UNIVERSITY OF CALIFORNIA,
IRVINE

Applications of Information Dynamics to the Study of Nanopores

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Physics

by

Claire Gilpin

Dissertation Committee:
Professor Craig Martens, Chair
Professor Zuzanna Siwy
Associate Professor Michael Yassa

2018

DEDICATION

To My Family

TABLE OF CONTENTS

Page

LIST OF FIGURES **v**

LIST OF TABLES **vii**

ACKNOWLEDGMENTS **viii**

CURRICULUM VITAE **ix**

ABSTRACT OF THE DISSERTATION **xii**

1 Introduction **1**
- 1.1 Dynamical Systems . . . 1
- 1.2 Background on Dynamical Systems . . . 2
  - 1.2.1 Types of Data . . . 2
  - 1.2.2 Categorization of Methods for Analyzing Dynamical Systems . . . 2
- 1.3 Information Dynamics . . . 6
  - 1.3.1 Early Developments . . . 6
  - 1.3.2 Modern Developments . . . 7
- 1.4 Nanopores . . . 8
  - 1.4.1 Background on Nanopores . . . 8
  - 1.4.2 Experimental Nanopore . . . 10
  - 1.4.3 Nanopore Model . . . 11
  - 1.4.4 Solving the Nanopore Equations . . . 13
- 1.5 Time series notation . . . 14
- 1.6 Probability Densities . . . 17
- 1.7 Overview of the Thesis . . . 24

2 Shannon Entropy and Entropy Rates **25**
- 2.1 Differential Entropy . . . 25
  - 2.1.1 Definition of Differential Entropy . . . 25
  - 2.1.2 Conditional Differential Entropy . . . 27
  - 2.1.3 Definition of Entropy Rate . . . 29
  - 2.1.4 Local Entropy Rate . . . 30
  - 2.1.5 Specific Entropy Rate . . . 31
  - 2.1.6 Summary of Entropy Rate Definitions . . . 32

2.2 Model Order Selection and Probability Density Estimation 33
2.2.1 Kernel Density Estimation 33
2.2.2 Kernel Nearest Neighbor Estimation 35
2.2.3 Model Order Selection in This Work 37
2.2.4 k^{th} Nearest Neighbor Estimation 39
2.2.5 Computing Entropy Rate Estimators 40

3 Local and Specific Entropy Rates in Nanopore Simulation Study 42

4 Q-Step Specific Entropy Rate In a Nanopore 50
4.1 What is q-step Specific Entropy Rate 50
4.2 Application of q-step Specific Entropy Rate to Nanopore Simulation Data . . 54
4.3 Application of q-step Specific Entropy Rate to Experimental Nanopore Data 59
4.3.1 Considerations in the Application of normalized q-step specific Entropy Rate 60
4.3.2 Preliminary normalized q-step specific Entropy Rate Results for Experimental Nanopore Data 60

5 Detection of Nonlinear Structure in Nanopore Interevent Intervals 66
5.1 Design and Aim of the Study 66
5.1.1 Methods Overview 67
5.2 Application to Simulation Nanopore Data 69
5.2.1 Computing Interevent Intervals 69
5.2.2 Results and Discussion 70
5.3 Application to Experimental Nanopore Data 71
5.3.1 Preparation of the Data 71
5.3.2 Computing Interevent Intervals 74
5.3.3 Results and Discussion 77

6 Future Directions and Conclusions 82
6.1 Future Directions for Normalized q-step Specific Entropy Rate 82
6.2 Future Directions for Interevent Interval Analysis 83
6.3 Computational Suggestions for Future Work 83
6.4 Conclusions 84

Bibliography 85

A Additional Normalized q-step Specific Entropy Rate Results 92

B Additional Surrogate Analysis 101

LIST OF FIGURES

Page

1.1 Nanopore Model Potential 12
1.2 Nanopore Simulation Parameter Values 13
1.3 Nanopore Model Transition Trace 14
1.4 Ensemble Space 15
1.5 Ensemble Space with a Binary Alphabet 16
1.6 Ensemble Space for More Complicated Alphabet 17
1.7 Probability Density Functions for Gaussian Distributions with Different Standard Deviations 19
1.8 Probability Density Functions for Gaussian Distributions with Different Standard Means 20
1.9 Bivariate Gaussian Joint and Conditional Probability Densities 22
1.10 Marginal Density $f(x)$ for Bivariate Gaussian 23
1.11 Marginal Density $f(y)$ for Bivariate Gaussian 23

2.1 I-diagram 28
2.2 Realization of a Continuous Valued Time Series 31
2.3 Realization of a Continuous Valued Time Series 32
2.4 Example KDE 35
2.5 Example Behavior Space 36

3.1 Simulation Nanopore Data with Local and Specific Entropy Rates 43
3.2 Reconstucted State Space Plots for Simulation Nanopore Data 46
3.3 Cross-section of Nanopore Reconstructed State Space 48
3.4 Nanopore Transition Trajectory Direction 49

4.0 Example Behavior Space for Constructing Joint Density Estimator 53
4.1 Example Behavior Space for Constructing Marginal Density Estimator . . . 54
4.2 Normalized q-step specific Entropy Rate in Simulation Data 59
4.3 Reconstructed State Space for Experimental Nanopore Data 61
4.4 Normalized q-step specific Entropy Rate in Experimental Data 65

5.1 Transitions in Nanopore Simulation Data 70
5.2 Estimated Total Entropy Rate Histogram for Nanopore Simulation Data . . 71
5.3 Power Spectral Density 72
5.4 Data Smoothing 73

5.5 Choosing a Threshold 75
5.6 Choosing a Threshold 76
5.7 Estimated Total Entropy Rate Histograms for Experimental Nanopore Interevent Intervals 80

LIST OF TABLES

Page

1.1 Parameters Chosen for an Example Bivariate Gaussian Distribution. 21

5.1 Surrogate Generation Steps 68
5.2 Results of Hypothesis Testing for Experimental Nanopore Interevent Intervals 77

ACKNOWLEDGMENTS

I would like to say a sincere and heartfelt thank you to the many people who have helped me grow and learn as a person and as a scientist. The first thank you goes out to my adviser, Craig Martens, for agreeing to work with me as a transfer student. I'm very grateful to you for always treating me like a colleague and valuing my contributions. I'm also grateful for the guidance you have given me. Thank you to Zuzanna Siwy for your collaboration on this research and also for your guidance throughout my time in the Physics department at UC Irvine. Thank you very much to Michael Yassa for being an encouraging committee member. I would also like to acknowledge and thank Jun Allard and Ilya Krivorotov for taking time out of their schedules to sit on my advancement committee. Your input was valuable to me. A big thank you also to David Darmon. You have been an invaluable mentor and your guidance has helped me break into a very interesting and inspiring area of research.

I would also like to thank my previous mentors and professors for believing in me and encouraging me. The study of physics is intellectually fruitful, but challenging, and I have greatly appreciated your support.

Thank you to the journal Entropy for permitting inclusion of our previously published results.

Thank you to UC Irvine as a whole for becoming my home. I have made great friends here and have enjoyed my time in the physics program. I am excited to broaden my horizons through continued future study here. I am also very grateful for people who have become like family at NBRL. You have also helped shape my career goals and an thankful for that.

Thank you to my wonderful friends. I am a very lucky person to have your support and it means the world to me. Thank you for always lending your ears and for walking through life with me.

Lastly, thank you to my parents and family for the opportunities you have given me throughout my life. I could never thank you enough for your love and support. You are amazing! Also, thanks to my favorite four-legged creature, Helena for the love and comic relief.

CURRICULUM VITAE

Claire Gilpin

EDUCATION

PhD: Physics - Chemical and Materials Physics (ChaMP)
Thesis title: Applications of Information Dynamics to the Study of Nanopores
Adviser: Craig Martens, PhD
University of California, Irvine
2018

MS: Physics - Chemical and Materials Physics (ChaMP)
Thesis title: Fabrication and Electronic Studies of PbSe Nanoparticle Superlattices
Adviser: Matthew Law, PhD
University of California, Irvine
2016

BA: Astrophysics
Franklin & Marshall College
2012

RESEARCH EXPERIENCE

Information Dynamics of Nanopores
Adviser: Craig Martens, PhD
University of California, Irvine
2016-2018

PbSe Nanoparticle Superlattices
Adviser: Matthew Law, PhD
University of California, Irvine
2013-2016

Surface Physics of CdSe Nanoparticles
Adviser: Robert Meulenberg, PhD
University of Maine, Orono
2012-2013

Dye-Sensitized and Depleted Heterojunction Colloidal Quantum Dot Solar Cells
Adviser: J. Kenneth Krebs, PhD
Franklin & Marshall College
2011-2012

ASTRON Summer Student Programme
Advisers: Jason Hessels, PhD, Joeri van Leeuwen, PhD, Vlad Kondrateiv, PhD
Dwingeloo, The Netherlands
2011

Pulsar Survey and Timing Studies
Advisers: Fronefield Crawford, PhD, Andrea Lommen, PhD
Franklin & Marshall College
2008-2011

TEACHING EXPERIENCE

Graduate Teaching Assistant (Physics)
Taught laboratory and discussion for introductory physics classes
University of California, Irvine
2013-2015

Graduate Teaching Assistant (Physics)
Taught laboratory and discussion for introductory physics classes
University of Maine, Orono
2012-2013

Undergraduate Teaching Assistant (Physics and Astronomy)
Assisted professors in teaching introductory physics and astronomy laboratory
Franklin & Marshall College
2009-2012

REFEREED JOURNAL PUBLICATIONS

Claire Gilpin, David Darmon, Zuzanna Siwy, and Craig Martens, (2018) Information Dynamics of a Nonlinear Stochastic Nanopore System, *Entropy*, 20, 221.

Deborah Schmidt, Fronefield Crawford, Glen Langston, and Claire Gilpin, (2013) A Search For Rapidly Spinning Pulsars and Fast Transients in Unidentified Radio Sources with the NRAO 43-Meter Telescope, *Astronomical Journal*, 145, 4.

SELECTED HONORS AND AWARDS

NSF Graduate Student Research Fellowship - Honorable Mention
University of California, Irvine
2014

Cum Laude
Franklin & Marshall College
2012

Departmental Honors
Franklin & Marshall College
Thesis: Synthesis and Electrical Characterization of Alternative Photovoltaics Including Dye-Sensitized Solar Cells and Depleted Heterojunction Colloidal Quantum Dot Solar Cells
2012

Sigma Pi Sigma (National Physics Honors Society)
Franklin & Marshall College
2011

John Kershner Scholar - Physics
Franklin & Marshall College
2011 & 2012

ABSTRACT OF THE DISSERTATION

Applications of Information Dynamics to the Study of Nanopores

By

Claire Gilpin

Doctor of Philosophy in Physics

University of California, Irvine, 2018

Professor Craig Martens, Chair

Over the previous three decades both experimental and theoretical research into nanopores has been gaining momentum. It has been discovered that nanopores play an important role in controlling important molecular and cellular scale physiological processes. It has also been discovered that both synthetic and biotic nanopores may have groundbreaking potential for both biomedical devices and scientific research instruments. In particular, nanopores are currently being studied for their potentially cost effective application to DNA sequencing and protein, drug, and pathogen sensing. Additionally, research into the time-dependent electrical properties of nanopores may aid in our understanding and ability to model the behavior of physiological nanoscale membrane ion channels. Recent advances in information theory, particularly the development of time-dependent measures of Shannon entropies, have opened the door to studying these nanoscale systems from a new angle. In this work we will share results of the novel application of these techniques, highlighting their ability to track autonomous fluctuations in nanopore currents. We will also discuss a proposed extension of these techniques that may allow short-time scale prediction of current fluctuations in the future. Additionally, we will discuss a process for testing for potentially interesting nonlinear structure in nanopore interevent interval sequences, where the events are current fluctuations. Lastly, we will discuss some potential future research directions in light of what we have learned.

Chapter 1

Introduction

1.1 Dynamical Systems

Understanding dynamical systems is fundamental to understanding our world. A dynamical system is any system whose behavior changes as a function of time. Some prominent examples of dynamical systems most of us think about regularly are systems like the stock market and weather patterns. For as many dynamical systems as we can observe with the naked eye, there are many that operate beneath our visual limits. Among the more obscure dynamical systems is the time-dependent electrical behavior of nanopores; nanometer-scale channels that are currently being investigated for their potential biomedical and scientific applications.

In this introduction we will cover the historical highlights of scientific research on dynamical systems including commonly used analysis techniques. We will then explore the advent of information theory and how recently developed time-dependent measures of entropy rate make it possible to explore dynamical systems in a new light.

1.2 Background on Dynamical Systems

1.2.1 Types of Data

To understand dynamical systems research we need to understand the types of data we might be analyzing. Time series data may take the form of symbol sequences (an example of which would be the left and right bar presses in a rat behavior experiment), interevent interval sequences (such as inter-beat intervals of the heart, neural spike trains, and as we will see, sequences created by successive current fluctuations in a nanopore), or continuous valued waveforms (such as seismograph data) [39, 86]. We acknowledge that 'continuous valued' is continuous valued within the limits of digital measurement technology.

It is possible to convert between some of these types of data. For example, interevent interval sequences and continuous valued data can be converted to symbol sequences by setting threshold values and assigning values in certain ranges to a particular symbol. Continuous valued waveforms may also be converted to interevent interval sequences by detecting transitions [9, 10]. Transition detection is still an active area of research and there is no general consensus on an ideal method [1].

1.2.2 Categorization of Methods for Analyzing Dynamical Systems

There are multiple categories of analysis techniques that can be applied to dynamical systems research. Each has its own scope of applicability. We will discuss the predominant categories briefly, acknowledging which techniques are the focus of this research.

Statistical Time Series Analysis

Statistical analysis techniques have broad applications to dynamical systems. Statistical characterization can be useful for characterizing the distribution of data from dynamical systems, determining the quality of the data, and can also be useful for model development applications. For example, it can be useful for determining if something has gone awry during a measurement if, for example, a much broader distribution that is expected for a particular variable is seen. In modeling applications, statistics such as regression analysis are often used to inform the development of and refine models for a particular process/variable [90, 19].

From the point of view of time series analysis statistical analysis provides the ability to model complicated sets of data either through using well characterized parametric models or through non-parametric modeling when a well defined set of parameters isn't known. Parametric statistical analyses, such as linear regression and autoregressive moving average have the advantage of being computationally efficient, but require a priori assumptions that a well characterized model class is a good fit for the data. While more computationally burdensome, non-parametric techniques do not require such a priori assumptions. Non-parametric modeling is used in this work for the purposes of model order selection (see section 2.2) [7].

Spectral Analysis

Spectral analysis can be a useful tool in time series analysis, particularly from the point of view of model validation or from the point of view of matching contributions from a signal to a well established model [53]. For example, in the process of modeling a dynamical system the power spectral density may be estimated for experimental data and simulated data from the model [4]. Comparing these power spectra, which show the signal amplitude in the frequency domain can be part of model validation or the identification of errors in modeling.

One drawback of spectral analysis is that its inherent assumption is that the dynamical system in question can be modeled by a group of oscillators. This is often a good approximation for many physical dynamical systems, but it may overlook important behavior in others. Explicit procedures exist in the literature for comparing spectra [4].

Additionally, spectral analysis can be useful in analyzing noisy experimental data by using spectra-based filtering methods, such as band pass filters [61, 79]. In this work we use spectral analysis for this purpose.

Time-Frequency Analysis

Time-frequency analysis is a supplementary technology to spectral analysis. In spectral analysis we can learn about and characterize a signal in the frequency domain, but it does not tell us which frequencies are active in the signal at specific times. This is possible by using time-frequency analysis technologies such as wavelet, Gabor, and S transforms [27, 84, 20]. These transformations aim to provide a time-dependent frequency analysis. These techniques are extremely useful when analyzing signals where it is known that certain time-dependent frequencies have important behavioral implications, such as analysis of electrocardiogram data where the timing of different frequency components of the signal is an indicator of cardiac function [52, 68].

One notable drawback to time-frequency analysis is that there is a tradeoff between temporal and frequency precision. This tradeoff is akin to a Heisenberg uncertainty principle in that the more temporal precision you achieve the less frequency precision you will achieve (the reverse is true as well) [26].

Dynamical Systems Theory

The development of dynamical systems theory has spanned more than 100 years in time, and an early lead in to the subject might be thought of as Henri Poincare's attempt to construct analytic solutions to the three body problem [64]. His analysis showed that differential equations could have solutions that are represented by geometric objects (what later became known as attractors) and that deterministic systems could present what are now known to be chaotic solutions. A later parallel occurred as Dutch scientist Balthasar van der Pol observed an odd mode-locking behavior in vacuum tube-containing electrical circuits at certain frequencies [91]. This became a possible problem for electrical engineers who were concerned about the impact of aperiodic oscillations in their circuits.

Mary Cartwright and John Littlewood were recruited to address these concerns in the context of their application to radar systems and their work became one of the foundations for the development of chaos theory [8, 55]. Another critical foundational contribution to the development of chaos theory occurred when Edward Lorenz was modeling weather patterns [51]. In his simulation studies he opted to begin a simulation using conditions at the midpoint of his previous simulation as his new initial conditions. He was surprised by the markedly different results he obtained and he explored other initial conditions. This lead to the discovery that slight changes in initial conditions result in different behavior in the solutions to some nonlinear differential equations.

Chaos theory has developed into its own discipline over the years and there is active research into its applicability to physical and biological systems. Common measures emerging from the chaos theory community are correlation dimension, Lyapunov exponents, and Hurst exponents. Correlation dimension is a measure of the dimension of the space occupied by a set of points (where points on a line have correlation dimension of 1, points on a plane have correlation dimension of 2, etc.), where non-integer values are possible [25, 24]. Lyapunov

exponents provide a quantification for the divergence of trajectories for infinitesimally close initial conditions [2, 21]. Hurst exponents are measures of persistence in time series data, where persistence means tendency to follow the current trajectory (i.e. to increase following past increases and decrease following past decreases) [34].

Correlation dimension, Lyapunov exponents, and Hurst exponents have some practical drawbacks to the analysis of experimental data. In particular nonstationary data complicate estimation of the correlations dimension and it is necessary to have long, stationary data sets [88, 87, 22]. Lyapunov and Hurst exponents for experimental data analysis have the added drawback of sensitivity to noise and identification of false values. There is not yet a clear consensus for how to circumvent these issues for a clean analysis of non-stationary experimental data.

1.3 Information Dynamics

1.3.1 Early Developments

Following preceding work by Harry Nyquist in 1924 [59] and Ralph Hartley in 1928 [28], in 1948, while working at Bell Labs, Claude Shannon published a groundbreaking paper [76] entitled "The Mathematical Theory of Communication," in which he proposed the following definition for information entropy, $H[X]$,

$$H[X] = -\sum_{x \in \mathcal{X}} p(x) \log p(x). \tag{1.1}$$

This definition, based on probability theory, quantifies the uncertainty in a message X, where H is the entropy, x is the realized form of the message, $\mathcal{X}$ represents the possible "alphabet" from which successive entries in X were realized, and $p(x)$ is the probability of any particular

realized element or subset of the message. The definition above concerns discrete variables, however his work also included a generalization to continuous variables, differential entropy

$$H[X] = -E[\log f(x)] = -\int_{-\infty}^{\infty} f(x) \log f(x) dx \tag{1.2}$$

where $f(x)$ is a probability density associated with a continuous variable X [76, 12]. As an aside, we will be working with continuous variables in this work and as such will address the mathematical details of differential entropy in more detail in the next chapter. Shannon's paper is commonly regarded as the start to the formal study of information theory. Since its creation this basic equation has been adapted and built upon in an attempt to describe how information is stored, processed, and transmitted in physical systems. A notable contribution was made in the 1950s by Andrey Kolmogorov and Yakov Sinai that was directly applicable to chaotic systems. Generally, Kolmogorov-Sinai entropy is the rate at which information about the initial conditions as more data are observed [42, 80]. One important point about the Kolmogorov-Sinai entropy is that it is only finite and non-zero for chaotic systems and goes to infinity for stochastic systems [60].

1.3.2 Modern Developments

Modern developments in information dynamics have seen the creation of new extensions of entropy to time dependent measures usable on stochastic dynamical systems. These notable developments include Shannon entropy variants local entropy rate, which quantifies the time-dependent surprise associated with a particular known future state given knowledge of the past states, and specific entropy rate, which quantifies the time-dependent uncertainty in an unknown future state given knowledge of the past states [49, 48, 47, 14, 17].

These measures are the primary focus of this dissertation research and are applicable to the system of interest, an autonomously fluctuating conical nanopore. We explore them in

greater mathematical detail in chapter 2.

1.4 Nanopores

1.4.1 Background on Nanopores

What is a Nanopore

Nanopores are nanometer-scale channels that allow the passage of certain substances. They may be present in biological systems, such as membrane pores created by pore forming proteins or synthetic, such as pores created in a polymer film via track etching. They are unique and have distinct physical behavior not seen in larger pores because the physical impact of the properties of the pore walls plays a significant role at the nanometer size scale [37, 83, 65].

Brief History of Nanopore Research

Nanopores have become a topic of interest in recent decades as researchers began studying their potential advantages for prominent, real world applications such as genome sequencing and biosensing, and in developing a mechanistic understanding of cellular and molecular scale physiological processes [63, 40, 43, 32, 36, 71, 67, 6, 5, 31, 29, 56, 96, 93, 72, 89, 62]. We will briefly discuss each of these applications so as to construct an overview of important advances in nanopore research.

Since the advent of the Human Genome Project in the late 1900s, there has been a push in the scientific community to develop faster and more cost efficient methods for DNA sequencing [70]. We have determined that our the sequence of our base pairs, the nitrogenous

subunit of DNA, is critical to understanding the human body, and in particular, states of disease [11]. Understanding these sequences is also the first step to developing gene therapies that may hold the key to combatting many disease processes. One of the primary barriers to genome research was the cost of DNA sequencing. Just after the turn of the century the original endeavor of the Human Genome Project was announced as complete, but it was not yet cost efficient enough for DNA sequencing to be used widely for research. Finding more cost effective, high accuracy DNA sequencing techniques became one of the subsequent goals in the scientific community [18].

Nanopores were identified as having promise for cost efficient DNA sequencing. It was discovered that different bases traveling through a nanopore result in different electrical currents [58, 92, 31, 93, 82, 33]. This information was used to create nanopore-based DNA sequences that are currently on the market, such as those available from Oxford Nanopore Technologies. The high throughput sequencing technology uses an enzyme, which attaches to the DNA strand and docks on the nanopore, to ratchet DNA strands through the nanopore. The current is measured and the values are paired with the corresponding nitrogenous base.

Additionally, recent advances in nanopore research have lead to the ability to detect single molecules, also by detecting changes in the nanopore current [54, 83, 37, 65, 30]. This paves the way for applications of nanopore in both scientific research and medical diagnostics with nanopores potentially providing portable rapid analysis of samples ranging from biological fluids to environmental samples. Detection could be customized to include targets from proteins, to drugs, to single viruses.

