. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevResearch.3.013263

[32] P. C. Humphreys, B. J. Metcalf, T. Gerrits, T. Hiemstra, A. E. Lita, J. Nunn, S. W. Nam, A. Datta, W. S. Kolthammer, and I. A. Walmsley, "Tomography of photon-number resolving continuous-output detectors," *New J. Phys.*, vol. 17, no. 10, p. 103044, oct 2015. [Online]. Available: https://dx.doi.org/10.1088/1367-2630/17/10/103044

[33] Z. H. Levine, T. Gerrits, A. L. Migdall, D. V. Samarov, B. Calkins, A. E. Lita, and S. W. Nam, "Algorithm for finding clusters with a known distribution and its application to photon-number resolution using a superconducting transition-edge sensor," *J. Opt. Soc. Am. B*, vol. 29, no. 8, pp. 2066–2073, Aug 2012. [Online]. Available: https://opg.optica.org/josab/abstract.cfm?URI=josab-29-8-2066

[34] S. Rajendran, W. Meert, D. Giustiniano, V. Lenders, and S. Pollin, "Deep learning models for wireless signal classification with distributed low-cost spectrum sensors," *IEEE Transactions on Cognitive Communications and Networking*, vol. 4, no. 3, pp. 433–445, 2018.

[35] H. P. Nautrup, N. Delfosse, V. Dunjko, H. J. Briegel, and N. Friis, "Optimizing quantum error correction codes with reinforcement learning," *Quantum*, vol. 3, p. 215, Dec. 2019. [Online]. Available: http://dx.doi.org/10.22331/q-2019-12-16-215

[36] D. Jin and J. Yuan, "Fitting an Elephant with Four non-Zero Parameters," Jul. 2024.

[37] J. Mayer, K. Khairy, and J. Howard, "Drawing an elephant with four complex parameters," *American Journal of Physics*, vol. 78, no. 6, pp. 648–649, Jun. 2010.

[38] "Seaborn.kdeplot — seaborn 0.13.2 documentation," https://seaborn.pydata.org/generated/seaborn.kdeplot.html, 2024-10-23.

[39] D. Fukuda, G. Fujii, T. Numata, K. Amemiya, A. Yoshizawa, H. Tsuchida, H. Fujino, H. Ishii, T. Itatani, S. Inoue, and T. Zama, "Titanium-based transition-edge photon number resolving detector with 98% detection efficiency with index-matched small-gap fiber coupling," *Opt. Express*, vol. 19, no. 2, pp. 870–875, Jan 2011. [Online]. Available: https://opg.optica.org/oe/abstract.cfm?URI=oe-19-2-870

[40] N. Dalbec-Constant, G. Thekkadath, D. England, B. Sussman, T. Gerrits, and N. Quesada, "Accurate Unsupervised Photon Counting from Transition Edge Sensor Signals," Nov. 2024.

[41] G. S. Thekkadath, "Preparing and characterizing quantum states of light using photon-number-resolving detectors," http://purl.org/dc/dcmitype/Text, University of Oxford, 2020.

[42] L. A. Morais, "Probing the Unknown: Measurements in the Quantum Realm."

[43] "Short Article - Quantum Efficiency," https://www.photometrics.com/learn/imaging-topics/quantum-efficiency.

[44] A. Lamas-Linares, B. Calkins, N. A. Tomlin, T. Gerrits, A. E. Lita, J. Beyer, R. P. Mirin, and S. Woo Nam, "Nanosecond-scale timing jitter for single photon detection in transition edge sensors," *Applied Physics Letters*, vol. 102, no. 23, p. 231117, Jun. 2013.

[45] L. You, "Superconducting nanowire single-photon detectors for quantum information," *Nanophotonics*, vol. 9, no. 9, pp. 2673–2692, Sep. 2020.

[46] A. Goetz, "The Possible Use of Superconductivity for Radiometric Purposes," *Physical Review*, vol. 55, no. 12, pp. 1270–1271, Jun. 1939.

[47] D. H. Andrews, W. F. Brucksch, W. T. Ziegler, and E. R. Blanchard, "Superconducting Films as Radiometric Receivers," *Physical Review*, vol. 59, no. 12, pp. 1045–1046, Jun. 1941.

[48] D. H. Andrews, W. F. Brucksch, Jr., W. T. Ziegler, and E. R. Blanchard, "Attenuated Superconductors I. For Measuring Infra-Red Radiation," *Review of Scientific Instruments*, vol. 13, no. 7, pp. 281–292, Jul. 1942.

[49] D. H. Andrews, R. D. Fowler, and M. C. Williams, "The Effect of Alpha-particles on a Superconductor," *Physical Review*, vol. 76, no. 1, pp. 154–155, Jul. 1949.

[50] K. M. Morgan, "Hot science with cool sensors," *Physics Today*, vol. 71, no. 8, pp. 28–34, Aug. 2018.

[51] W. Seidel, G. Forster, W. Christen, F. von Feilitzsch, H. Göbel, F. Pröbst, and R. L. Mößbauer, "Phase transition thermometers with high temperature resolution for calorimetric particle detectors employing dielectric absorbers," *Physics Letters B*, vol. 236, no. 4, pp. 483–487, Mar. 1990.

[52] K. D. Irwin, S. W. Nam, B. Cabrera, B. Chugg, G. Park, R. P. Welty, and J. M. Martinis, "A self-biasing cryogenic particle detector utilizing electrothermal feedback and a squid readout," *IEEE Transactions on Applied Superconductivity*, vol. 5, pp. 2690–2693, 1995. [Online]. Available: https://api.semanticscholar.org/CorpusID:9547941

[53] B. Cabrera, R. Clarke, A. Miller, S. W. Nam, R. Romani, T. Saab, and B. Young, "Cryogenic detectors based on superconducting transition-edge sensors for time-energy-resolved single-photon counters and for dark matter searches," *Physica B: Condensed Matter*, vol. 280, no. 1, pp. 509–514, May 2000.

[54] S. R. Bandler, E. Figueroa-Feliciano, C. K. Stahle, K. Boyce, R. Brekosky, J. Chervenak, F. Finkbeiner, R. Kelley, M. Lindeman, F. S. Porter, and T. Saab, "Design of transition edge sensor microcalorimeters for optimal performance," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 520, no. 1, pp. 285–288, Mar. 2004.

[55] N. Mujica-schwahn, "Position Dependence of High Efficiency Single Photon Detectors: A Route to Better Understanding of Transition Edge Sensors."

[56] H. F. C. Hoevers, "Thermal physics of transition edge sensor arrays," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 559, no. 2, pp. 702–705, Apr. 2006.

[57] J. E. Sadleir, "Superconducting Transition-Edge Sensor Physics," Ph.D. dissertation, University of Illinois at Urbana-Champaign, Jan. 2011.

[58] A. Kozorezov, A. A. Golubov, D. D. E. Martin, P. A. J. de Korte, M. A. Lindeman, R. A. Hijmering, J. van der Kuur, H. F. C. Hoevers, L. Gottardi, M. Yu. Kupriyanov, and J. K. Wigmore, "Modelling the resistive state in a transition edge sensor," *Applied Physics Letters*, vol. 99, no. 6, p. 063503, Aug. 2011.

[59] M. Lorenz, C. Kirsch, P. E. Merino-Alonso, P. Peille, T. Dauser, E. Cucchetti, S. J. Smith, and J. Wilms, "GPU Supported Simulation of Transition-Edge Sensor Arrays," *Journal of Low Temperature Physics*, vol. 200, no. 5, pp. 277–285, Sep. 2020.

[60] K. Hattori, T. Konno, Y. Miura, S. Takasu, and D. Fukuda, "An optical transition-edge sensor with high energy resolution," *Superconductor Science and Technology*, vol. 35, no. 9, p. 095002, Jul. 2022.