Biological nanopores are also found in the membranes of our cells and are responsible for controlling fluxes of ions into and out of our cells [83, 65]. Studying both biotic and synthetic nanopores in laboratory environments, particularly their electronic behavior can be used to provide insight into the physiological processes they model. In particular it has been noted that nanopores in contact with ion solutions can exhibit bias dependent autonomous

current fluctuations [83, 65]. In other words, application of a sufficient external bias voltage to a nanopore immersed in an ion solution can result in autonomous current fluctuations. The theorized mechanism of action behind these autonomous fluctuations is the alternating precipitation/dissolution of components of the ionic solution due to their interaction with charges on the nanopore walls [65, 37]. Understanding theses current fluctuations is an area of active research within the UC Irvine community.

In this project we conduct new explorations of these current fluctuations both for a model system and experimental data using time-dependent information dynamics tools. In particular we apply new time dependent measures of entropy rate to identify and study features of interest in the current time series for single conical nanopores. We additionally begin efforts to identify evidence of nonlinear structure in the sequences of interevent intervals (where the events of interest are the current fluctuations). We begin by describing both the experimental and model systems researched in our work below.

1.4.2 Experimental Nanopore

In this work we will be considering a system comprised of a single nanopore current oscillator in contact with an ion solution [65, 37]. Experimentally, these conical nanopores are fabricated by irradiating polyethylene terephthalate (PET) with heavy ions to form tracks and subsequently subjecting those tracks to controlled chemical etching. The resulting pores are between 2 and 6 nm in diameter [65].

The pores are immersed in a solution containing 0.1 M KCl and 0.3 mM $CaCl_2$. With sufficient applied external bias voltage, these pores exhibit autonomous current fluctuations between positive and negative conductance states. There is evidence that these fluctuations are the result of $CaCl_2$ nanoprecipitate formation/dissolution on the narrow walls of the nanopore (due to the dynamic interactions of the Ca^{2+} and Cl^- ions with the surface charges

on the nanopore walls) [65, 37]. A model for nanopore current fluctuations under similar experimental conditions was proposed in 2015 and we will discuss it in the next section below [35].

1.4.3 Nanopore Model

The following is an excerpt from [23]: "A single nanopore current oscillator, which is observed in a single state x as a function of time t, can be modeled as a coupled nonlinear oscillator with bistable potential $U(x, y)$ forced by dynamical noise:

$$\begin{aligned} dX_t &= -\frac{1}{\gamma_x}\frac{\partial U}{\partial x}(X_t, Y_t)\, dt + \sigma_x\, dW_x \\ dY_t &= \frac{1}{\gamma_y}[k_+\Theta(X_t) - k_-\Theta(-X_t)]dt + \sigma_y dW_y \end{aligned} \tag{1.3}$$

where X_t is the nanopore current, Y_t is an unobserved state variable that controls the opening and closing behavior of the nanopore, γ is a coefficient of friction, Θ is a Heaviside function with amplitude determined by rate constants k_+ and k_-, and W_x and W_y are standard Brownian motions representing other unaccounted for inputs to the system. The double-well potential $U(x, y)$ is taken to be

$$U(x, y) = \frac{1}{4}ax^4 - \frac{1}{2}b(V)x^2 + cxy \tag{1.4}$$

where $b(V)$ is the voltage-dependent parameter that determines the barrier height:

$$b(V) = b_0\Big(\frac{V - V_c}{V_c}\Big) \tag{1.5}$$

with V_c as the critical voltage. The behavior of the potential as it relates to both X and Y is critical to understanding the behavior of the nanopore current and should be examined

in more detail. When X takes positive values, Y will be a random walk with drift $\frac{k_+}{\gamma_y}$ plus dynamical noise. As Y drifts to more positive values, the potential tips towards the negative well, making it likely that a transition will occur from a positive to a negative current. This effect can be seen in Figure 1.1A below. Conversely, when the nanopore current X takes negative values, Y will be a random walk with drift $\frac{-k_-}{\gamma_y}$ plus dynamical noise. Eventually, Y drifts negative to a point of tipping the potential towards the positive well, making it likely that a transition will occur from a negative current to a positive current. This effect can be seen in Figure 1.1B below. Between fluctuations, Y will pass through $Y = 0$.

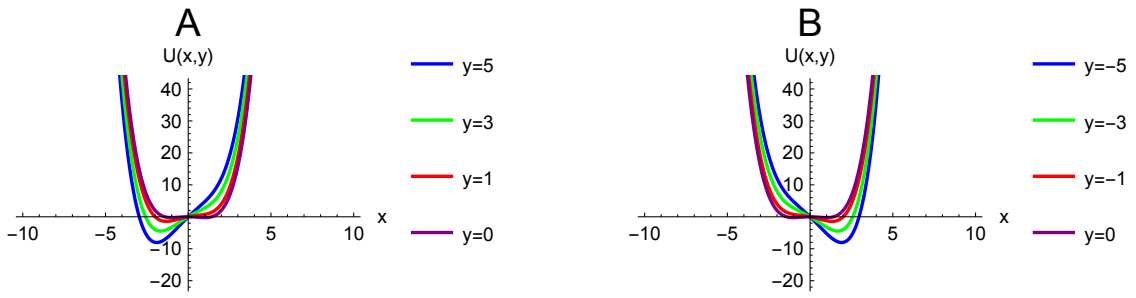


Figure 1.1: Potential at fixed positive values of y (**A**) and fixed negative values of y (**B**). The $y = 0$ configuration is shown on both plots. As y becomes more positive, it causes the potential to skew towards a transition to negative x. As y becomes more negative, it causes the potential to skew towards a transition to positive x. Physically, positive values of x in this graph correspond to positive current values.

The structure of the potential term as it relates to X and Y acts to ensure that the system undergoes transitions frequently and never becomes stuck indefinitely in one of the wells. Similar behavior of the nanopore current should therefore be expected across realizations of data simulated using this model, despite expected differences in the profiles of individual transitions due to the dynamical noise."

We should briefly note that the values shown in the graphs of the potential above are exaggerated to show the change in shape of the potential. They are not intended to show the range of Y values seen in our simulations, which may be smaller.

1.4.4 Solving the Nanopore Equations

We used a stochastic Runge-Kutta method to compute realizations of 1.3 [69]. The method was employed through the SRI2 integrator from the sdeint package for python. We used the following parameters in our simulations for this work, consistent with those used in [35].

Variable	Value
a	1
b	1
c	1
k_+	1
k_-	5
γ_x	1
γ_y	100
σ_x	0.1
σ_y	0.1
Time step	0.25
Down sampled time step	0.5

Figure 1.2: Parameter values for simulation of nanopore data. The external bias voltage is taken to be constant.

We can look directly at a trace of both X and Y with some displayed transitions in the example below. We see that as Y drifts to more positive values it eventually hits a value extreme enough to trigger a shift in the potential associated with a probable transition to the negative conductance state. As Y then drifts to more negative values it eventually hits a value extreme enough to trigger a shift in the potential associated with a probable transition to the positive conductance state.

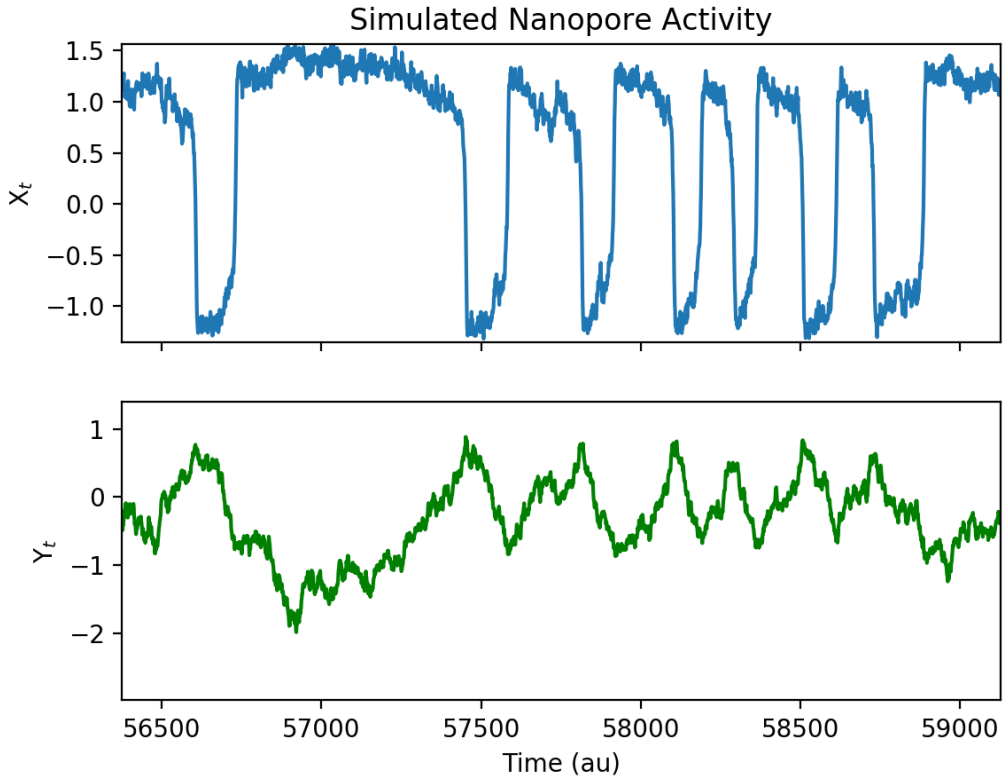


Figure 1.3: X and the unobserved variable Y for simulated transitions in a conical nanopore. We can observe that as the value of Y increases, eventually it reaches a tipping point where a transition then occurs. This corresponds to a change in shape to the potential similar to those seen in 1.1A. As the value of Y decreases, we can also see that it reaches a value low enough to trigger a transition back to the positive state. This corresponds to a change in shape to the potential similar to those seen in 1.1B.

We computed a total of five realizations for analysis in this work.

1.5 Time series notation

We will be using standard statistical notation throughout this work. Any exceptions will be explicitly identified in text.

$\{X_t\}$ will represent a stochastic process modeling some observable/random variable, X. We

can think of X_t as representing the state of this observable in the immediate future and X_{t-1} as representing the state in the immediate past. It is perhaps natural to think of X_t as representing the value of the observable at the present moment, but here we construct a mental framework which only contains pasts and futures. We can represent a block of states by using a subscript/superscript notation $X_m^n = (X_m, X_{m+1}, ..., X_{n-1}, X_n)$.

Realizations, or observed values, of the stochastic process $X_1, X_2, X_3, ..., X_T$ (where T is the length of the time series) will be denoted using lower case notation $x_1, x_2, x_3, ..., x_T$. Pictorially, we can imagine that we have a "realization space" (or "ensemble space") as shown below. We've called the realizations shown α, β, and γ in order to distinguish them from each other and other realizations.

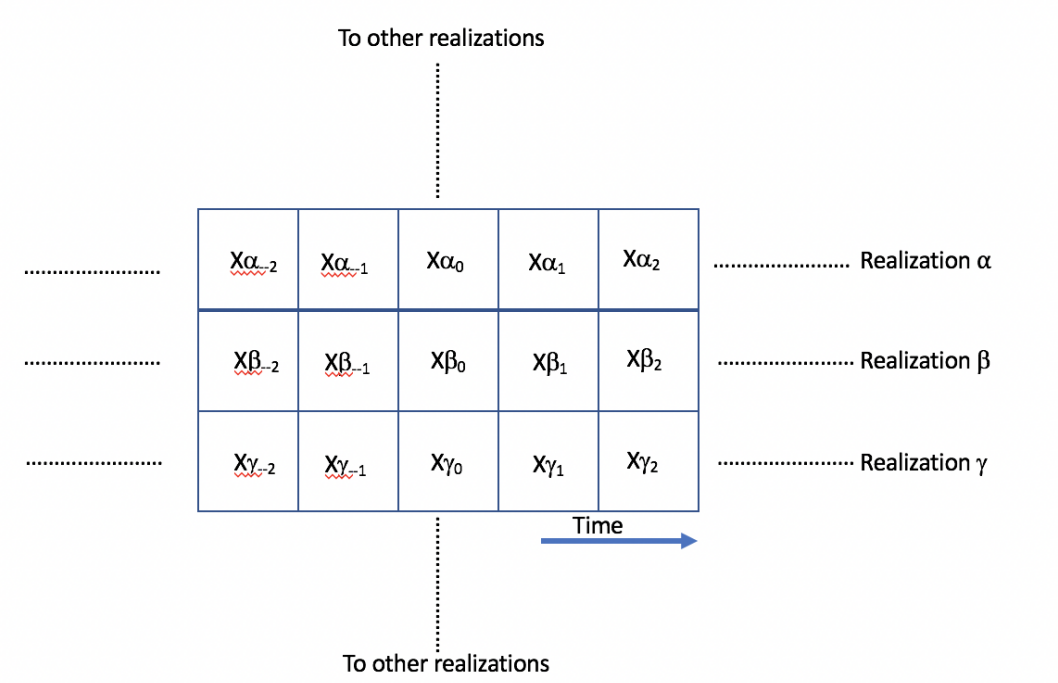


Figure 1.4: Pictorial representation of ensemble (realization) space. A single realization is a string of measurements taken over time. Three distinct realizations are shown in this figure.

We can also explicitly denote the set of accepted values for an observable as $\mathcal{X}$, which represents an alphabet of possible values for the observable. In a binary process, for example,

this alphabet contains only options 0 and 1 (i.e. $\{0, 1\}$).

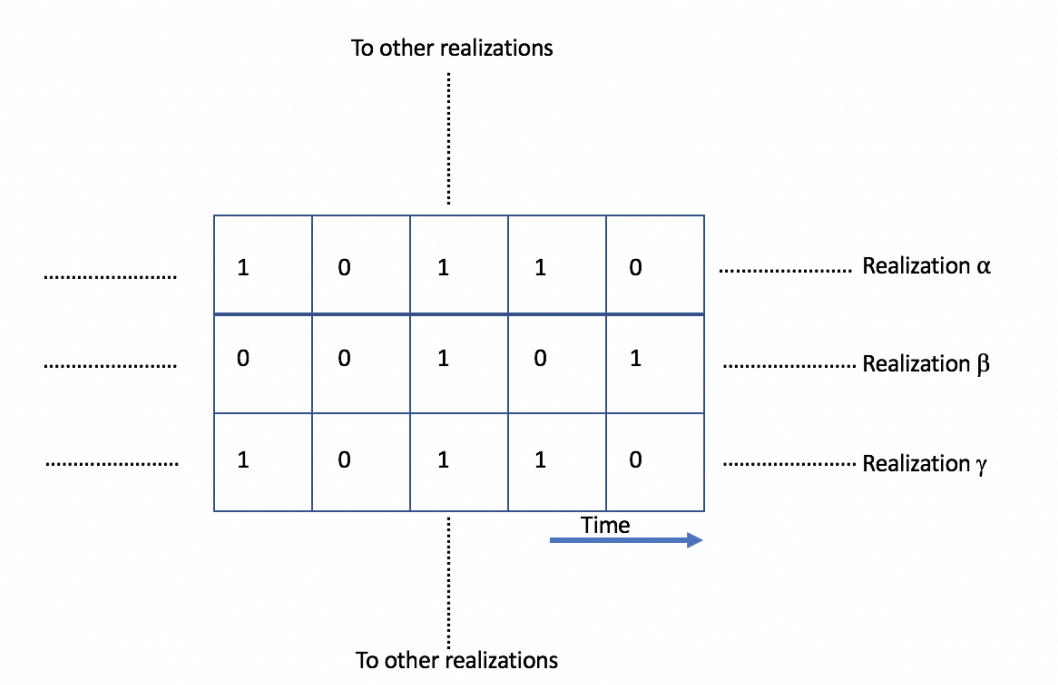


Figure 1.5: Example of three possible realizations for a binary alphabet.

The alphabet can become much larger for continuous-valued observables. Let us imagine that we are observing a random variable, X to three decimal places of precision and its possible alphabet ranges from 0 to 2. A realization in this case might look like the following

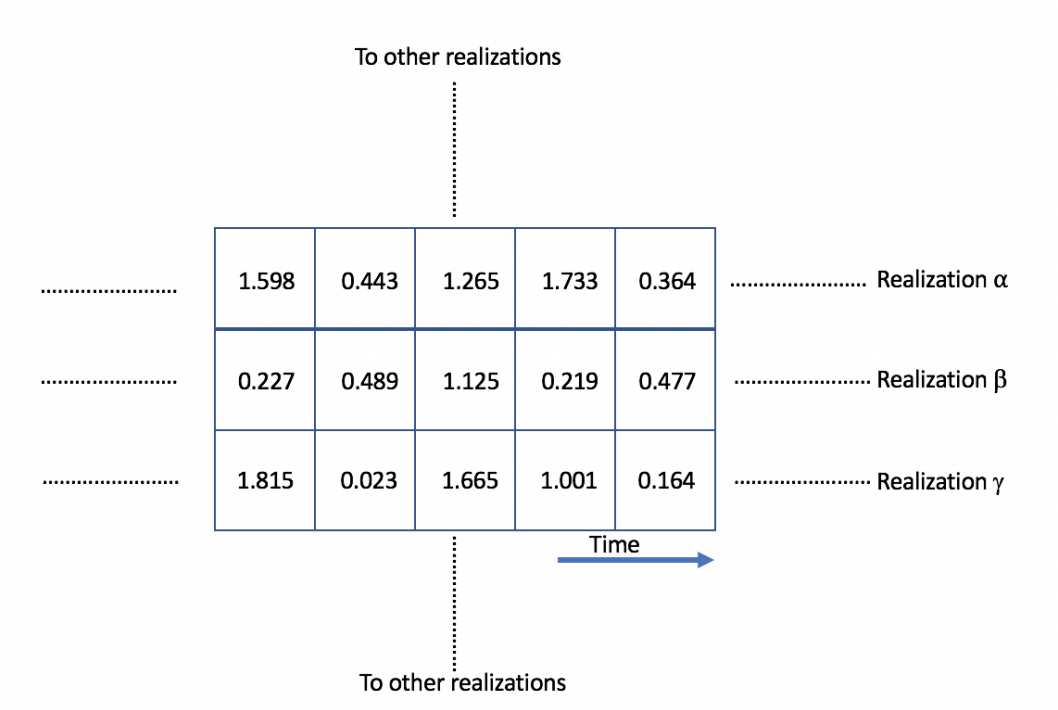


Figure 1.6: Example of three possible realizations for a more complicated alphabet where measurements are collected to three decimal places of precision and can range from 0 to 2.

1.6 Probability Densities

Probability densities will appear extensively in this work. As is standard practice, $p(x)$ or $f(x)$ will denote a probability mass function and density function for a discrete-valued or continuous-valued variable, X, respectively [94, 85, 12]. This density is often referred to in statistics as the marginal density of X. For ease of writing, this general discussion be formulated for continuous-valued variables. The probability density function $f(x)$ must satisfy the following conditions [12]

$$f(x) \geq 0 \tag{1.6}$$

for all x (i.e. that the probability density must be non-negative), and the normalization condition

$$\int_{-\infty}^{\infty} f(x)dx = 1. \tag{1.7}$$

It will also be important to understand the concept of joint and conditional probability densities. Joint probability densities can be thought of as the probability density associated with observing both x and y in a process with a joint density $f_{X,Y}$. That joint probability density is written $f_{X,Y}(x, y)$. Joint probability densities are subject to a normalization condition analogous to the marginal probability density above

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x, y)dxdy = 1. \tag{1.8}$$

Conditional probability densities can be thought of as the probability density associated with observing Y given that X was observed (or X given that Y was observed) in a process with conditional probability density $f_{Y|X}$ (or $f_{X|Y}$). In this case the ordering of the observations is crucial. Conditional probability density is written to reflect the order as [12]

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x, y)}{f_X(x)} \tag{1.9}$$

or

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}. \tag{1.10}$$

We can think, graphically, of conditional probability densities as being the cross-section of joint probability density at a fixed value of one or the other variable, normalized by the marginal probability density of that variable [94].

One example that is commonly introduced to aid in the understanding of marginal, joint, and conditional probability densities is a Gaussian (normal) distribution. For a univariate (single-variable) Gaussian, the marginal probability density associated with a continuous-valued variable X is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1.11}$$

where μ is the distribution's mean, σ is the standard deviation, and σ^2 is the variance. μ can be thought of as the expected value of X and σ determines the degree of spread of the distribution. Three examples of the marginal probability density function for Gaussian distributions with $\mu = 0$ and different values of σ, a small, medium, and large value, are shown in the plot below. This plot serves the purpose of highlighting the general shape of the distribution, and furthermore, highlights how the impact of the variance on that shape. Intuitively, we expect to see a larger spread in the marginal probability density, centered around mean μ as the standard deviation increases, and the plot below is consistent with that intuition.

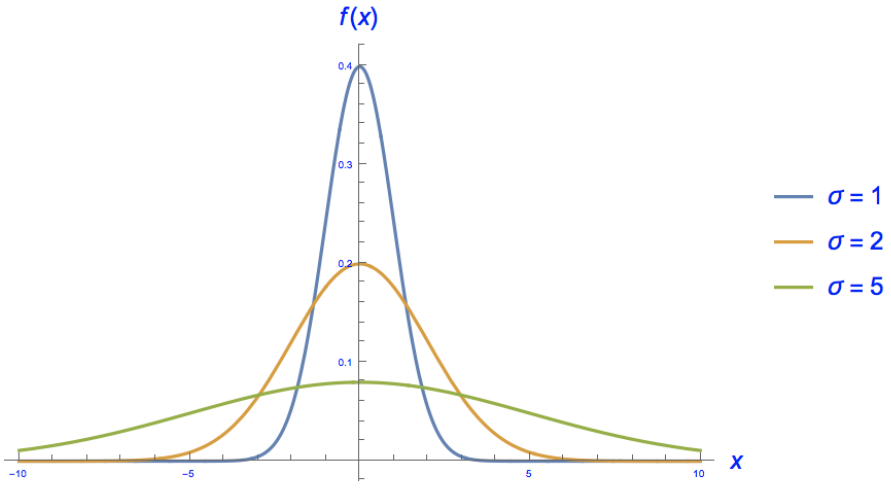


Figure 1.7: Examples of Gaussian (normal) distributions with mean 0 and different values of the standard deviation, 1, 2, and 5. Plotted together, they illustrate how the spread of the marginal probability density changes with σ.

Changing the mean, μ of the distribution results in a shift of the marginal probability density along the horizontal axis, as shown below.

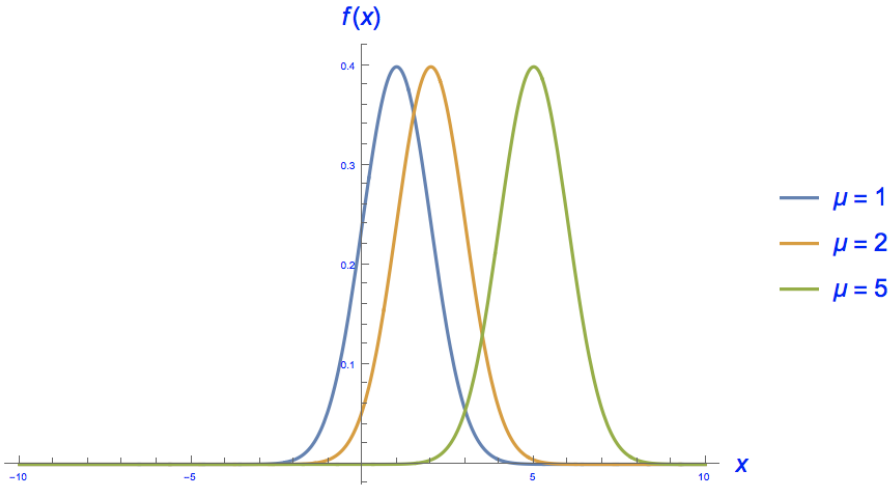


Figure 1.8: Examples of Gaussian (normal) distributions with means 1, 2, and 5 and standard deviation 1. Plotted together, they illustrate how the marginal probability density shifts with changing mean.

For multivariate (multiple variable) Gaussian distributions we can discuss joint and conditional probability densities. Taking the simplest example, a bivariate Gaussian, with variables X and Y, the known joint probability density is expressed below

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{\frac{-z}{2(1-\rho^2)}} \tag{1.12}$$

where z is defined as the following

$$z = \frac{(x-\mu_x)^2}{\sigma_x^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}. \tag{1.13}$$

In the above equation, μ_x and μ_y are the means of the X and Y distributions respectively.

Σ is the covariance matrix [12]

$$\Sigma = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix} \tag{1.14}$$

and ρ is the correlation between X and Y. Choosing the following parameter values,

μ_x	10
μ_y	10
σ_x	4
σ_y	4
ρ	0.5

Table 1.1: Parameters for example bivariate Gaussian distribution. Parameter values were chosen to aid in visualization of the joint probability density,

chosen for ease of visualization of the joint density, we arrive at the following graphical representation of the joint density. Additionally, the conditional density, as we know is obtained by taking a cross-section of the joint probability density at a fixed value of one of the random variables, X or Y. Below, we also show the conditional probability for a fixed value of X, $x = 10$.

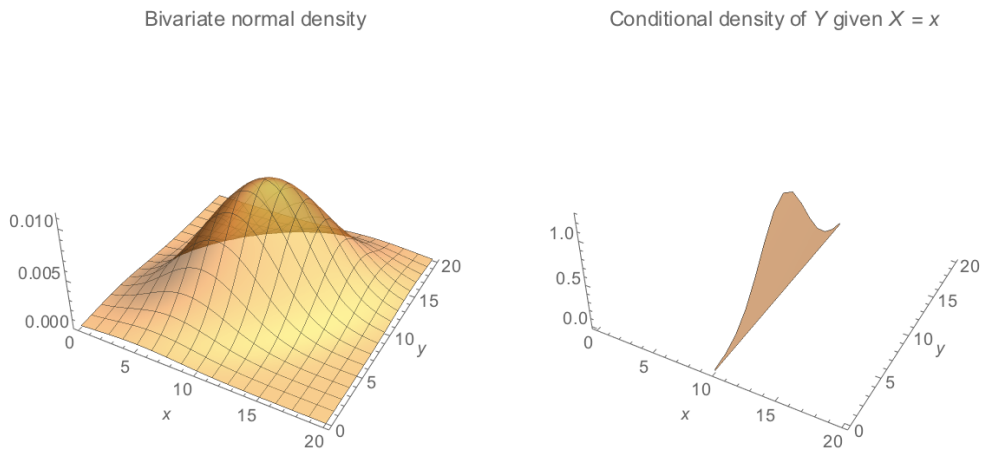


Figure 1.9: Example of a bivariate Gaussian joint probability density (left) and conditional probability density when $X = x = 10$ (right). The conditional probability density can be thought of as a renormalized cross-section of the joint probability density at a fixed value of one of the random variables. The parameter values used can be found in table 1.1.

The marginal probability densities associated with variables X and Y can also be found directly from the joint probability density function through integration of the joint probability density over each of the random variables respectively. The marginal probability density of X is

$$f(x) = \int_{-\infty}^{\infty} f(x, y)dy \tag{1.15}$$

and the marginal probability density of Y is

$$f(y) = \int_{-\infty}^{\infty} f(x, y)dx. \tag{1.16}$$

Using the example bivariate Gaussian above, we can obtain the following marginal probability densities through use of these equations

$$f(x) = \frac{e^{\frac{-1}{32}(-10+x)^2}}{4\sqrt{2\pi}} \tag{1.17}$$

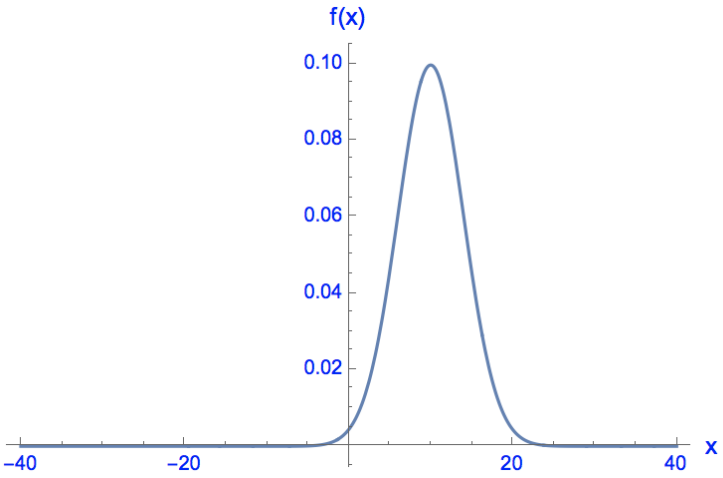


Figure 1.10: Marginal probability density $f(x)$ from the example bivariate Gaussian above.

$$f(y) = \frac{e^{\frac{-1}{32}(-10+y)^2}}{4\sqrt{2\pi}} \tag{1.18}$$

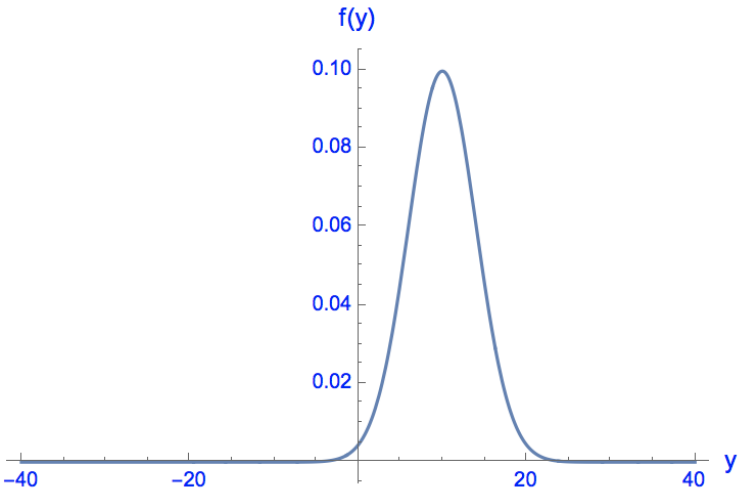


Figure 1.11: Marginal probability density $f(y)$ from the example bivariate Gaussian above.

The above discussion of marginal, joint, and conditional probability densities gives sufficient foundational knowledge to understand how the marginal, joint, and conditional entropies used in this work are constructed.

1.7 Overview of the Thesis

In this work we use a variations of Shannon entropy, the total entropy rate, local entropy rate, and specific entropy rate, to explore the information dynamics of a nonlinear stochastic nanopore system. This work represents the first such application of information dynamics to nanometer-scaled objects. Through study of simulated and empirical nanopore data, we demonstrate the utility of both local and specific entropy rate in tracking the dynamics of the nanopore current oscillations. We additionally probe for deeper, previously unknown structure in these nanopore current oscillations by analyzing the intervals between oscillations.

In this work we will discuss information dynamics in the continuous case, building a foundation for understanding from the ground up. We will then discuss application of the noted Shannon entropy measures to the simulated and experimental nanopore data and discuss the discoveries we have made therein. Our discoveries to date have opened doors to several interesting pathways for further research including application of an extension of specific entropy rate, the normalized q-step specific entropy rate, and preliminary testing to identify non-linear structure to nanopore interevent interval sequences through surrogate time series analysis. Finally, we will summarize our findings and future research plans in the conclusion. Appendices can also be found after the text to provide further information and results.

Chapter 2

Shannon Entropy and Entropy Rates

2.1 Differential Entropy

2.1.1 Definition of Differential Entropy

Differential entropy, $H[X]$ is the entropy associated with a continuous random variable, X [12]. This Shannon entropy is defined in terms of the marginal probability density $f(x)$ as the following, if such a density function exists for that variable AND the above integral exists

$$H[X] = -E[\log f(x)] = -\int_{-\infty}^{\infty} f(x) \log f(x) dx \tag{2.1}$$

where the logarithm will be taken to be in base e in this work.

In contrast to discrete entropy, which is always positive, differential entropy can be positive or negative. A simple example of a uniform distribution can be used to illustrate this point.

Consider the following uniform distribution

$$f(x) = \begin{cases} c & \text{if } x \text{ in } [0, a] \\ 0 & \text{otherwise.} \end{cases}$$

The standard normalization condition lets us ascertain that $c = \frac{1}{a}$ [12]. Plugging this into the equation for differential entropy, we obtain the following

$$H[X] = -\int_{-\infty}^{\infty} f(x) \log f(x) dx = -\int_{0}^{a} \frac{1}{a} \log \frac{1}{a} dx. \tag{2.2}$$

When we perform the integration, we see that

$$H[X] = -\frac{a}{a} \log \frac{1}{a} + \frac{0}{a} \log \frac{1}{a} = -\log \frac{1}{a} = \log a. \tag{2.3}$$

a is greater than zero, but it is not necessarily greater than 1. If a is indeed less than 1, $H[X]$ will be negative.

Now that we know the definition of differential entropy and have seen it applied to a specific example, the uniform distribution, it might be logical to ask what it tells us. Conceptually, it tells us, on average, how surprised we are to have seen X (a generic X), given that we know the probability density $f(x)$ [12]. It is an overall measure and is sequence-insensitive. Sequence-insensitive means that changing the order of values in a realization of a time series representing a continuous valued variable will result in the same differential entropy as the original ordering.

2.1.2 Conditional Differential Entropy

In the section above, we defined differential entropy, discussed the question it addresses, and explored a key difference between it and discrete entropy through an example. What if, however, we are not interested in an overall, sequence-insensitive measure of statistical surprisal? Perhaps instead we are interested in a time-dependent (i.e. a sequence-sensitive) measure of entropy. Sequence sensitive measures of entropy are also referred to as conditional entropies.

The simplest conditional entropy can be constructed without explicit time dependence. We can ask the question "what is the entropy of Y given a known value for X?" In similar form to the differential entropy, conditional entropies are expressed in terms of the conditional probability density. In the case of continuous variables, this takes the following form

$$H[Y \mid X] = -E[\log f_{Y|X}(Y \mid X)] = -\int_{x=-\infty}^{\infty} \int_{y=-\infty}^{\infty} f_{X,Y}(x,y) \log f_{Y|X}(y \mid x) dy dx \tag{2.4}$$

where the weighting is over the joint probability density as it is representative of all possible pairings of X and Y. This statement tells us what the average surprise is at seeing Y given that we already know X, averaged over all possible values of X [12].

It is also common to see the conditional entropy expressed in terms of a difference between the joint and marginal entropy [12]

$$H[Y \mid X] = H[X,Y] - H[Y] \tag{2.5}$$

and in practice, conditional entropies may be computed using this definition for convenience.

As an aside, the joint entropy can also be expressed in integral notation as the following

$$H[X,Y] = -E[\log f_{X,Y}(X,Y)] = -\int_{x=-\infty}^{\infty}\int_{y=-\infty}^{\infty} f_{X,Y}(x,y)\log f_{X,Y}(x,y)dydx. \quad (2.6)$$

The statement $H[Y \mid X] = H[X,Y] - H[Y]$ is simplest to see graphically, with an I-diagram, a Venn diagram for information theory relationships [85, 95]. In an I diagram, circles representing the entropy associated with two distinct random variables, X and Y are shown overlapping. The area of overlap is known as the mutual information, $I[X;Y]$, between X and Y. The outline of the overlapping circles represents the joint entropy $H[X,Y]$ and the left and right crescent shapes represent the conditional entropies $H[X \mid Y]$ and $H[Y \mid X]$ respectively. We show a detailed version of an information theory I diagram below.

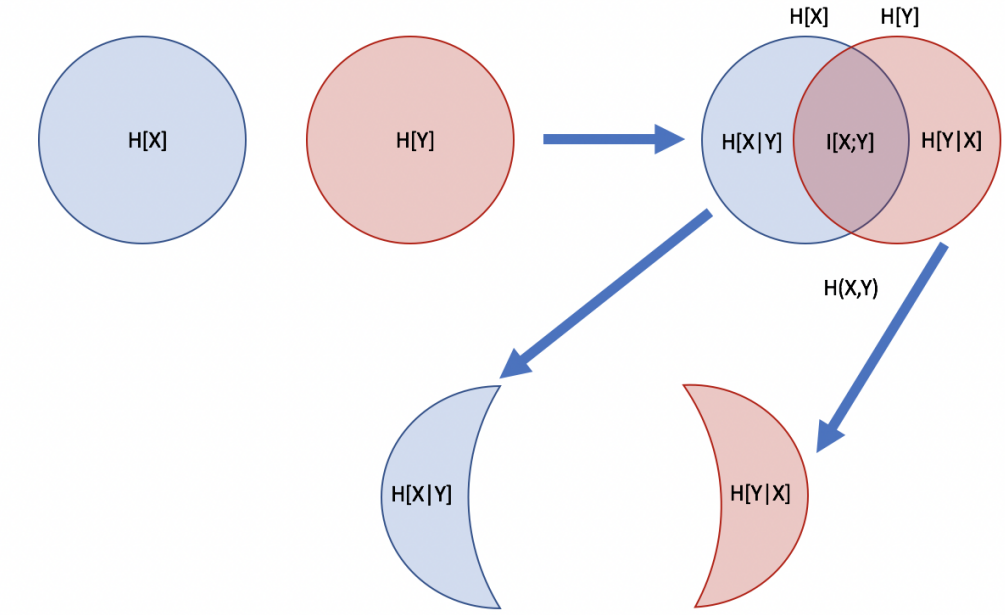


Figure 2.1: Pictorial representation of Shannon entropies using an I-diagram. A typical I-diagram is shown in the top right (overlapping circles). The additional content in the figure is added for step-by-step conceptual clarification. We begin with two variables X and Y, each of which have an associated marginal entropy $H[X]$ and $H[Y]$, represented by the separate circles on the top left. The intersection of the two circles is the mutual information $I[X;Y]$. The left and right crescent shapes represent the conditional entropies $H[X \mid Y]$ and $H[Y \mid X]$ respectively. The external outline of the I-diagram represents the joint entropy $H[X,Y]$.

Using the I-diagram, we can readily see the relationship between the marginal, joint, and conditional entropies.

2.1.3 Definition of Entropy Rate

We have now constructed a sufficient foundation to construct entropy rates, which are conditional entropies with explicit time dependence. The simplest example of an entropy rate arises when we ask the question "what is the average surprise at seeing the immediate future X_t given my knowledge of the past vector $X_{-\infty}^{t-1}$?" In this case, we need not fear the notational complexity - X_t simply refers to Y from 2.6 and $X_{-\infty}^{t-1}$ refers to X from 2.6. To answer the question above, we need to construct the entropy rate

$$H[X_t \mid X_{-\infty}^{t-1}] = -E[\log f_{X_t|X_{-\infty}^{t-1}}(X_t \mid X_{-\infty}^{t-1})] \tag{2.7}$$

$$= -\int_{x_t \in \mathbb{R}} \int_{x_{-\infty}^{t-1} \in \mathbb{R}^\infty} f_{X_t, X_{-\infty}^{t-1}}(x_t, x_{-\infty}^{t-1}) \log f_{X_t|X_{-\infty}^{t-1}}(x_t \mid x_{-\infty}^{t-1}) dx_{-\infty}^{t-1} dx_t \tag{2.8}$$

again, if it exists [14]. When computing an entropy rate on a tangible set of data we cannot look infinitely far into the past. We may look at a past vector that proceeds a particular number of steps p into the past. This gives us the following entropy rate equation

$$H[X_t \mid X_{t-p}^{t-1}] = -E[\log f_{X_t|X_{t-p}^{t-1}}(X_t \mid X_{t-p}^{t-1})] \tag{2.9}$$

$$= -\int_{x_t \in \mathbb{R}} \int_{x_{t-p}^{t-1} \in \mathbb{R}^p} f_{X_t, X_{t-p}^{t-1}}(x_t, x_{t-p}^{t-1}) \log f_{X_t|X_{t-p}^{t-1}}(x_t \mid x_{t-p}^{t-1}) dx_{t-p}^{t-1} dx_t \tag{2.10}$$

where X_{t-p}^{t-1} is the block of states from p steps to 1 step in the past [14]. 2.8 can be thought of as the limit of $H[X_t \mid X_{t-p}^{t-1}]$ as p approaches ∞. This gives us the average surprise at seeing the immediate future X_t given the past vector X_{t-p}^{t-1}. Another way to think about the

knowledge gained from computing a p-step entropy rate is that it tells us how uncertain we are about particular pairings of the past and future obtained by averaging over all possible pairings of the past and future.

2.1.4 Local Entropy Rate

The time-dependent entropy rates previously discussed are averages over all possible pasts and all possible values of the immediate future. In practice, if we take a set of measurements of a random variable X for any given point in time we will already know the realized values of the past vector x_{t-p}^{t-1}. If we then view x_t, we can ask the question "how surprised am I to have seen this particular, realized value of x_t given that I know the realized past x_{t-p}^{t-1}?" To answer this question we need not evaluate any averages over pasts or futures. We need only look at the expectand of 2.9 (i.e. the content within the expectation value). This quantity is called the local entropy rate, $H^L(x_t \mid x_{t-p}^{t-1})$ as it is local to that particular time point x_t [49, 48, 47]. The local entropy rate is defined as

$$H^L(x_t \mid x_{t-p}^{t-1}) = -\log f_{X_t|X_{t-p}^{t-1}}(x_t \mid x_{t-p}^{t-1}) \tag{2.11}$$

which can be re-expressed in terms of the joint and the marginal entropy as the following

$$H^L(x_t \mid x_{t-p}^{t-1}) = -\log \frac{f_{X_t,X_{t-p}^{t-1}}(x_t, x_{t-p}^{t-1})}{f_{X_{t-p}^{t-1}}(x_{t-p}^{t-1})}. \tag{2.12}$$

The figure below offers a visual representation of a possible question which would require calculation of local entropy rate to answer.

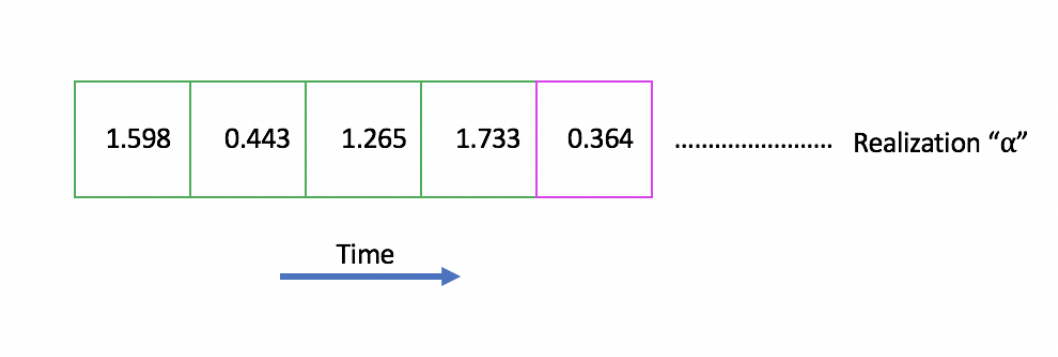


Figure 2.2: This figure shows a realization of a continuous valued time series. The measurements boxed in green represents the past $p = 4$ steps and the measurement boxed in pink represents the immediate future. Local entropy rate, LER, asks the question "how surprised are we to have seen the measurement in pink given that we observed the measurements in green?"

2.1.5 Specific Entropy Rate

The previously discussed expressions of entropy rates require observation of the immediate future and allow us to see how surprised we were to see that future given knowledge of the past. What happens if instead of asking how surprised we are to have seen a particular future, we ask how uncertain we are about the immediate future we have not yet seen given our knowledge of the past? This question may be answered by constructing a strategic average of local entropy rate values, where the average is taken only over future states [14]. This new quantity, the average of local entropy rate values over the future space, is known as the specific differential entropy rate, or specific entropy rate (SER) and is defined in terms of the local entropy rate as

$$H^S(x_{t-p}^{t-1}) = -E[H^L(X_t \mid X_{t-p}^{t-1}) \mid X_{t-p}^{t-1} = x_{t-p}^{t-1}] \tag{2.13}$$

which can be expressed in terms of the conditional probability density as follows

$$H^S(x_{t-p}^{t-1}) = -\int_{x_t \in \mathbb{R}} f_{X_t|X_{t-p}^{t-1}}(x_t \mid x_{t-p}^{t-1}) \log f_{X_t|X_{t-p}^{t-1}}(x_t \mid x_{t-p}^{t-1}) dx_t. \tag{2.14}$$

The following figure represents the question specific entropy rate aims to address.

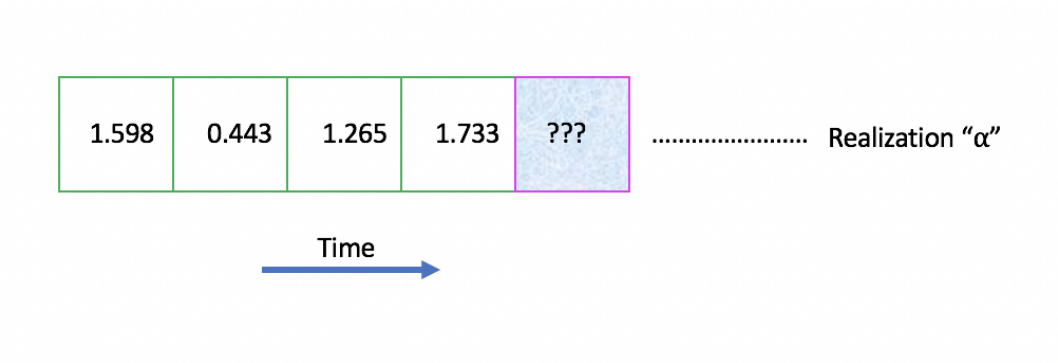


Figure 2.3: This figure shows a realization of a continuous valued time series. The measurements boxed in green represents the past $p = 4$ steps and the measurement boxed in pink represents the immediate future, which is unknown. Specific entropy rate, SER, asks the question "how uncertain are we about the measurement in pink given that we observed the measurements in green?"