[61] D. A. Bennett, R. D. Horansky, A. S. Hoover, N. J. Hoteling, M. W. Rabin, D. R. Schmidt, D. S. Swetz, L. R. Vale, and J. N. Ullom, "An analytical model for pulse shape and electrothermal stability in two-body transition-edge sensor microcalorimeters," *Applied Physics Letters*, vol. 97, no. 10, p. 102504, Sep. 2010.

[62] D. Fukuda, H. Takahashi, Y. Kunieda, N. Zen, M. Ohkubo, and M. Nakazawa, "Noise and signal analysis of Ir/Au TES with asymmetrical slits parallel to the electric current," *IEEE Transactions on Applied Superconductivity*, vol. 15, no. 2, pp. 522–525, Jun. 2005.

[63] D. Alberto, M. Rajteri, E. Taralli, L. Lolli, C. Portesi, E. Monticone, Y. Jia, R. Garello, and M. Greco, "Optical Transition-Edge Sensors Single Photon Pulse Analysis," *IEEE Transactions on Applied Superconductivity*, vol. 21, no. 3, pp. 285–288, Jun. 2011.

[64] D. Alberto, "Digital Signal Processing applied to Physical Signals," Ph.D. dissertation, INFN, Turin, 2011.

[65] "Single-Photon Avalanche Diode (SPADs) | MEETOPTICS Academy," https://www.meetoptics.com/academy/single-photon-avalanche-diode?srsltid=AfmBOoqgq9wXQoFCtFfiRZlEzLmz0_780EHdkeyTUiimLvFmqs8gBd5#silicon-photomultipliers-&-multiple-pixel-photon-counters.

[66] J. Lee, L. Shen, A. Cerè, T. Gerrits, A. E. Lita, S. W. Nam, and C. Kurtsiefer, "Multipulse fitting of transition edge sensor signals from a near-infrared continuous-wave source," *Review of Scientific Instruments*, vol. 89, no. 12, p. 123108, Dec. 2018.

[67] K. P. Sinaga and M.-S. Yang, "Unsupervised K-Means Clustering Algorithm," *IEEE Access*, vol. 8, pp. 80 716–80 727, 2020.

[68] D. Arthur and S. Vassilvitskii, "K-means++: The Advantages of Careful Seeding."

[69] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Information Sciences*, vol. 622, pp. 178–210, Apr. 2023.

[70] Z. H. Levine, T. Gerrits, A. L. Migdall, D. V. Samarov, B. R. Calkins, A. E. Lita, and S. W. Nam, "An algorithm for finding clusters with a known distribution and its application to photon-number resolution using a superconducting transition-edge sensor," *NIST*, vol. 29, no. 8, pp. 2066–2073, Jul. 2012.

[71] Y. Ichinohe, S. Yamada, R. Hayakawa, S. Okada, T. Hashimoto, H. Tatsuno, H. Suda, and T. Okumura, "Application of Deep Learning to the Evaluation of Goodness in the Waveform Processing of Transition-Edge Sensor Calorimeters," *Journal of Low Temperature Physics*, vol. 209, no. 5, pp. 1008–1016, Dec. 2022.

[72] Y.-C. Jeong, K.-H. Hong, and Y.-H. Kim, "Bright source of polarization-entangled photons using a PPKTP pumped by a broadband multi-mode diode laser," *Optics Express*, vol. 24, no. 2, pp. 1165–1174, Jan. 2016.

[73] L. J. Salazar, D. A. Guzmán, F. J. Rodríguez, and L. Quiroga, "Quantum-correlated two-photon transitions to excitons in semiconductor quantum wells," *Optics Express*, vol. 20, no. 4, p. 4470, Feb. 2012.

[74] G. Graciani and F. Amblard, "Super-resolution provided by the arbitrarily strong superlinearity of the blackbody radiation," *Nature Communications*, vol. 10, p. 5761, Dec. 2019.

[75] "Hanbury Brown and Twiss effect," *Wikipedia*, May 2024.

[76] H. Brown, "6.1 Introduction: The intensity interferometer."

[77] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain." *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958.

[78] "Linear/Fully-Connected Layers User's Guide," https://docs.nvidia.com/deeplearning/performance/d
performance-fully-connected/index.html.

[79] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, Jan. 1989.