2.1.6 Summary of Entropy Rate Definitions

We have now constructed a mathematical framework that can be used to answer several different interesting questions about a time series representing a random continuous valued variable. It should be noted, however, that as written, all of the previously discussed equations require a system with a known probability density. In practice we may not have an analytical statement of the probability density function and may just have realizations from a stochastic process. In such cases the probability density much be estimated from the data. The next section will highlight the density estimation techniques used in this work and will discuss their application.

2.2 Model Order Selection and Probability Density Estimation

In this section we will address the methods used in this work to estimate probability density functions from data. As in other places in this work it is important to build a foundation to enhance understanding. We will begin by exploring accepted techniques for density estimation, including kernel density estimation, kernel nearest neighbor estimation, and k^{th} nearest neighbor estimation (also referred to as k nearest neighbor estimation).

The first step to probability density estimation involves identifying the number of steps of past information necessary to appropriately estimate the probability density at a given point in time. This number of steps in the past, which we have referred to as p in previous sections, is known formally as the model order [94]. We will also specifically discuss the process of selecting the model order [15]. It will be important for us to proceed into this discussion with the express mindset that model order selection and probability density estimation are two distinct activities.

Because model order selection is the first step to entropy rate estimation, it might make sense to discuss it in detail first. As it turns out, however, the process of identifying the model order relies on density estimation techniques. For that reason we will enter the discussion by presenting methods of density estimation as stand-alone techniques before explaining their role in context and discussing model order selection.

2.2.1 Kernel Density Estimation

Kernel density estimation (KDE) is a technique that uses a kernel function K to smooth the data around each point in a data set. The kernel functions at each point are then

summed to produce an estimate of the probability density. Kernel functions are positive-valued functions that must integrate to 1 and they can take many different forms (uniform distribution, Gaussian distribution, etc.) [94, 74, 77]. Because most kernels have comparable efficiency, the functional form is a matter of choice. We have selected a Gaussian kernel for this work, which takes the general form for a variable u

$$K(u) = \frac{1}{2\pi} e^{-\frac{u^2}{2}}. \tag{2.15}$$

We can think of this as placing a Gaussian function on top of any particular data point [94, 74, 77].

To estimate the probability density function from data generated by a stochastic process at a particular point we can sum the contributions from all of the kernels at the particular point in question. This method gives us the following estimate for the normalized probability density function

$$\hat{f}_X(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{X_i - x}{h}\right), \tag{2.16}$$

where u from 2.15 becomes the distance between the i^{th} data point and x, N is the total number of data points, and h is the bandwidth. The bandwidth h can be thought of as a smoothing parameter that determines the degree of smoothing the kernel will do. While the choice of kernel function is not critical to obtaining a good estimate of the density, the choice of the bandwidth is. The simple example of points from a random, normally distributed sample below highlights how the bandwidth impacts the estimation.

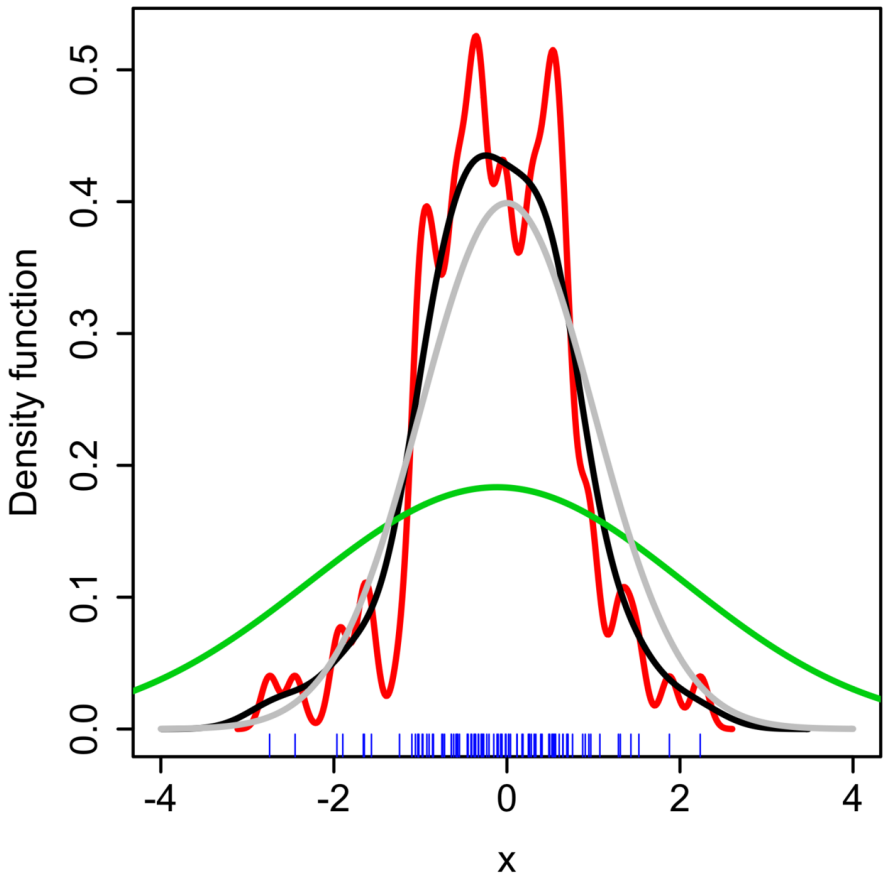


Figure 2.4: This figure from [45] shows points from a normally distributed sample and shows the impact of bandwidth selection on kernel density estimation performed to estimate the probability density function. The grey is the true normal probability density and the red, black, and green are obtained using kernel density estimation with kernel bandwidths h=0.05, 0.337, and 2 respectively.

Consequently, the proper bandwidth should be chosen via an optimization technique. This will be discussed briefly in a subsequent section.

2.2.2 Kernel Nearest Neighbor Estimation

While kernel density estimation is a well accepted technique for estimating probability density functions, one of its major drawbacks is that is it computationally expensive to compute the distance between each point and all other points. The contribution to the overall kernel

from points that are spatially far away from a particular x can be expected to be negligible. What do we mean by "spatially far" in this context? We mean behaviorally different and this is easiest to understand with a visual aid. The figure below shows a three dimensional space constructed from points in the form (X_{t-2}, X_{t-1}, X_t).

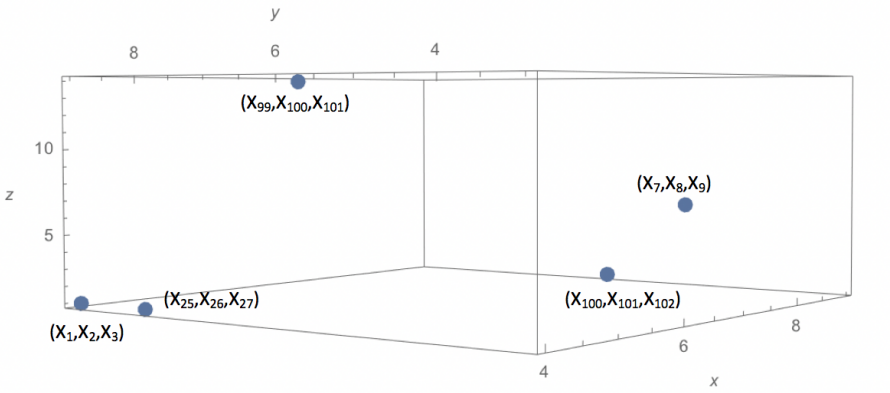


Figure 2.5: This figure shows points plotted in the form (X_{t-2}, X_{t-1}, X_t). Points spatially close in this behavior space will be most contributory to the density at a particular point of interest, often called the evaluation point. The figure highlights that these spatially nearby points need not be temporally nearby.

Points that are temporally close in time may not be behaviorally similar. We seek a set of behaviorally similar points nearest to a particular point. These behaviorally similar points will be most helpful in estimating the probability density at our point of interest. With this in mind, we use a hybrid density estimator that combines kernel density estimation with a separate estimation technique nearest neighbor regression [46, 15]. The hybrid kernel nearest neighbor technique allows us to include only points that make a non-negligible contribution to the density at any particular x. As we might expect, the number of points used will also be a parameter obtained via an optimization technique. This will also be discussed in the following subsection, where it will be notated as J.

2.2.3 Model Order Selection in This Work

The following has been reproduced from [23]: "In analyzing stochastic systems, the goal is to determine whether information contained in the past is useful to understanding the immediate future. If we do not look back far enough, we may miss useful information from the past that could help us determine the future. If we look too far back, we may inadvertently include information not relevant to determining the immediate future. As a concrete example, consider the case where the observable is well-modeled by a Markov process of some order p. In that case, after knowing the previous p values of the process, the future is independent of any values further in the past, so those values do not aid in the prediction of the process. However, they do increase the burden of the associated estimation problem, through the curse of dimensionality [78]. That is, geometrically more data is necessary to achieve the same level of precision in the estimate of the density. In the case where the process is not Markovian, a similar argument applies, except with the added consideration for balancing between the contribution of including more of the past and its increasing burden to estimation process. We seek the model order that results in a minimum uncertainty. Using the information theoretic criterion from [15], the model order is chosen to minimize the negative log predictive likelihood (NLPL)

$$\mathrm{NLPL}(p) = -\frac{1}{N-p}\sum_{i=p+1}^{N} \log \hat{f}_{-i}(X_i \mid X_{i-p}^{i-1}) \tag{2.17}$$

where N is the number of points in the time series, and $\hat{f}_{-i}$ is an estimator of the predictive density estimated holding out the block X_{i-p}^{i}. We use a kernel-nearest neighbor estimator for f, which performs kernel density estimation over the set of nearest neighbors in the future space [46, 38]. The estimator takes the form

$$\hat{f}_{-i}(X_i \mid X_{i-p}^{i-1}) = \frac{1}{J}\sum_{m \in \mathcal{N}_J(X_{i-p}^{i-1})} K_h(X_i - X_m) \tag{2.18}$$

where K_h is a Gaussian kernel, J is the number of nearest neighbors to X_{i-p}^{i-1}, $\mathcal{N}_J(X_{i-p}^{i-1})$ is the index set of the J-nearest neighbors, and h is a bandwidth for the density estimator over futures. Equation (2.17) over h, J, and p is optimized using the constrained Nelder–Mead method from the NLopt library, where h is constrained to $(0, \infty)$ and J is constrained to $\{1, ..., J_{max}\}$, where $J_{max} \ll N$ [57]."

We should note to clarify the above statements that the Nelder-Mead optimization routine is performed at fixed values of p in a predefined range. At each value of p the optimization can be performed over the variables J and h, the number of nearest neighbors and the kernel bandwidth respectively. Ultimately, the sole purpose of this optimization process is to identify the model order that minimizes the negative log predictive likelihood, and as such, the value of p which accomplishes this goal is selected. An upper bound may be set on the number of nearest neighbors to use for computational efficiency but it is important that we not set the upper bound too low. If we do, we risk identifying a local rather than a global minima of the negative log predictive likelihood. One way to avoid this unwanted consequence is to carefully monitor the value of J chosen by the optimization routine and ensure that it is sufficiently below the upper bound.

We should also explicitly address the mathematical space in which the evaluation point and its neighbors exist. Because we use the kernel nearest neighbor technique described in the previous section to construct the probability density, the space in question can be thought of as a p-dimensional embedding space with points taking the form $(x_{t-p}, ..., x_{t-2}, x_{t-1}, x_t)$. We note that the nearest neighbors used for computing the kernel nearest neighbor estimated density are the points spatially nearest the evaluation point in this embedding space, not the points temporally nearest x_t. The graphic shown in the previous subsection, Figure 2.5 serves as a visual reminder.

Once we have computed the optimal model order, we do not utilize the same density estimator to perform subsequent entropy rate calculations, as the primary goal of this procedure

was to construct a sufficient density estimator for use in determining the optimal model order. The following subsection will introduce the density estimation technique that will be used to construct the probability density estimator for computing estimators of entropy rates.

2.2.4 k^{th} Nearest Neighbor Estimation

We compute the estimator of the probability density which will be used to compute local entropy rate by using a variant k nearest neighbor method, wherein the density estimate is inversely proportional to volume of a p dimensional sphere with radius equal to the distance between the evaluation point and its k^{th} nearest spatial/behavioral neighbor [81, 50, 44].

k is pre-determined and must be thought of in the context of the expected use of the local entropy rate. If the density estimator is intended to be used to compute a local entropy rate estimator en route to computing specific entropy rate estimator, k can be chosen to be small because computing the estimator for specific entropy rate involves averaging many local entropy rate estimators, thereby averaging out the impact of variance in the estimator. If, however, the density estimator is intended to be used to compute a free-standing estimator of the local entropy rate (in other words, a superiorly estimated local entropy rate), where no subsequent averaging will be performed, k must be chosen to scale with a power of N in order to ensure consistency of the estimator (i.e. that the estimator converges to the true probability density) [3]. The expression for estimating the density for a block of states X_{i-p}^{i-1} using a k^{th} nearest neighbor method is as follows

$$\hat{f}_{X_{i-p}^{i-1}}(x_{i-p}^{i-1}) = \frac{k}{N-d+1}\frac{1}{V_{r_d}} \tag{2.19}$$

where N is the total number of points, d is the dimension of the past vector X_{i-p}^{i-1}, and V_{r_d}

is the volume of a d dimensional sphere

$$V_{r_d} = \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)} r^d \tag{2.20}$$

where r is the Euclidian distance between the evaluation point x_{i-p}^{i-1} and its k^{th} nearest neighbor in the distribution of the past vectors, X_{i-p}^{i-1}. $d = p$ in the case of estimation of a marginal probability density. To compute the probability density estimator for a conditional probability (i.e. $f_{X_i|X_{i-p}^{i-1}}$, we recall the fact that a conditional probability density can be written as the joint probability density divided by the marginal probability density. k and N will be the same for both the joint and marginal probability densities, but d for the joint probability density will be one higher than d for the marginal probability density. We will have the following:

$$\hat{f}_{X_i|X_{i-p}^{i-1}}(x_i \mid x_{i-p}^{i-1}) = \frac{V_{r_{d+1}}}{V_{r_d}} \tag{2.21}$$

where we assume d corresponds to the dimension of the marginal (past) distribution.

2.2.5 Computing Entropy Rate Estimators

Once the optimal model order has been selected and the density estimator has been computed, we used the following expressions to compute the entropy rate estimators [16]

Local Entropy Rate Estimator

The estimator of local entropy rate at time t_i, $\hat{H}_{t_i}^L$ is [49, 47, 48, 14]:

$$\hat{H}_{t_i}^L = -\log\left\{\frac{\hat{f}_{X_i,X_{i-p}^{i-1}}(x_i,x_{i-p}^{i-1})}{\hat{f}_{X_{i-p}^{i-1}}(x_{i-p}^{i-1})}\right\}. \tag{2.22}$$

Specific Entropy Rate Estimator

For practical purposes, the integral in 2.14 won't be directly calculable. The estimator for specific entropy rate at each time t_i is calculated in practice by taking k^* values of the estimated local entropy rate and averaging them. The particular estimated values of the local entropy rate are chosen such that

$$\hat{H}^S_{t_i} = \frac{1}{k^*} \sum_{j \in \mathcal{N}_{k^*}(X^{i-1}_{i-p})} \hat{H}^L(X_j \mid X^{j-1}_{j-p}) \tag{2.23}$$

where $\mathcal{N}_{k^*}(X^{i-1}_{i-p})$ is the index set of the k^* nearest neighbors of X^{i-1}_{i-p} and k^* is set at $\sqrt{N}$ (to ensure consistency) [14].

Total Entropy Rate Estimator

We will find in this work that it is also useful to have an estimate for an overall entropy rate for a realization of a stochastic process. By taking the arithmetic average of the local entropy rate measures computed for each time point we arrive at the estimated total entropy rate for the realization. This estimated total entropy rate will be utilized in this work to help identify non-linear structure in a series of nanopore interevent intervals (see chapter 5). The estimated total entropy rate is

$$\hat{H}^T = \frac{1}{N-p} \sum_{i=1}^{N-p} \hat{H}^L_i \tag{2.24}$$

Chapter 3

Local and Specific Entropy Rates in Nanopore Simulation Study

We report and discuss the results of a simulation study in which estimators for local and specific entropy rate were computed for data obtained using the methods and parameters discussed in chapters 1 and 2. The results are presented as an excerpt from for a single simulated realization [23].

"Figure 3.1 shows the nanopore current, the LER, and the SER for a group of transitions in the nanopore system (pore open/close events). All three panels are aligned in time. The uppermost panel represents the nanopore current as a function of time. Each orange point is one measurement. Open/close (transition) events can be seen in the rapid switching of the nanopore current from positive to negative values or vice versa. The middle panel is the estimated free-standing LER, computed using Equation (2.22). We can see that there are peaks in the free-standing LER aligned with the transition events of the nanopore. This indicates that information is generated by these events, and there was some surprise associated with their occurrence. The bottom panel is the SER estimate computed via

Equation (2.23). We see that the SER also increases around the transition events in the nanopore, indicating an increase in uncertainty about future states near the transitions. These results are a subset of results from a single realization, but are representative of the behavior seen in these information measures during transition events across all five realizations, as is expected for this model.

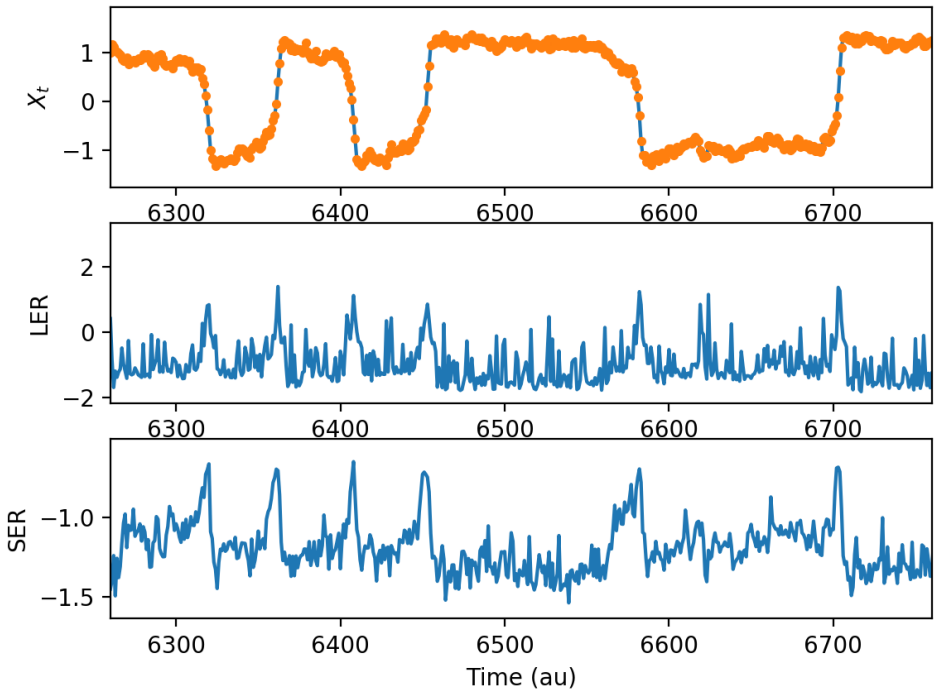


Figure 3.1: Top: the nanopore current, with each orange dot representing a measurement. Middle: the estimate of the local entropy rate (LER) of the nanopore system as a function of time. Bottom: the estimate of the specific entropy rate (SER) of the nanopore system as a function of time. This is a representative excerpt from a 40,000 point time series containing on the order of 100 transitions.

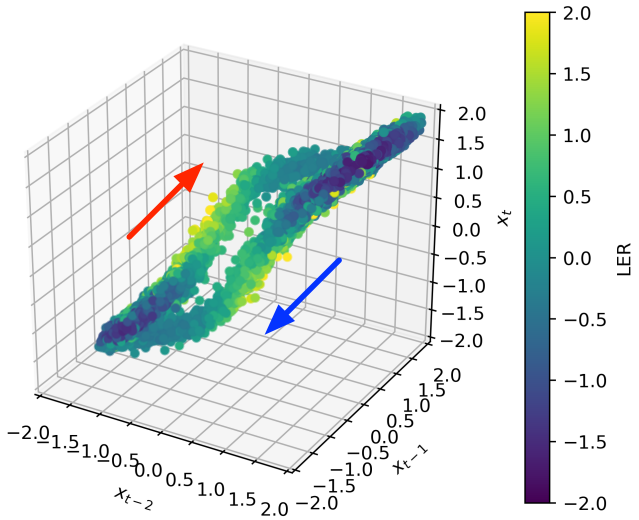
The peaks in the LER and SER corresponding to transitions should be considered in the context of the model. It will be easiest to make a transition between positive and negative currents when the slope of the potential is greater (i.e., when the magnitude of y is further from zero). Additionally, when the slope of the potential is large, any small kick from the dynamical noise could lead to a wider array of possible futures (noise amplification).

There will accordingly be a greater uncertainty in an unseen future (higher SER) during the transition events. There will also be an elevated LER in these regions because, of the many possible outcomes for current values, when the potential slope is large, individual outcomes may occur only rarely. This will translate to a relatively high surprise.

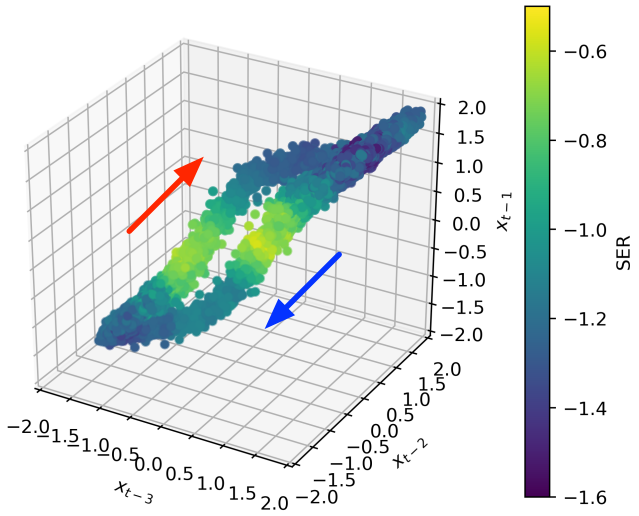
It should also be noted that there are substantial peaks in the LER that are not always associated with transitions. The LER metric is sensitive to viewing any atypical future. In the relatively flatter (low variation in nanopore current) regions between transitions, any variation above the noise level from the anticipated trajectory may result in high surprise, even though a transition may not occur. This is particularly prominent about 6430 au and 6620 au in Figure 3.1. If the future is unseen, as in the SER, in these relatively flatter regions there will be low uncertainty about the future. In other words, variation above the noise level is not expected. This is why similar peaks not associated with transitions are rarely seen in the SER.

To further investigate transitions, we consider how the LER and SER vary as a function of the reconstructed state space of the nanopore system. Figure 3.2 shows a 3D projection of the $p = 4$ reconstructed state space, where each point is shaded by the LER (left) and SER (right). We use the projection (X_{t-2}, X_{t-1}, X_t) for the LER and $(X_{t-3}, X_{t-2}, X_{t-1})$ for the SER. The arrows indicate the direction of the transitions with respect to time. We know that, if the nanopore is in a closed state, it is likely to remain closed and that, if it is in an open state, it is likely to remain open. We thus see relatively low surprise (a low LER) and relatively low uncertainty (a low SER) under those conditions, corresponding to the points in the bottom left and top right of the reconstructed state space. When a transition event is occurring, corresponding to the points along the central "tubes", we are relatively more surprised (a higher LER) and relatively more uncertain about the immediate future (a higher SER). It should also be noted that, if all transitions in this system were identical, the reconstructed state-space trajectory would not show spread about the average path (i.e.,

the "fuzziness" is due to differences in the profile of the transitions). Although the LER and SER have similar regions of relatively high/low values in their corresponding measures, it is important not to conflate them. The LER measures information generated by seeing the future, and it should not surprise us to see a high degree of symmetry in the LER plot, with maximal information generated for more atypical transitions (on the outsides of the transition "tubes"). The SER, by contrast, measures uncertainty in the future, given a known past, and we might expect high uncertainty in the region of all transitions. It is additionally noteworthy that there is some anti-symmetry between the two transition tubes in the SER plot with respect to the location of the onset of elevation in the SER. This, together with the arrows indicating the direction of the trajectory with time, shows that uncertainty is highest at the beginning of a transition. Uncertainty decreases as the transition proceeds to completion. This is not apparent from examination of the time series.



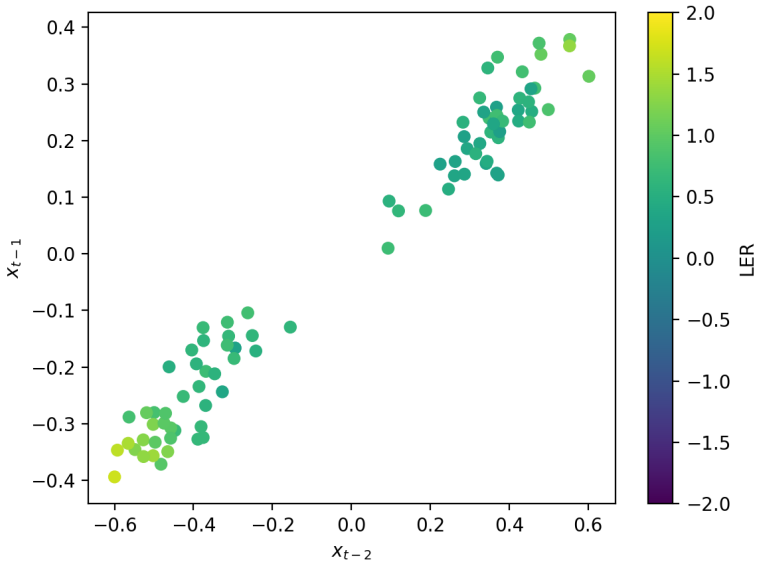
(a)



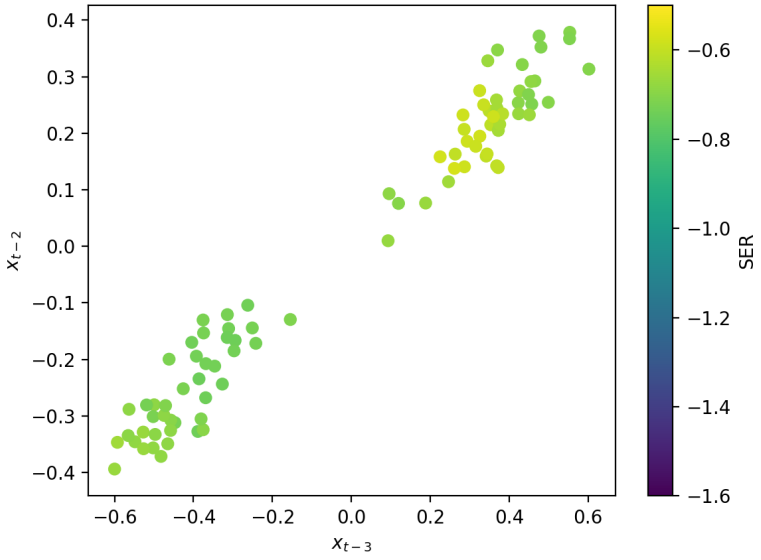
(b)

Figure 3.2: A projection of the reconstructed state space for the nanopore system shaded by the estimates of the LER (a) and SER (b) associated with the overall state. The plots reveal a clear trajectory in the reconstructed state space, and the arrows indicate the direction along the transitions between open and closed states. Along this trajectory, regions of relatively low surprise (LER) and low uncertainty (SER) occur when the system is in an open/closed state. Conversely, in the central regions, corresponding to transitions, we see increases in both the LER and SER. Anti-symmetry is noted in the onset of increase in SER, which shows that uncertainty is highest at the beginning of a transition and decreases as the transition proceeds to completion. (**a**) LER; (**b**) SER.

It may also be helpful to look directly at the transition region in both the LER and SER schemes, i.e., to look inside the trajectory in the transition regions. To do so, we take a cross section of the plots in Figure 3.1 at $x_t = 0$ and $x_{t-1} = 0$, respectively, and include points that fall within a tolerance of $\epsilon = \pm\ 0.05$. This cross section is shown in Figure 3.3. We can see that the LER is highest in the regions corresponding to less typical transitions (i.e., on the outside of the tubes), as previously mentioned. Additionally, as expected, all transitions in the SER scheme are associated with a similarly elevated SER.



(a)



(b)

Figure 3.3: A 2D cross section of the reconstructed state space constructed from points within $\epsilon = \pm\ 0.05$ of the $x_t = 0$ (a) and $x_{t-1} = 0$ (b) planes for each plot, respectively. These plots show that information is generated most heavily around atypical transition events (the highest LER visible on the periphery of the transition tubes in the LER plot), and there is relatively uniform, high uncertainty for all transitions in the SER plot. (**a**) LER; (**b**) SER."

We should explicitly state that the direction of the reconstructed state space trajectory is

identified by choosing a single transition event and plotting its associated reconstructed state space using a color scale for the time variable. An example for a simulated transition is shown below and is consistent with the direction arrows shown in 3.2a and 3.2b.

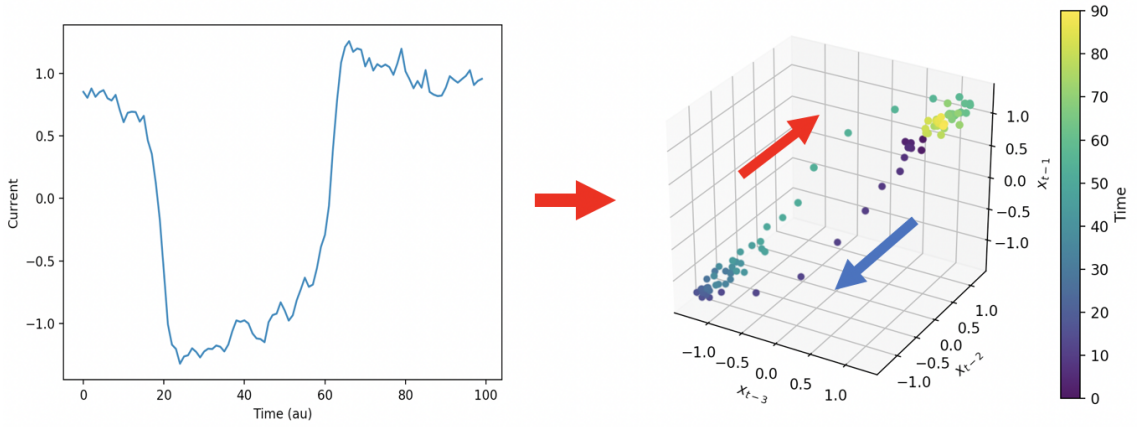


Figure 3.4: Left: A single simulated nanopore transition and Right: its corresponding reconstructed state space. The purpose of this plot is to demonstrate how the trajectory relates to features of the transition. The color is scaled by time, allowing us to follow the transition through the reconstructed state space trajectory.

This study demonstrated the utility of both local and specific entropy rate in tracking the dynamical behavior of this nanopore system. It also set the stage for further investigation into specific entropy rate, both in simulation and experimental nanopore data. In the next chapter we explore an extension of specific entropy rate, q-step specific entropy rate (cite personal communication with Dave). We will discuss the technique and its applications to simulation and experimental nanopore data.

Chapter 4

Q-Step Specific Entropy Rate In a Nanopore

4.1 What is q-step Specific Entropy Rate

To this point, our discussion of specific entropy rate has been limited to evaluating the uncertainty associated in the immediate future. We can think of this as evaluation of the uncertainty associated with the 0-step future. In this chapter, we explore a possible extension of the 0-step specific entropy rate and its utility for analyzing simulation and experimental nanopore data.

This extension is referred to as the normalized q-step specific entropy rate and its purpose is to provide a method for computing the divergence between the predictive probability density associated with q steps in the future X_{t+q} and the immediate future X_t. It is defined in terms of the 0-step conditional probability density as the Kullback-Leibler divergence from the q-step to the 0-step conditional probability density. A Kullback-Leibler Divergence quantifies the difference between an expected distribution and a secondary distribution [94, 12]. In this

case, the expected distribution is that associated with the 0-step future. The normalized q-step specific entropy rate is defined as the following

$$\tilde{H}(x;q) = \int_{\mathbb{R}} f_0(y \mid x) \log \frac{f_0(y \mid x)}{f_q(y \mid x)} dy \tag{4.1}$$

or

$$\tilde{H}(x;q) = \int_{\mathbb{R}} f_0(y \mid x) \log f_0(y \mid x) dy - \int_{\mathbb{R}} f_0(y \mid x) \log f_q(y \mid x) dy \tag{4.2}$$

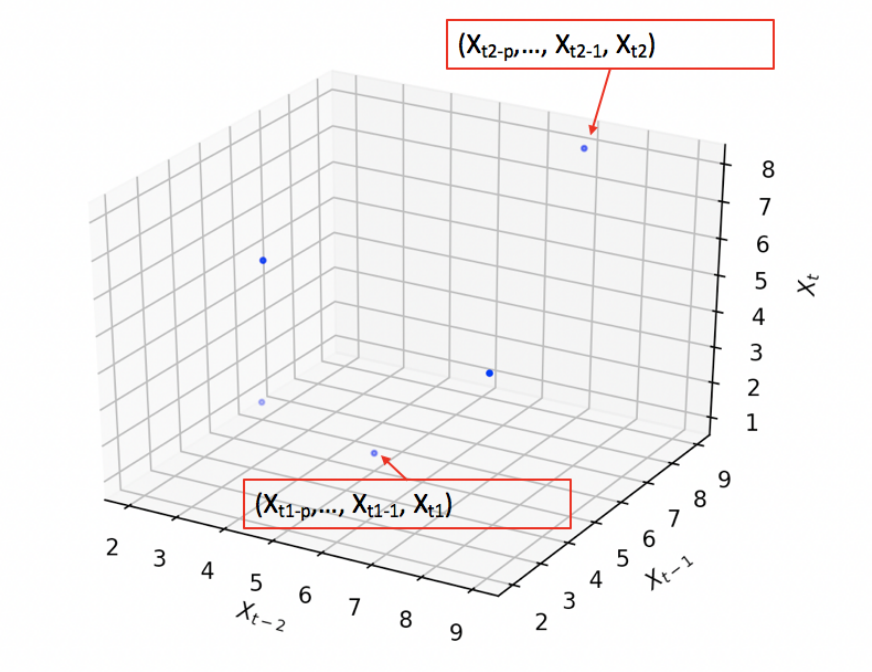
where $f_0(y \mid x)$ is the conditional probability density associated with the immediate futures given the past x and $f_q(y \mid x)$ is the conditional probability density associated with the q step future given the past x [13]. The normalized q-step specific entropy rate is identically zero when the q-step future looks the same as the 0-step future (i.e. when $f_0(y \mid x) = f_q(y \mid x)$), for almost all y, and is otherwise positive. We should again reiterate that the normalized q-step specific entropy rate is a relative measure taken to compare the q-step future to the 0-step future.

As is the case for computing 0-step specific entropy rate, we must compute estimators for the conditional density $f_0(y \mid x)$. The estimators for both the 0-step conditional probability density $f_0(y \mid x)$ and the q-step conditional probability density $f_q(y \mid x)$ are computed using the k^{th} nearest neighbor estimation procedure outlined in chapter 2 using the model order selected for the 0-step specific entropy rate. The known past vector for both conditional probability densities is thus X_{t-p}^{t-1}. The estimator for normalized q-step specific entropy is expressed as the following

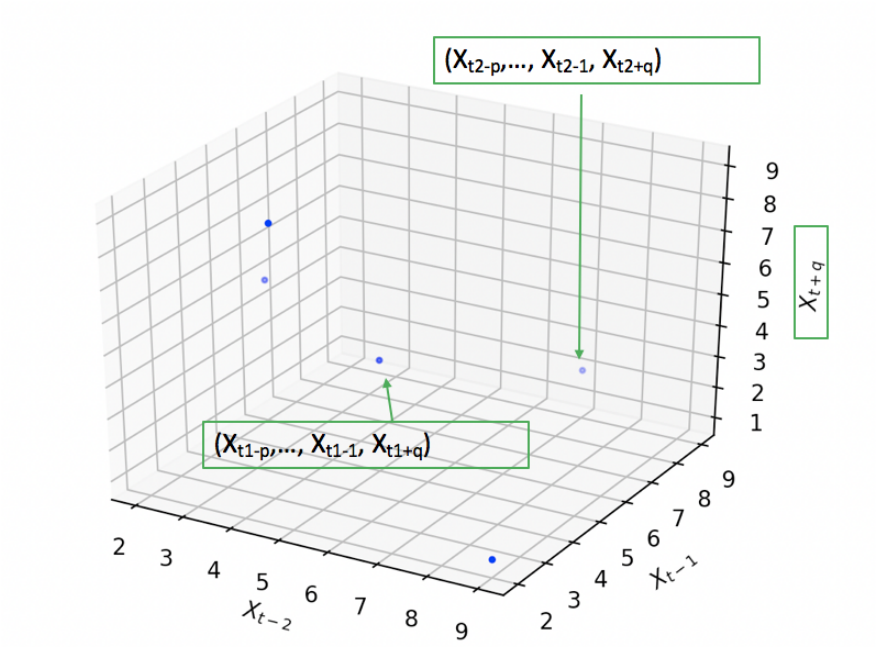
$$\hat{\tilde{H}}(x;q) = E_{\hat{f}_0}\left[\log \frac{\hat{f}_0(Y \mid X)}{\hat{f}_q(Y \mid X)} \middle| X = x \right] \tag{4.3}$$

where, x is the known past vector X_{t-p}^{t-1} [13]. We should take a moment before moving into the applications section to solidify understanding of this equation, in particular to solidify

understanding of the two distinct conditional probability density estimators. When we compute $\hat{f}_0(Y \mid X)$, we do so using the nearest neighbor method described in section 2.2.5, where Y is the immediate future X_t and the nearest neighbors to the evaluation point for the joint probability density and marginal probability density ((X_{t-p}^{t-1}, X_t) and (X_{t-p}^{t-1}) respectively) are found in the spaces created from the data in the format $(X_{t-p}, X_{t-p+1}, ..., X_{t-1}, X_t)$ and $(X_{t-p}, X_{t-p+1}, ..., X_{t-1})$ respectively. To compute $\hat{f}_q(Y \mid X)$, where again Y is the immediate future X_t, the only difference is in the construction of the nearest neighbor space. The new nearest neighbor spaces will be $(X_{t-p}, X_{t-p+1}, ..., X_{t-1}, X_{t+q})$ and $(X_{t-p}, X_{t-p+1}, ..., X_{t-1})$ for the q step future and the past spaces respectively. The projection onto three dimensional space below can be used as a guide.



(a)



(b)

Figure 4.0: This figure shows an example of the space that would be constructed to find nearest neighbors for estimating the joint probability density for the 0-step case (a) and the q-step case (b). These estimated joint probability densities would be used to compute the estimator for their respective conditional probabilities. It should be noted that this example space is shown in 3D for convenience, but the dimension of the space constructed for each estimator calculation will be p+1 dimensional, where p is the model order. These are example points shown for conceptual understanding.

For completeness, we also show the space constructed to find the nearest neighbors to estimate the marginal probability density, which is the same for both the 0-step and q-step, because it is purely a p dimensional past space.

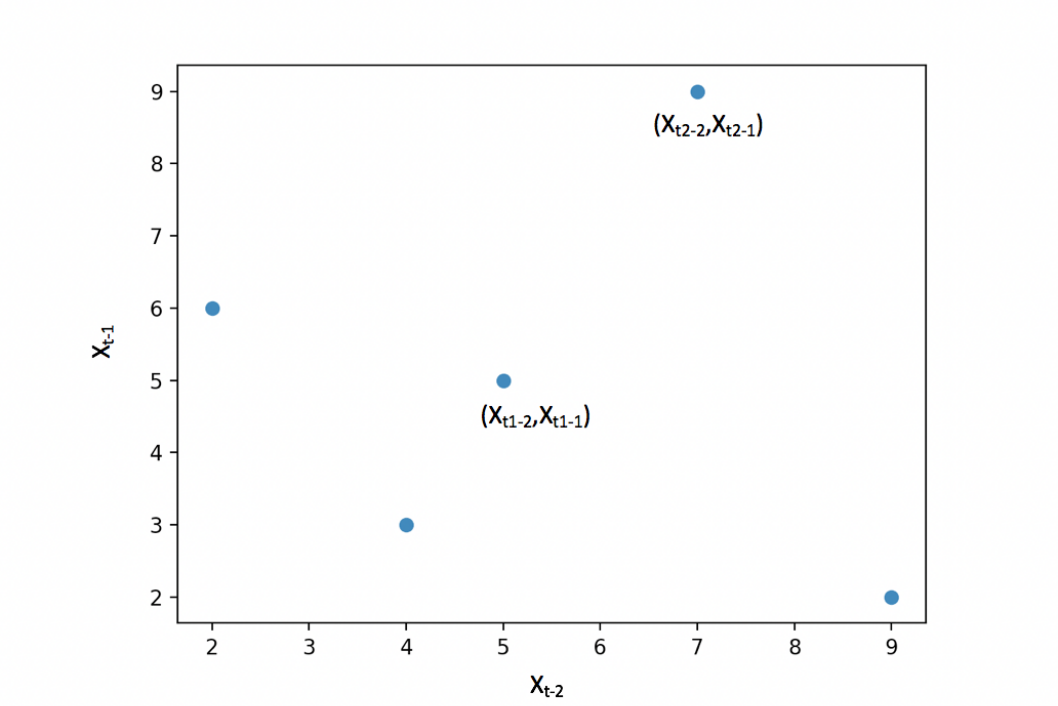


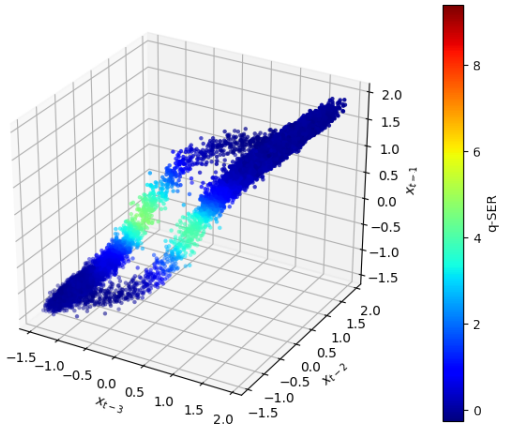
Figure 4.1: This figure shows an example of the space that would be constructed to find the nearest neighbors for estimating the marginal probability density for both the 0-step case and the q-step case. Again, this is shown in 2D for convenience, but the dimension of this space will be p dimensions. Again, these are example points only to aid in understanding, but they are consistent with the example plotted in the previous figure.

4.2 Application of q-step Specific Entropy Rate to Nanopore Simulation Data

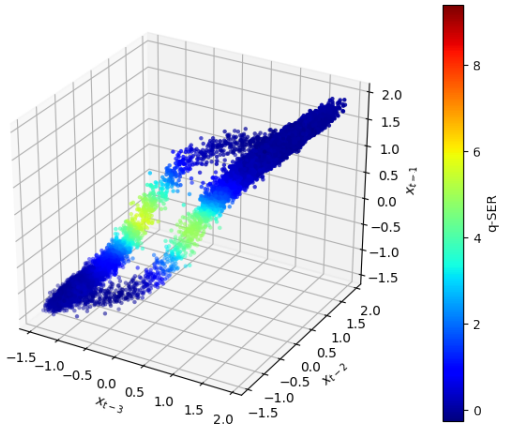
The addition of q into the parameter space gives us a new tool to use to track the dynamical behavior of the nanopore system. We began by exploring the estimated q step specific entropy rate applied to the nanopore simulation data for different values of q. We show

results for $q = 1$ to $q = 20$ for the data set used in analysis in chapter . We note that scaling is different than 0-step specific entropy rate.