[80] G. Gripenberg, "Approximation by neural networks with a bounded number of nodes at each level," *Journal of Approximation Theory*, vol. 122, no. 2, pp. 260–266, Jun. 2003.

[81] F. Dyson, "A meeting with Enrico Fermi," *Nature*, vol. 427, no. 6972, pp. 297–297, Jan. 2004.

[82] "Fitting an Elephant | Wolfram Demonstrations Project," https://demonstrations.wolfram.com/FittingAnElephant/.

[83] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, Dec. 1943.

[84] "Convolutional neural network," *Wikipedia*, Nov. 2024.

[85] "Chapter 5: Introduction to Convolutional Neural Networks," https://www.tomasbeuzen.com/deep-learning-with-pytorch/chapters/chapter5_cnns-pt1.html.

[86] A. Chaman and I. Dokmanic, "Truly shift-invariant convolutional neural networks," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, TN, USA: IEEE, Jun. 2021, pp. 3772–3782.

[87] A. Chaubey, "Downsampling and Upsampling of Images — Demystifying the Theory," Jan. 2024.

[88] A. Anwar, "Difference between AutoEncoder (AE) and Variational AutoEncoder (VAE)," https://towardsdatascience.com/difference-between-autoencoder-ae-and-variational-autoencoder-vae-ed7be1c038f2, Nov. 2021.

[89] T. Gerrits, S. Glancy, T. S. Clement, B. Calkins, A. E. Lita, A. J. Miller, A. L. Migdall, S. W. Nam, R. P. Mirin, and E. Knill, "Generation of optical coherent-state superpositions by number-resolved photon subtraction from the squeezed vacuum," *Physical Review A*, vol. 82, no. 3, p. 031802, Sep. 2010.

[90] C. You, M. Hong, P. Bierhorst, A. E. Lita, S. Glancy, S. Kolthammer, E. Knill, S. W. Nam, R. P. Mirin, O. S. Magaña-Loaiza, and T. Gerrits, "Scalable multiphoton quantum metrology with neither pre- nor post-selected measurements," *Applied Physics Reviews*, vol. 8, no. 4, p. 041406, Oct. 2021.

[91] M. Allaoui, M. L. Kherfi, and A. Cheriet, "Considerably improving clustering algorithms using UMAP dimensionality reduction technique: A comparative study," in *Image and Signal Processing*, A. El Moataz, D. Mammass, A. Mansouri, and F. Nouboud, Eds. Springer International Publishing, 2020, pp. 317–325.

[92] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 86, pp. 2579–2605, 2008.

[93] F. Jelinek, R. L. Mercer, L. R. Bahl, and J. K. Baker, "Perplexity—a measure of the difficulty of speech recognition tasks," *J. Acoust. Soc. Am.*, vol. 62, no. S1, pp. S63–S63, 08 2005. [Online]. Available: https://doi.org/10.1121/1.2016299

[94] L. McInnes, J. Healy, N. Saul, and L. Großberger, "UMAP: Uniform Manifold Approximation and Projection," *J. Open Source Softw.*, vol. 3, no. 29, p. 861, Sep. 2018.

[95] W. Dong, M. Charikar, and K. Li, "Efficient k-nearest neighbor graph construction for generic similarity measures," in *The Web Conference*, 2011. [Online]. Available: https://api.semanticscholar.org/CorpusID:207186093

[96] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science (New York, N.Y.)*, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000.

[97] T. Schapeler, N. Lamberty, T. Hummel, F. Schlue, M. Stefszky, B. Brecht, C. Silberhorn, and T. J. Bartley, "Electrical trace analysis of superconducting nanowire photon-number-resolving detectors," *Phys. Rev. Appl.*, vol. 22, p. 014024, 2024.

[98] I. Jolliffe, *Mathematical and Statistical Properties of Sample Principal Components*. New York, NY: Springer New York, 2002, pp. 29–61. [Online]. Available: https://doi.org/10.1007/0-387-22440-8_3

[99] B. Scholkopf, A. Smola, and K.-R. Müller, "Kernel principal component analysis," in *International Conference on Artificial Neural Networks*, 1997. [Online]. Available: https://api.semanticscholar.org/CorpusID:7831590

[100] T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," *Ann. Stat.*, vol. 36, no. 3, pp. 1171–1220, Jun. 2008.