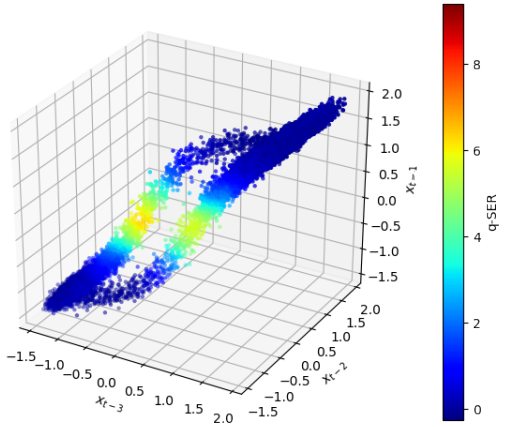
We see that as the value of q increases the q-step future begins to diverge increasingly from the 0-step future. This is visible in the increase in the normalized q-step specific entropy rate in the reconstructed state space trajectories shown below. Additionally, we note that this increase is more concentrated in certain regions of the reconstructed state space. We note that there is an increase in the normalized q-step specific entropy rate both during the transition from the positive to the negative state and from the negative to the positive state. During transitions, we expect the q-step future to look increasingly more different than the immediate 0-step future as q increases, which is consistent with our observations. As q changes, the region of increase in the normalized q-step specific entropy rate also shifts: with increasing q, the normalized q-step specific entropy rate increases earlier during both types of transitions. Lastly, transitions from the negative to the positive state show the greatest enhancement of the normalized q-step specific entropy rate. This indicates that each type of transition event produces distinct information, and furthermore, distinct information at different q values.


q-SER
x_{t-1}
x_{t-2}
x_{t-3}

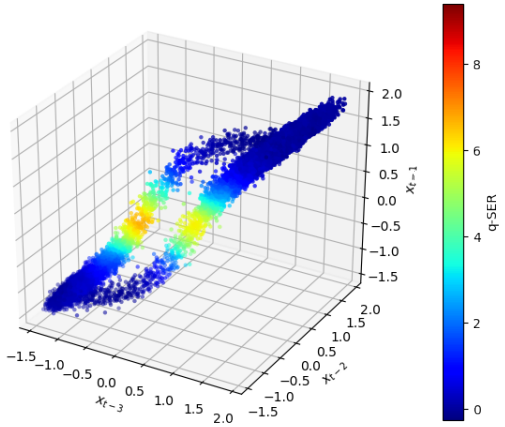
(a)


q-SER
x_{t-1}
x_{t-2}
x_{t-3}

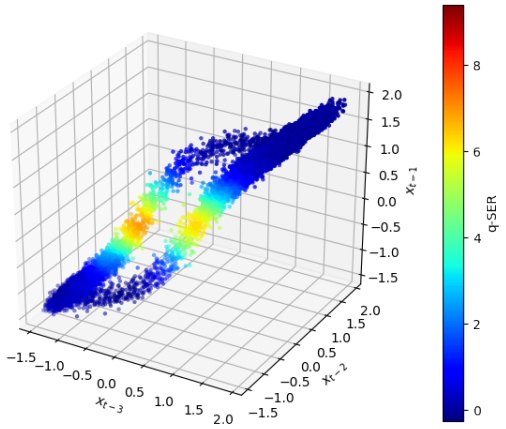
(b)


q-SER
x_{t-1}
x_{t-2}
x_{t-3}

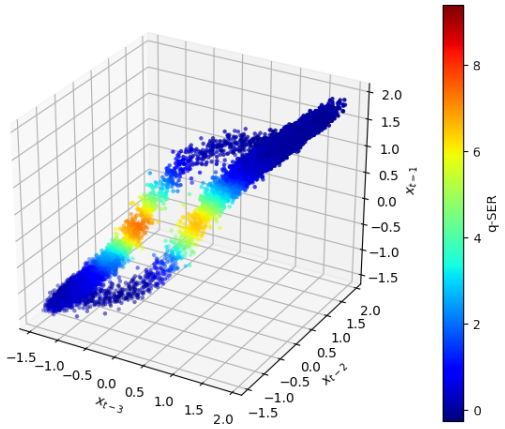
(c)


q-SER
x_{t-1}
x_{t-2}
x_{t-3}

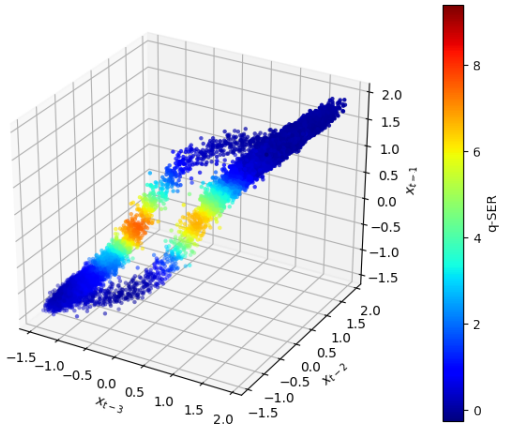
(d)


q-SER
x_{t-1}
x_{t-2}
x_{t-3}

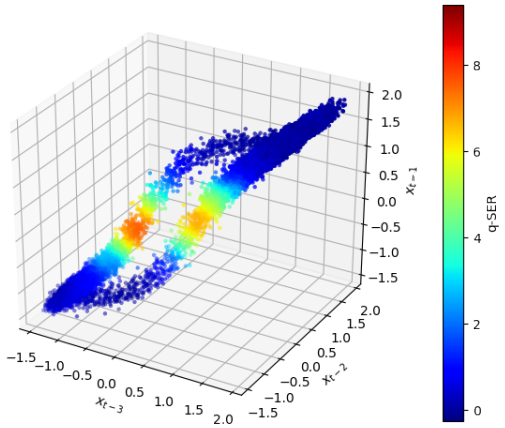
(e)


q-SER
x_{t-1}
x_{t-2}
x_{t-3}

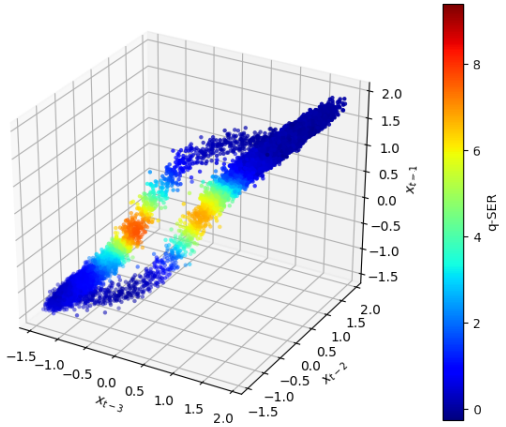
(f)


q-SER
x_{t-1}
x_{t-2}
x_{t-3}

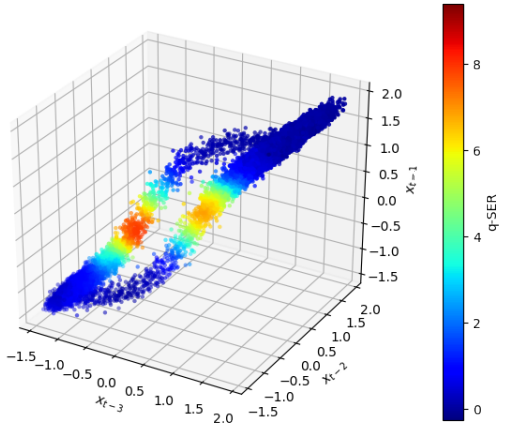
(g)


q-SER
x_{t-1}
x_{t-2}
x_{t-3}

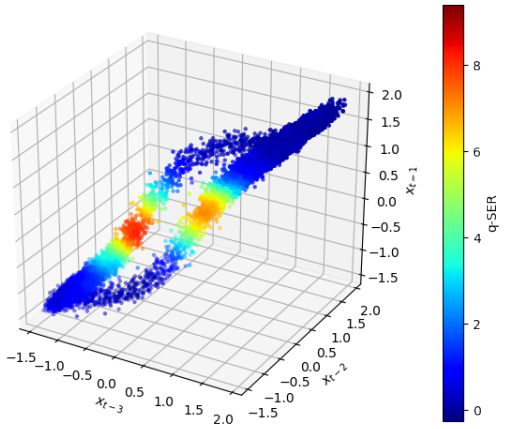
(h)


q-SER
x_{t-1}
x_{t-2}
x_{t-3}

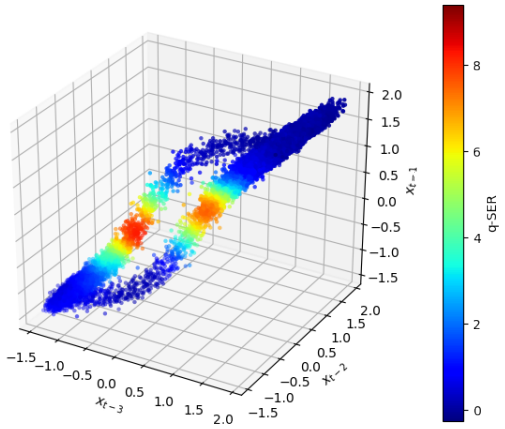
(i)


q-SER
x_{t-1}
x_{t-2}
x_{t-3}

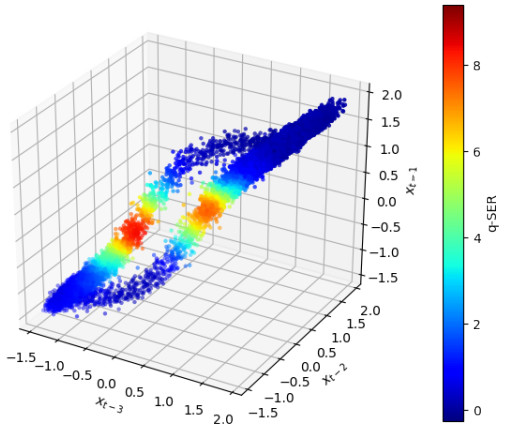
(j)


q-SER
x_{t-1}
x_{t-2}
x_{t-3}

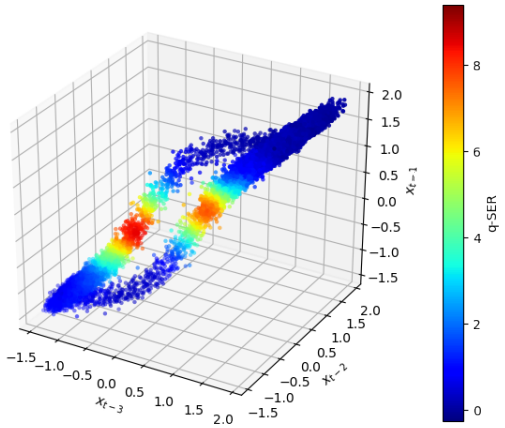
(k)


q-SER
x_{t-1}
x_{t-2}
x_{t-3}

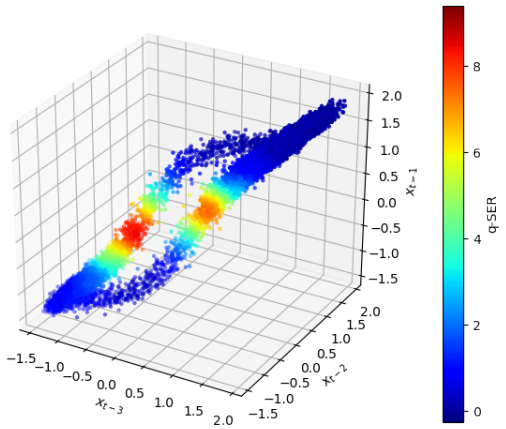
(l)


q-SER
x_{t-1}
x_{t-2}
x_{t-3}

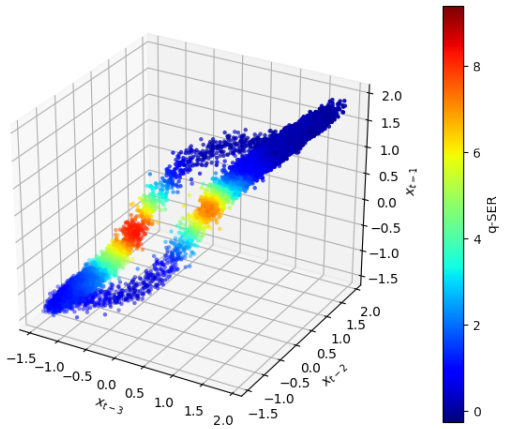
(m)


q-SER
x_{t-1}
x_{t-2}
x_{t-3}

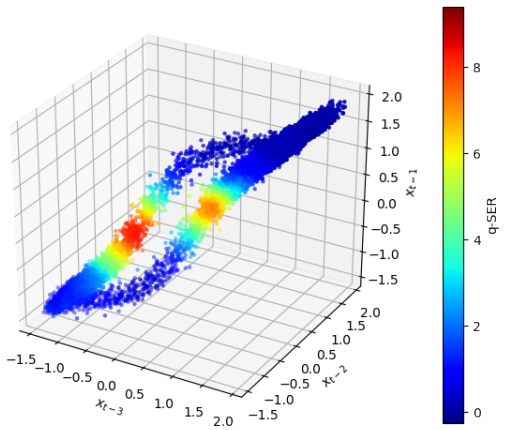
(n)


q-SER
x_{t-1}
x_{t-2}
x_{t-3}

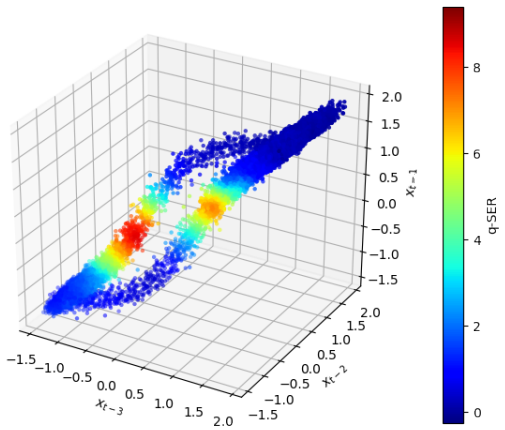
(o)


q-SER
x_{t-1}
x_{t-2}
x_{t-3}

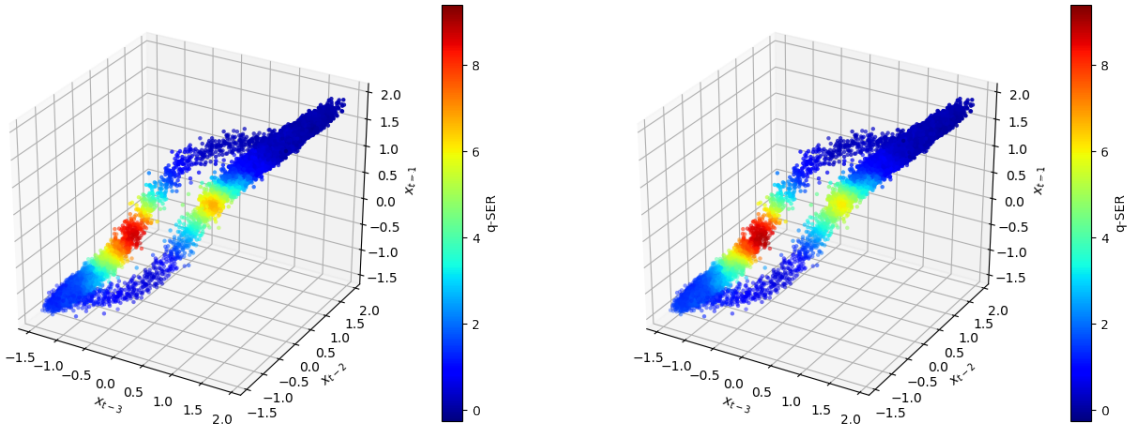
(p)


q-SER
x_{t-1}
x_{t-2}
x_{t-3}

(q)


q-SER
x_{t-1}
x_{t-2}
x_{t-3}

(r)



(s) (t)

Figure 4.2: This figure shows the evolution of the normalized q-step specific entropy rate with changing q. q changes from 1 (a) to 20 (t). The q step future looks increasingly more different than the zero step future with increasing q and the effect is most notable during transitions. We also note that negative to positive conductance state transitions show the greatest enhancement.

4.3 Application of q-step Specific Entropy Rate to Experimental Nanopore Data

In this section we will discuss the application of the normalized normalized q-step specific entropy rate to the experimental data discussed in section 1.4.2. Application of the technique to experimental data comes with some additional considerations and data preparation. We will first discuss how the data are prepared for analysis, then share and discuss some preliminary results.

4.3.1 Considerations in the Application of normalized q-step specific Entropy Rate

One potential challenge of analyzing experimental data is the precision of the measurements. It is extremely important for the purposes of model order selection that there be no repeated data values. Repeated values will be detrimental to the kernel nearest neighbor estimation. The precision in the simulation nanopore data makes repeated data values extremely unlikely; however, digitizers used to collect experimental data can lead to repeated values (which cause numerical instabilities in these analyses). This can be simply addressed using a standard technique of adding uniformly distributed noise to the time series with the a similar magnitude to the precision of the data. For the experimental data analyzed in this work, we add uniform noise between -0.5 and +0.5. Following this processing step, the next thing we need to consider is selection of a downsampling rate.

It should be noted that there is no established procedure for choosing a downsampling rate for these experimental data. Downsampling may help us reduce the contribution from experimental noise in the analysis, and it is therefore worthwhile to explore several reasonable downsampling rates. We explore the analysis for downsampling rates of 2, 4, and 8. We present the results from downsampling by 2 in the next subsection for discussion. The results obtained from downsampling by 4 and 8 can be found in appendix A.

4.3.2 Preliminary normalized q-step specific Entropy Rate Results for Experimental Nanopore Data

We obtained preliminary normalized normalized q-step specific entropy rate results for an empirical nanopore with experimental parameters discussed in chapter and externally biased at a voltage of $-1.00\ V$. We first verify the direction of the transition trajectory in the

reconstructed state space

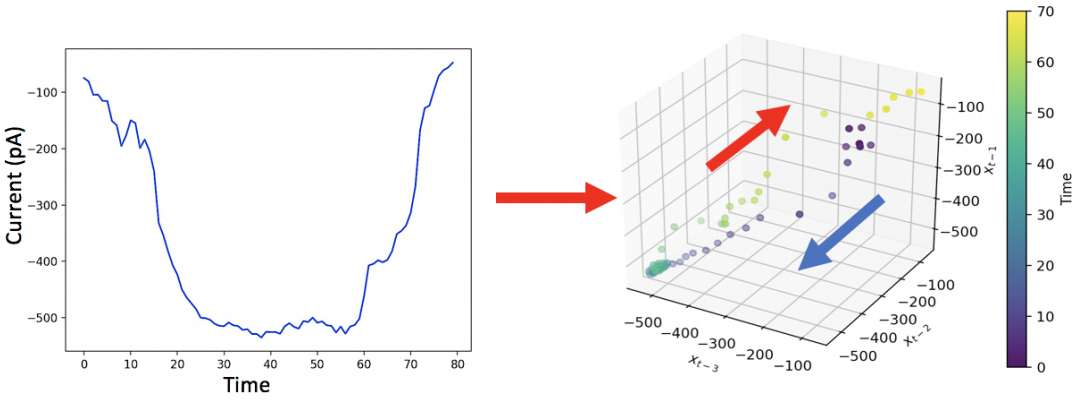
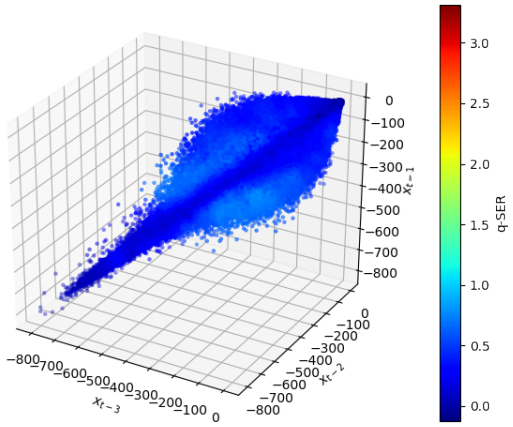
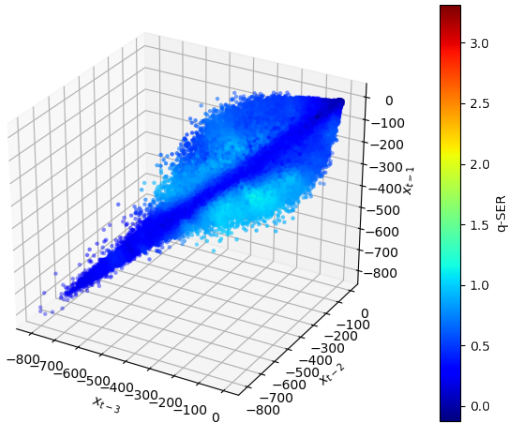


Figure 4.3: This figure shows the reconstructed state space (right) associated with a single transition (left). The color scale represents time. The purpose of this graph is to show the direction of the transition trajectory in reconstructed state space and will serve as a reference for the state space reconstructions shown in the figure below. The sampling rate is 10000 Hz. The data are downsampled by a factor of 2.

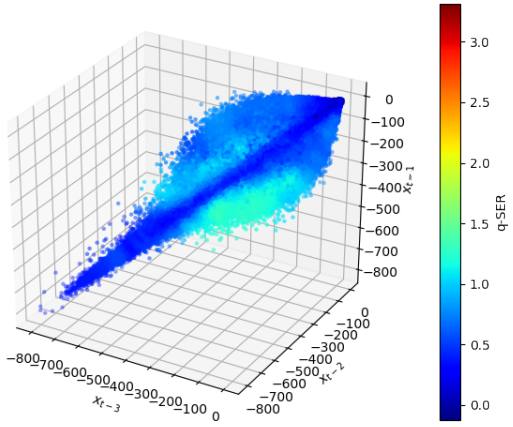
We see that there is also an evolution of the reconstructed state space, with the normalized normalized q-step specific entropy rate increasing in some portions of the space with increasing q. We also see that the transition from higher to lower conductance states is associated with the earliest increases in the normalized normalized q-step specific entropy rate with increasing q. We note that as q increases, we also begin to see enhancement of the normalized normalized q-step specific entropy rate for the most negative conductance states. This nanopore tends to spend relatively less time in the most negative conductance states than it does in higher conductance states, and therefore a higher normalized normalized q-step specific entropy rate with increasing q makes sense.