[101] M. Nijs, T. Smets, E. Waelkens, and B. De Moor, "A mathematical comparison of non-negative matrix factorization related methods with practical implications for the analysis of mass spectrometry imaging data," *Rapid Communications in Mass Spectrometry*, vol. 35, no. 21, p. e9181, 2021.

[102] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0377042787901257

[103] T. Gerrits, "Transition edge sensor signals (boulder)," https://doi.org/10.5281/zenodo.14101974, 2024.

[104] N. Dalbec-Constant, "Transition edge sensor signals (ottawa)," https://doi.org/10.5281/zenodo.14042152, 2024. [Online]. Available: https://doi.org/10.5281/zenodo.14042152

[105] V. D. Vaidya, B. Morrison, L. G. Helt, R. Shahrokshahi, D. H. Mahler, M. J. Collins, K. Tan, J. Lavoie, A. Repingon, M. Menotti, N. Quesada, R. C. Pooser, A. E. Lita, T. Gerrits, S. W. Nam, and Z. Vernon, "Broadband quadrature-squeezed vacuum and nonclassical photon number correlations from a nanophotonic device," *Science Advances*, vol. 6, no. 39, p. eaba9186, Sep. 2020.

[106] N. Dalbec-Constant, "Photon-number-classification," https://github.com/polyquantique/Photon-Number-Classification, 2024.

[107] (2024) UMAP: Uniform manifold approximation and projection for dimension reduction — umap 0.5 documentation. [Online]. Available: https://umap-learn.readthedocs.io/en/latest/index.html

[108] C. Malzer and M. Baum, "A hybrid approach to hierarchical density-based cluster selection," in *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. IEEE, Sep. 2020. [Online]. Available: http://dx.doi.org/10.1109/MFI49285.2020.9235263

[109] (2024) scikit-learn: machine learning in python — scikit-learn 1.5.1 documentation. [Online]. Available: https://scikit-learn.org/stable/index.html

[110] S. S. Lingala, S. Bedekar, P. Tyagi, P. Saha, and P. Shahane, "FPGA based implementation of neural network," in *2022 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*, 2022, pp. 1–5. [Online]. Available: https://ieeexplore.ieee.org/document/9752656

[111] J. Lee, L. Shen, A. Cere, and C. Kurtsiefer, "Multi-pulse fitting of transition edge sensor signals from a near-infrared continuous-wave source," in *2019 Conference on Lasers and Electro-Optics Europe & European Quantum Electronics Conference (CLEO/Europe-EQEC)*, 2019, pp. 1–1.

[112] I. A. Burenkov, A. Sharma, T. Gerrits, G. Harder, T. Bartley, C. Silberhorn, E. Goldschmidt, and S. Polyakov, "Full statistical mode reconstruction of a light field via a photon-number-resolved measurement," *Phys. Rev. A*, vol. 95, no. 5, p. 053806, 2017.

APPENDIX A NEURAL NETWORK

For both parametric implementations of t-SNE and UMAP, we use a simple feed-forward neural network defined as a series of blocks containing linear layers with ReLU activation functions followed by a batch normalization step. The results presented in this work use a neural network with 4 blocks, where each linear layer contains 300 inputs and outputs. While more complex architectures could be used for this task, we find that even an elementary neural network can achieve this task accurately, resulting in fast data transformation.

We note that we use the same neural network to predict data for both close-to-uniform and close-to-geometric cases. This is done to train the neural network on a balanced dataset, and in this process we guarantee that the neural network is never trained on test data. Using different distributions for the training step is not advantageous to parametric methods, since the close-to-uniform dataset contains fewer samples than in the close-to-geometric case. Additionally, other methods do not benefit from being trained using this data.

For more details about the implementation, the source code for parametric algorithms is available on the public repository provided in Ref. [106].