q-SER
x_{t-3}
x_{t-2}
x_{t-1}

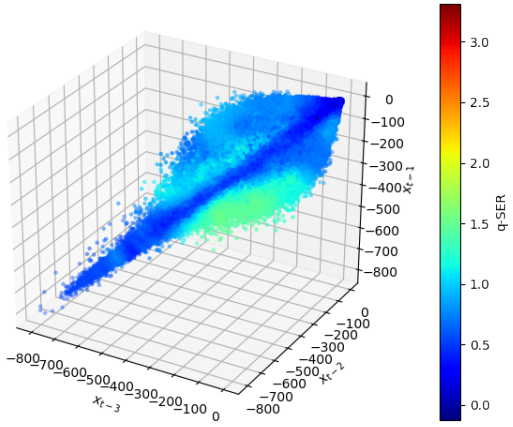
(a)


q-SER
x_{t-3}
x_{t-2}
x_{t-1}

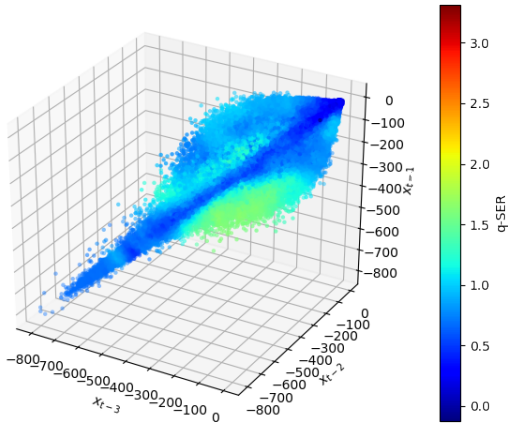
(b)


q-SER
x_{t-3}
x_{t-2}
x_{t-1}

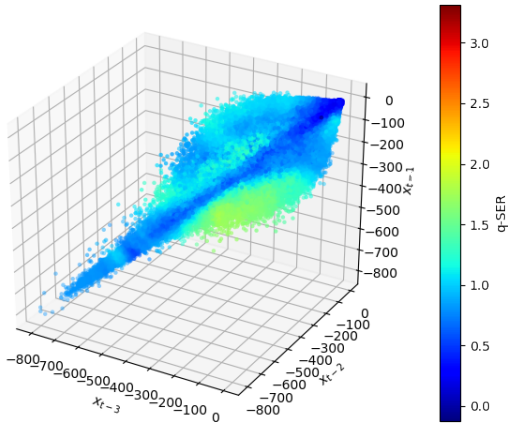
(c)


q-SER
x_{t-3}
x_{t-2}
x_{t-1}

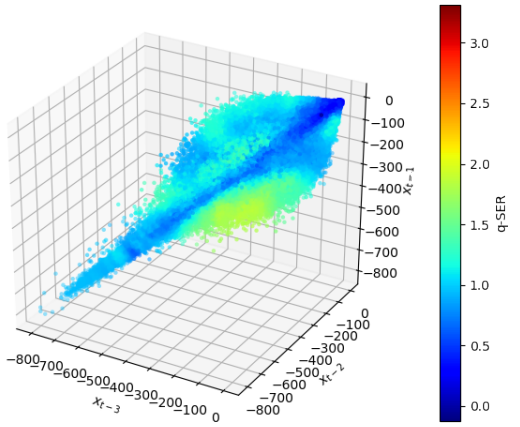
(d)


q-SER
x_{t-3}
x_{t-2}
x_{t-1}

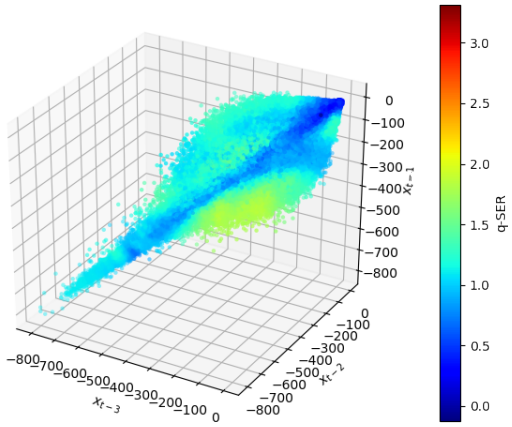
(e)


q-SER
x_{t-3}
x_{t-2}
x_{t-1}

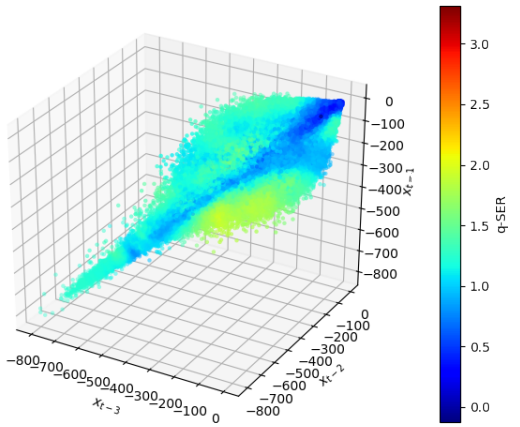
(f)


q-SER
x_{t-3}
x_{t-2}
x_{t-1}

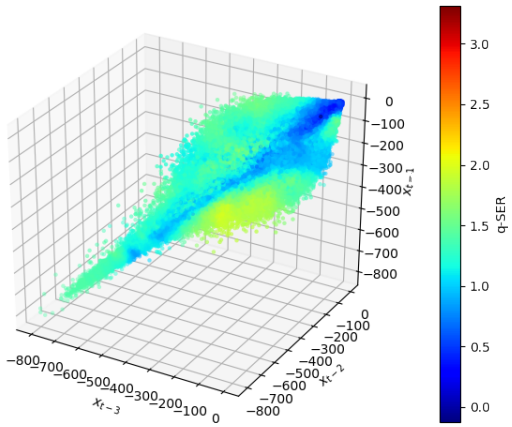
(g)


q-SER
x_{t-3}
x_{t-2}
x_{t-1}

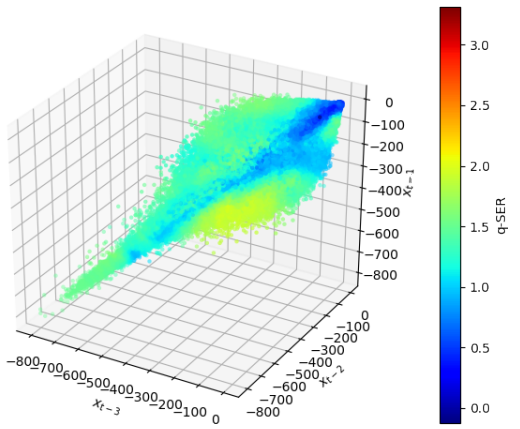
(h)


q-SER
x_{t-3}
x_{t-2}
x_{t-1}

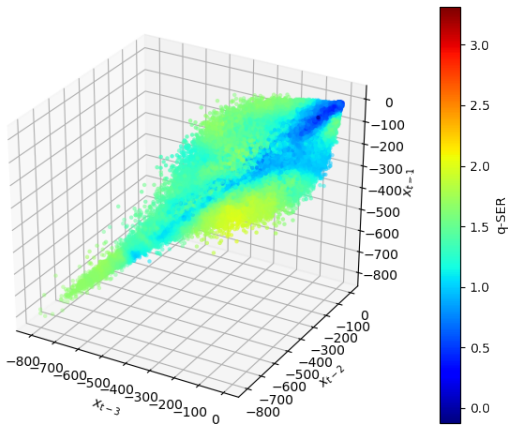
(i)


q-SER
x_{t-3}
x_{t-2}
x_{t-1}

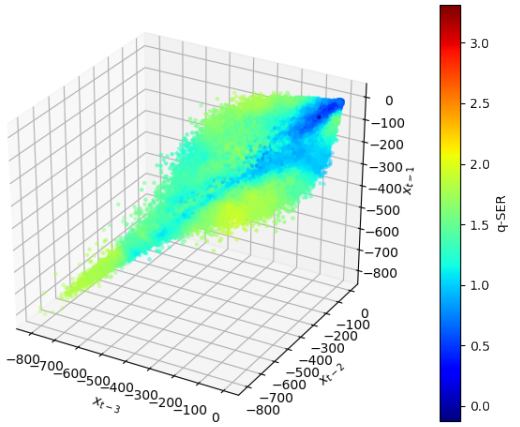
(j)


q-SER
x_{t-3}
x_{t-2}
x_{t-1}

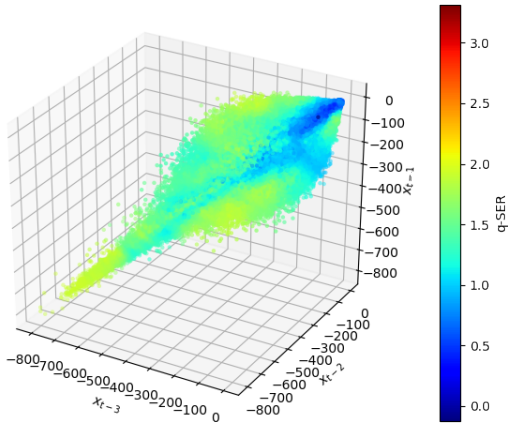
(k)


q-SER
x_{t-3}
x_{t-2}
x_{t-1}

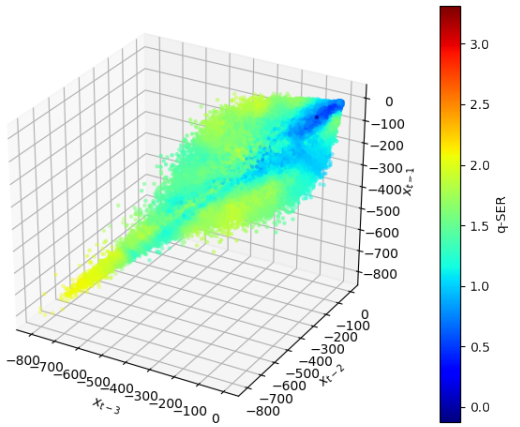
(l)


q-SER
x_{t-3}
x_{t-2}
x_{t-1}

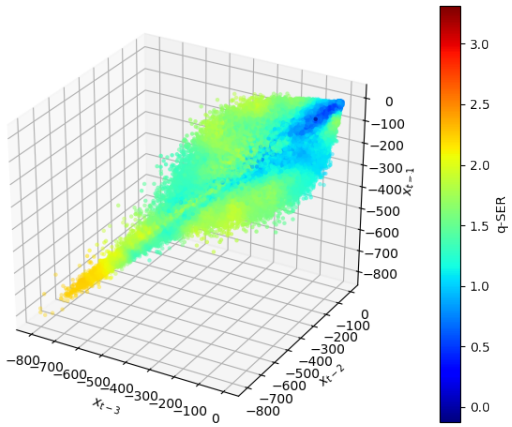
(m)


q-SER
x_{t-3}
x_{t-2}
x_{t-1}

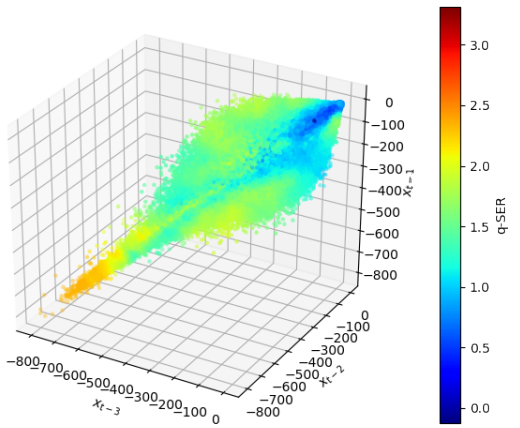
(n)


q-SER
x_{t-3}
x_{t-2}
x_{t-1}

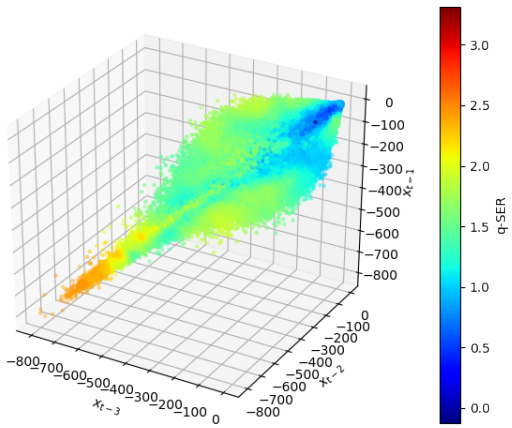
(o)


q-SER
x_{t-3}
x_{t-2}
x_{t-1}

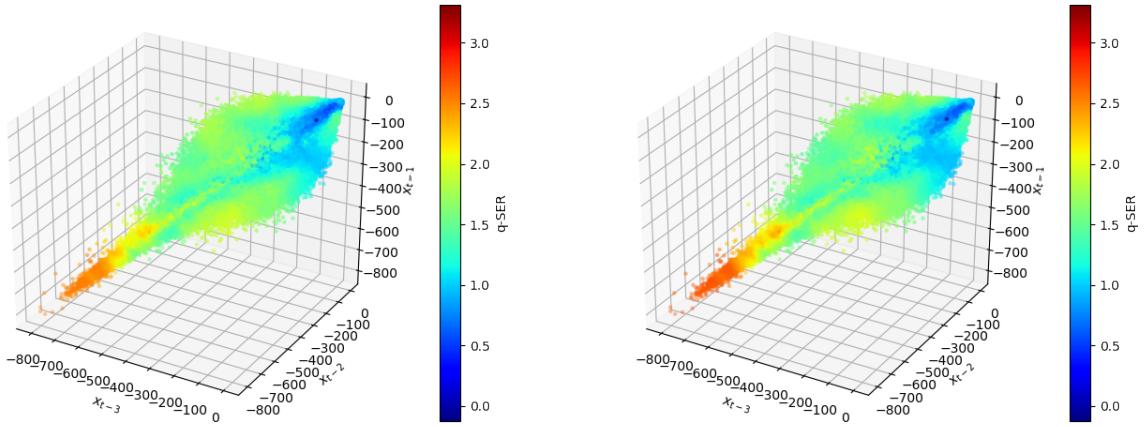
(p)


q-SER
x_{t-3}
x_{t-2}
x_{t-1}

(q)


q-SER
x_{t-3}
x_{t-2}
x_{t-1}

(r)



(s) (t)

Figure 4.4: This figure shows the evolution of the reconstructed state space for normalized q-step specific entropy rate for experimental nanopore data with changing q. Plots (a) through (t) are associated with $q = 1$ through $q = 20$. The transitions from high to low conductance states are associated with the earliest increase in the normalized q-step specific entropy rate with increasing q. With increasing q we also see increases in the normalized q-step specific entropy rate for the lowest conductance states.

These results may lead to interesting and more broadly applicable future exploration. We discuss the potential next steps for this research in chapter 6.

Chapter 5

Detection of Nonlinear Structure in Nanopore Interevent Intervals

5.1 Design and Aim of the Study

To this point we have discussed analysis methods that involved computations on the raw simulation or (minimally pre-processed) experimental nanopore data. It is possible that there is knowledge we can gain from looking at the interevent intervals (i.e. the time between successive high to low or low to high conductance state transitions). Mechanistically, we know that it is likely that each fluctuation may set the conditions for the next (recall the proposed formation/dissolution of nanoprecipitates inside of the pores in [66, 37]), and that there may thus be evidence of a non-linear structure to be found in these interevent intervals. In this chapter we discuss the application of surrogate data analysis as a method for identifying non-linear structure to the interevent intervals. In this section we will discuss the process of data smoothing, identifying interevent intervals, surrogate generation, and hypothesis testing. In the subsequent sections, we will discuss application of these methods to both simulation and

experimental nanopore data and discuss our preliminary results.

5.1.1 Methods Overview

The following is a conceptual overview of the methods we employed to search for nonlinearity in the interevent interval (II) data. Specific details for both simulation nanopore data and experimental nanopore data will be discussed in subsequent sections.

We convert the time series representing the nanopore current to a time series of interevent intervals. In this preliminary work, we chose to identify events as transitions from high to low conductance states. The new data represents the interval between successive events. We have discussed the possibility that there may be an interdependence of events in the experimental nanopore data, but we may not see this in simulation data. If we look at the model system in chapter 1, we note that the transition behavior is largely dominated by the unobserved variable Y, a noise process with drift. We therefore might hypothesize that the interevent interval sequence for simulation data generated from that model will be memoryless. A natural null hypothesis to test would be that the interevent intervals are generated by a process that is some monotonic nonlinear transformation of colored noise (such as cubing colored noise). Rejection of this null hypothesis can be interpreted as evidence of a nonlinearity in the interevent intervals.

We test this hypothesis using a surrogate data technique described in [73, 41]. This method entails using the original interevent interval sequence to create surrogates as though they arise from a monotonic nonlinear transformation of colored noise. The surrogate generation process has the following steps:

1	Initialize surrogate to random shuffle of II sequence
2	Compute Fourier Transform
3	Construct Fourier Series with same Fourier amplitudes as original II sequence and phases of shuffled II sequence
4	Compute inverse Fourier Transform of Series constructed in step 3 to obtain $\{X^T\}$
5	Take original II sequence and rank order entries to match rank order of $\{X^T\}$
6	Take new time series from step 5 it becomes the new input for step 2
7	Iterate steps 2-6 100 times.

Table 5.1: Steps to constructing a single surrogate data set. The algorithm used creates amplitude adjusted Fourier transform (AAFT) surrogates.

The initial 4 steps of the process can be thought of as a constrained colored noise process. Gaussian colored noise processes preserve the power spectral density but do not necessarily preserve the time series amplitudes (and thus will not preserve the marginal probability density). Our interevent interval sequences are not normally distributed, and if we were to generate surrogates through a Gaussian colored noise process we would be engineering a hypothesis test that was sure to reject the null each time. It is therefore important that we constrain the surrogate generation process to also preserve the time series amplitudes. Step 5 can be thought of as the application of a nonlinear monotonic transformation. Thus the surrogates have been constructed in such a way as to test the proposed hypothesis.

We then apply our target metric, the estimator for total entropy rate, to the original interevent interval data and its surrogates, compare the results, and compute a P value. Specifically, the P value is computed by determining what fraction of the surrogates have

a lower estimated total entropy rate than the original interevent interval sequence. If transitions influence future transitions and there is an associated nonlinear structure to the interevent interval sequence we would expect to see the original interevent sequence have a lower estimated total entropy rate (associated with more certainty) than its surrogates. In other words, the process of surrogate generation would be expected to break this nonlinear structure, if it exists, resulting in higher estimated total entropy rates for the surrogates. We use the standard $P < 0.05$ as the criteria for rejecting the null hypothesis.

We also need to consider the model order as we are computing the estimated total entropy rate. Model order selection is conducted in accordance with the procedures discussed in chapter 2 of this work. For a memoryless set of interevent intervals where there is no connection between each event and future events, we would expect the optimal model order to be 0. We therefore expect to see an optimal model order of 0 for the interevent interval sequence (as distinct from the raw time series represented by the variable X in the model, which certainly has memory) for the simulation data (and its surrogates). We do not necessarily expect to see any particular optimal model order selected for the interevent interval sequence or its surrogates for the experimental nanopore data.

5.2 Application to Simulation Nanopore Data

5.2.1 Computing Interevent Intervals

The first step is to take our raw simulation data and convert it into an interevent interval sequence. For the simulation nanopore data, we accomplish this by observing that all high to low conductance state transitions are associated with crossing the line of zero current. We identify the points on either side of that crossing and fit a line between them to find the approximate time of this crossing. The following figure is an example of transitions identified

in simulation nanopore data using this method.

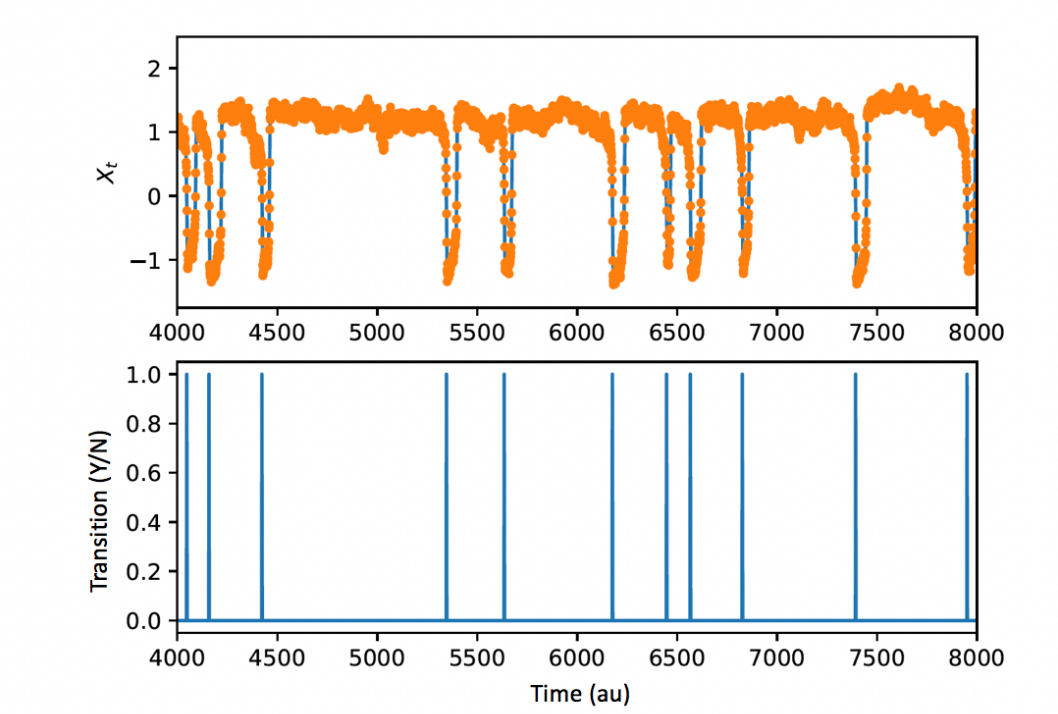


Figure 5.1: The top panel shows nanopore simulation data with the transitions marked by the blue delta functions in the bottom panel. Interevent interval sequences are constructed from the difference in time of the each successive event.

The interevent intervals are computed by taking the time differences between the times of successive events, followed by a log transform.

We generated 500 surrogates, computed the estimated total entropy rates for the original interevent interval sequence and the surrogates, and computed the associated P value. We present the results for a realization of the simulation data.

5.2.2 Results and Discussion

We show a histogram of the estimated total entropy rates for the surrogate data, with the estimated total entropy rate value of the original data labeled. The optimal model order

selected for the original interevent interval sequence and all surrogates is 0. The P value from the hypothesis testing (which should not be confused with p, the model order) is $P = 0.918$. We cannot reject the null hypothesis that the interevent intervals for the simulation nanopore data are generated by a transformation of a white noise process. Using this method, we do not detect evidence of a nonlinear structure in this interevent interval sequence.

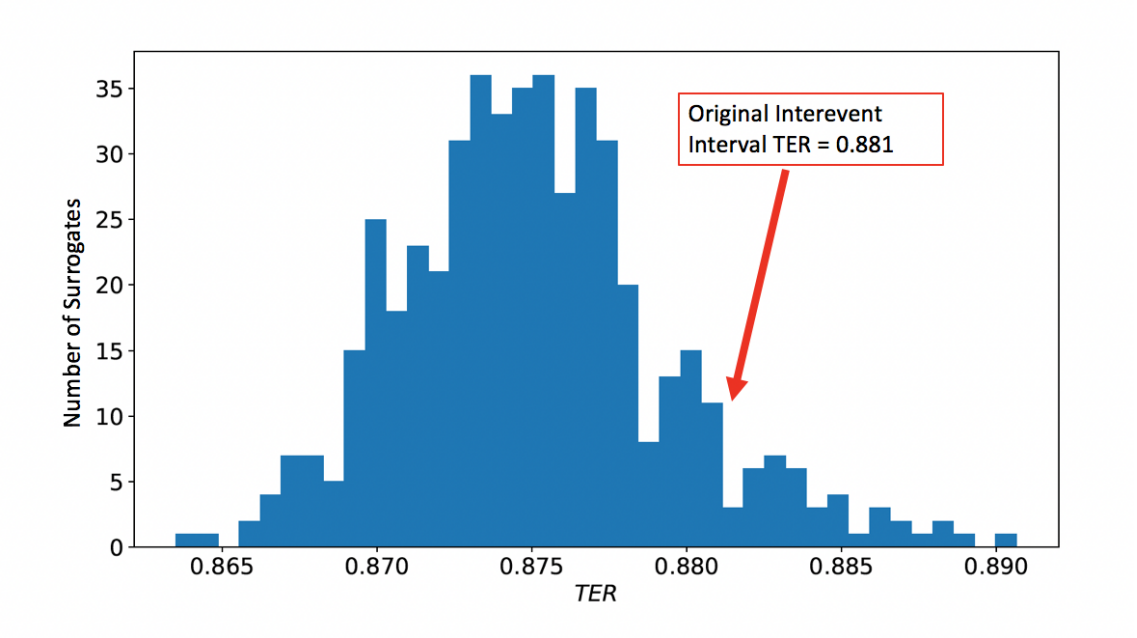


Figure 5.2: This figure is a histogram of the estimated total entropy rates computed for interevent interval surrogate data simulated from the nanopore model discussed in chapter 2. The estimated total entropy rate for the original interevent interval sequence is also indicated by the red arrow. The hypothesis test yields $P = 0.918$. We cannot reject the null hypothesis.

5.3 Application to Experimental Nanopore Data

5.3.1 Preparation of the Data

The experimental nanopore data must be smoothed to avoid double counting of transitions due to noise. We selected a Butterworth filter, a type of low pass filter, to attenuate high

frequency components of the signal [75]. The transition activity of interest occurs at relatively low frequencies, and therefore attenuation of higher frequencies would not be expected to impact results. We choose the parameters for the filter by looking at the power spectral density of the data as a function of frequency. In the figure below we show the semilog-scaled power spectral density for the experimental nanopore data at different bias voltages.

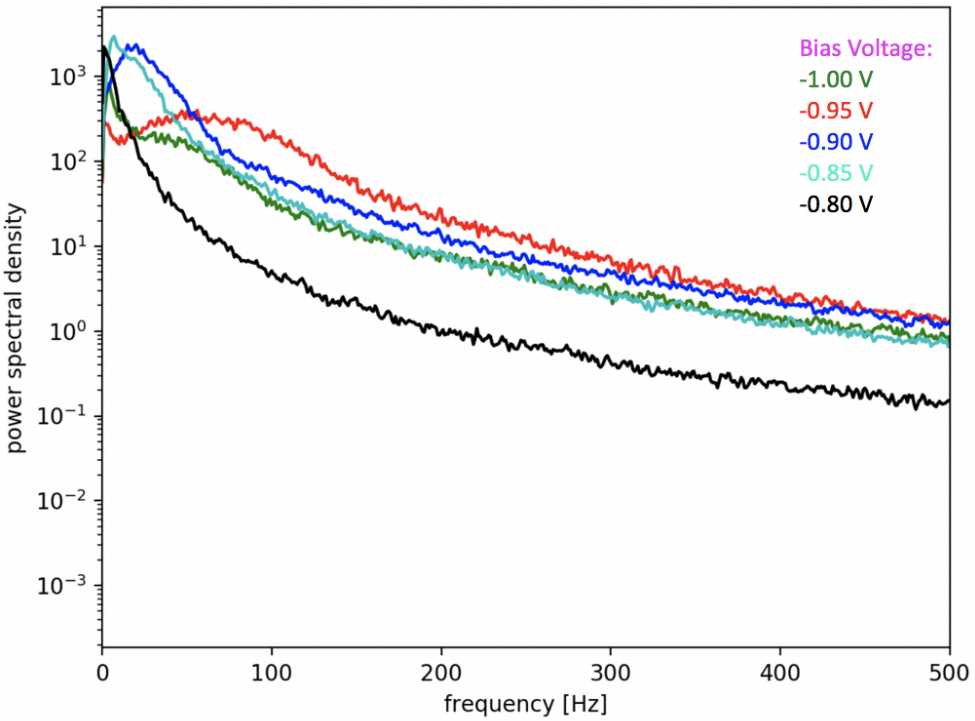
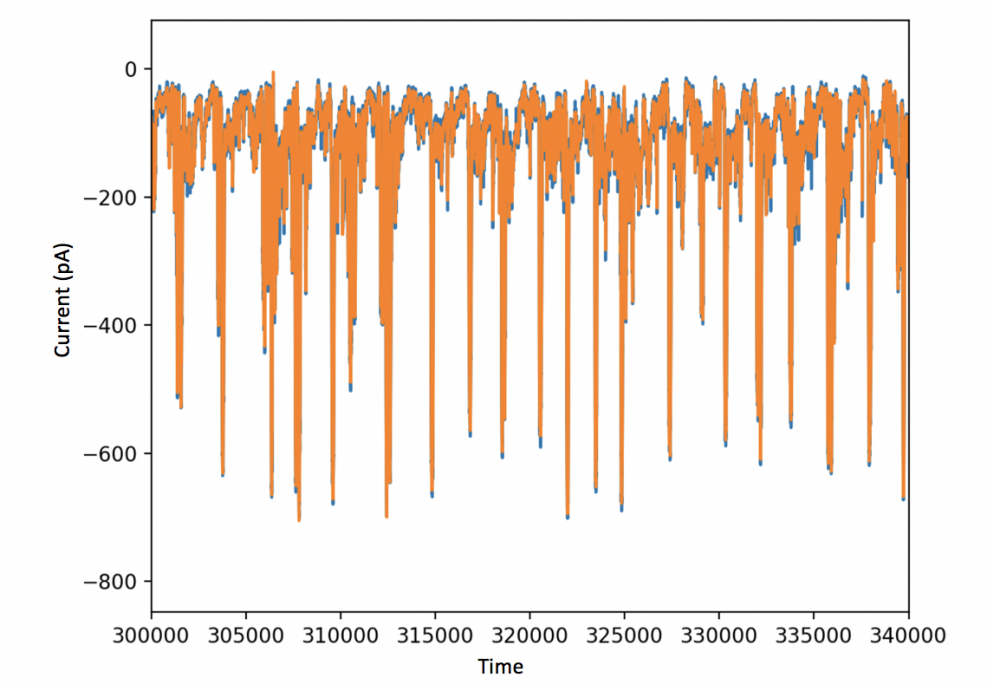
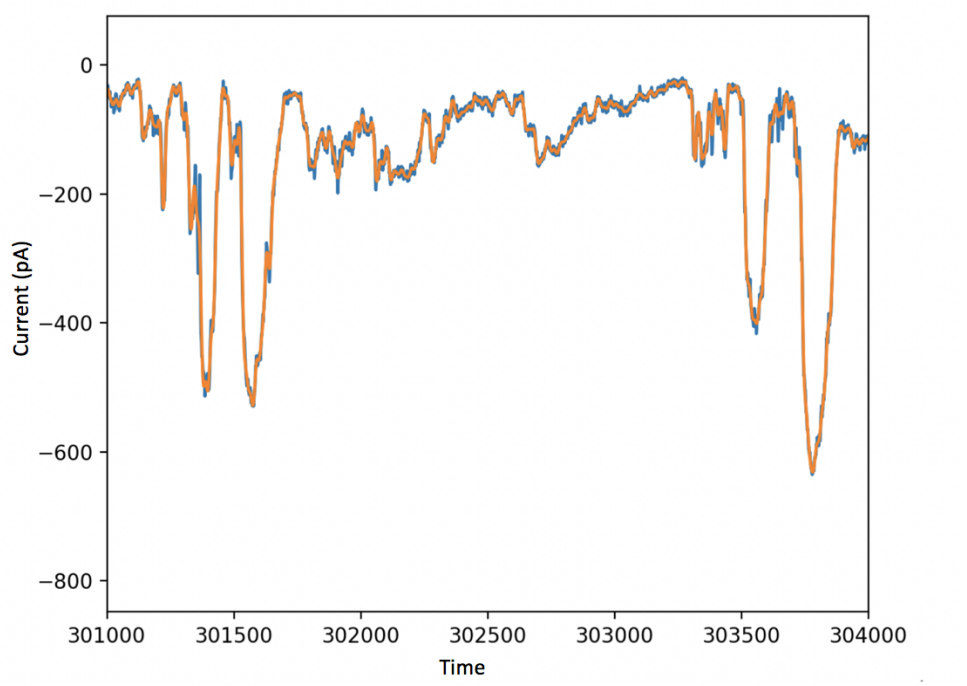


Figure 5.3: We show the power spectral density as a function of frequency for experimental nanopore data. The plot helps to determine an appropriate frequency for the Butterworth filter.

We attenuate all frequencies above $0.1 f_{Nyquist}$, where $f_{Nyquist}$ is the Nyquist frequency, or half the sampling frequency. The sampling frequency for these data is $10,000$ Hz, the Nyquist frequency is $5,000$ Hz, and the frequency cutoff for the Butterworth filter is 500 Hz. A plot shown at two different scales shows a snapshot of the filtered current time series atop the raw current time series.



(a)

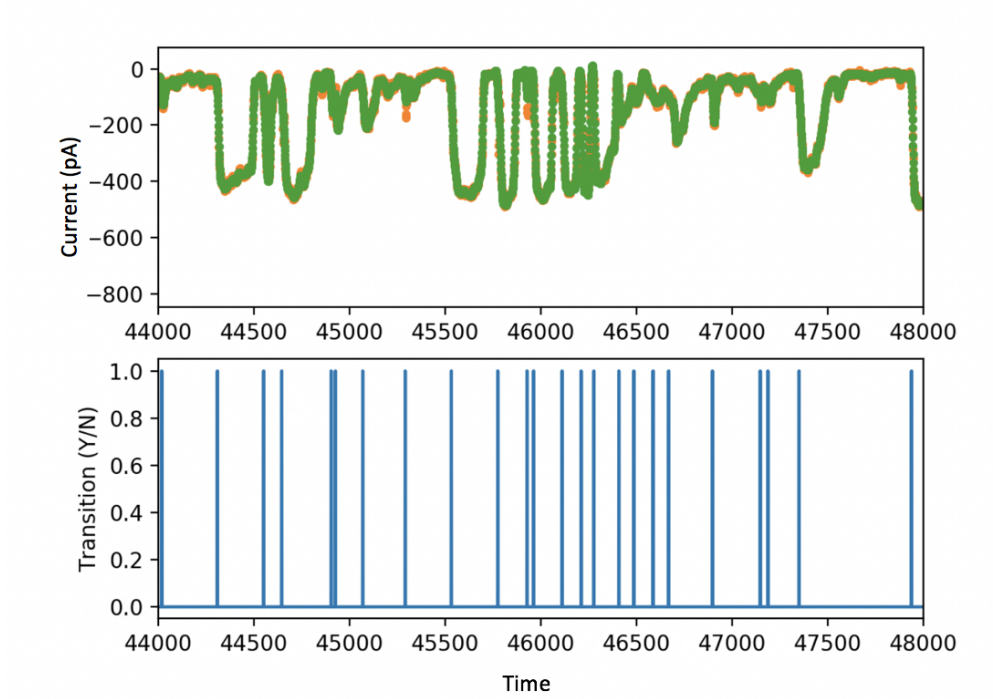


(b)

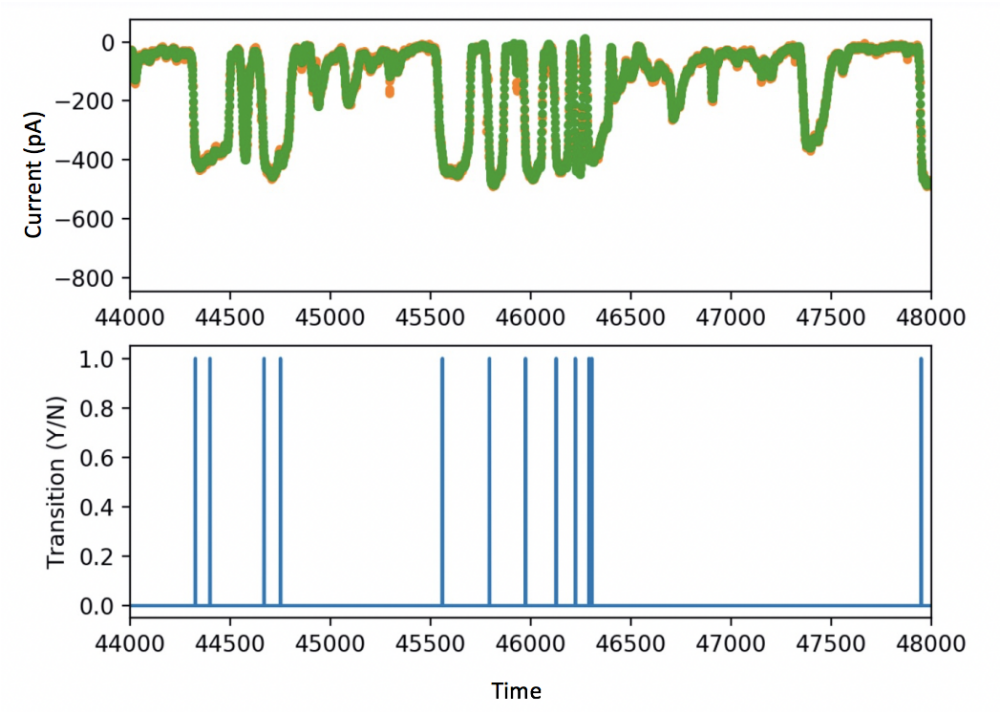
Figure 5.4: This figure depicts an example of the results of smoothing experimental nanopore data with a Butterworth low pass filter. The filter was selected to attenuate frequencies above 500 Hz. Raw data shown in blue, filtered data shown in orange. The filter smoothes high frequency noise, while leaving the lower frequency features of interest unchanged. Sampling frequency is 10000 Hz. Data are shown at two different time scales.

5.3.2 Computing Interevent Intervals

We compute interevent intervals by determining an appropriate threshold amplitude for events of interest, determining the time associated with the threshold crossing (also by linear interpolation between the two points on either side of the threshold), and computing the differences between successive events. We need to take more care in determining the threshold for each experimental data set than is needed for the simulation data, as the experimental data are not stationary. If we set the threshold too low or too high, we risk missing some events or double counting others due to wiggles in the data. The figure below shows examples of these concerns.



(a)



(b)

Figure 5.5: This snapshot highlights some potential problems encountered when choosing a current threshold for events of interest in the experimental nanopore data. Specifically, the two concerns are double counting of events due to small wiggles in the data about the threshold current value and missing events of interest by setting the threshold current too low. Filter data shown in green, raw data shown in orange.

Our goal is to find an optimal compromise that avoids double counting events and results in a minimal number of missed events, acknowledging that it is not possible to view each potential event by eye but is feasible to view multiple sections of the data to ensure this balance is adequately fulfilled. For the data above, this balance is best struck with a current threshold of -250 pA. Thresholds of between -250 pA and -400 pA were used for all interevent interval data sets analyzed in this work.

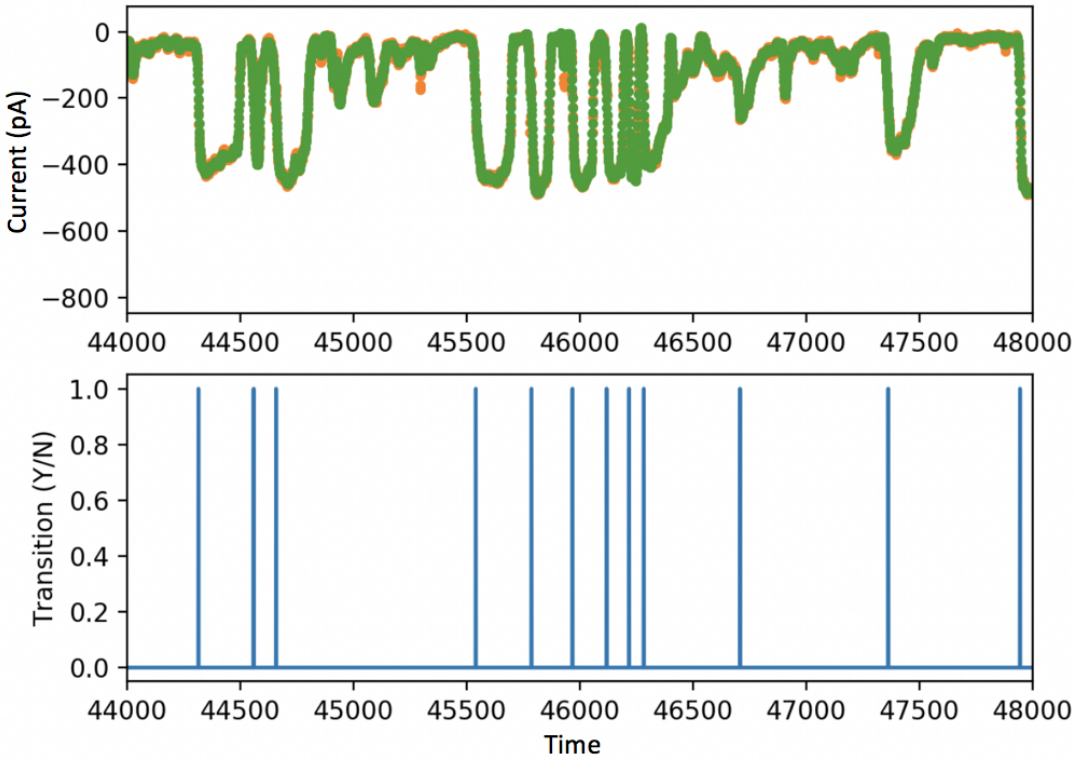


Figure 5.6: This snapshot shows the same data as are shown above with a threshold current chosen to strike the best balance between counting all events that might be important to understanding the dynamics of the interevent intervals and not double counting any events. We aim to strike this balance for each set of experimental nanopore data.

Once we have determined a threshold for events of interest we compute the interevent intervals, generate surrogates, compute the estimator for total entropy rate associated with each surrogate and the original interevent interval sequence, and compute the P value for the hypothesis test. We identified that 500 surrogates was more than necessary and for

computational efficiency 200 surrogates were computed for each experimental nanopore data set.

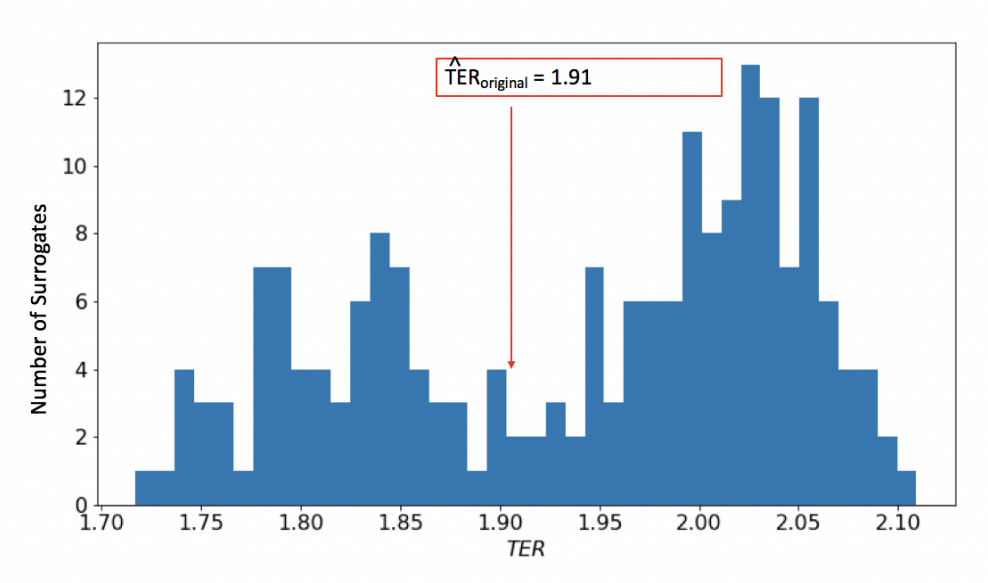
5.3.3 Results and Discussion

In the table below we state the parameters selected for hypothesis testing of five experimental nanopore data sets and the P value obtained from hypothesis testing with 200 surrogates. We find evidence of nonlinear structure in the interevent intervals associated with the highest two magnitudes of external bias voltage, $-0.95\ V$ and $-1.00\ V$.

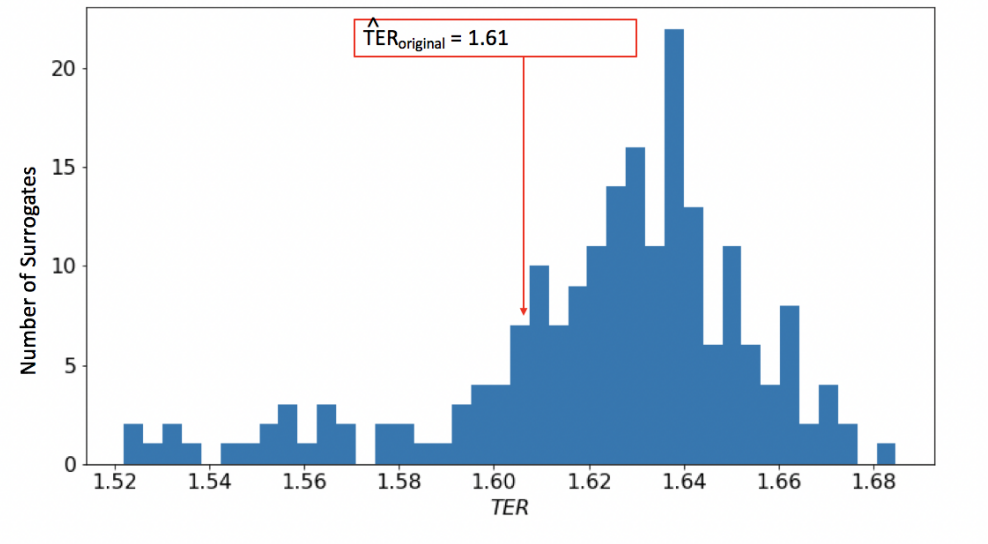
$Bias\ (V)$	$Threshold\ (pA)$	$P\ Value$
-0.80	-350	0.37
-0.85	-400	0.20
-0.90	-400	0.26
-0.95	-300	0.005
-1.00	-250	0.005

Table 5.2: The table shows the results of hypothesis testing for experimental nanopore interevent intervals. We reject the null hypothesis that the data come from a constrained colored noise process with an applied monotonic nonlinear transformation for the two highest magnitude externally applied bias voltage, indicating evidence of a potentially interesting nonlinear structure.

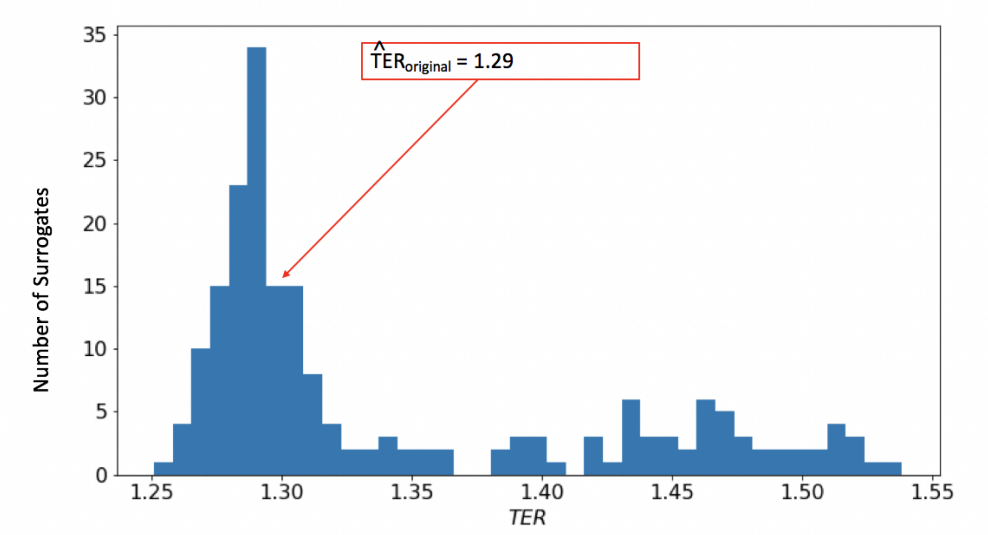
We also present the histograms for the estimated total entropy rates of the surrogates for each entry in table 5.2. We label the value of the total entropy rate. In the cases where the null hypothesis is rejected we see that the estimated total entropy rate for the original interevent interval sequence lies outside of the distribution of estimated total entropy rates for its surrogates, which is evidence the nonlinear structure of the original does not come about from a monotonic nonlinear transformation of colored noise (which would be an uninteresting source of nonlinear structure).



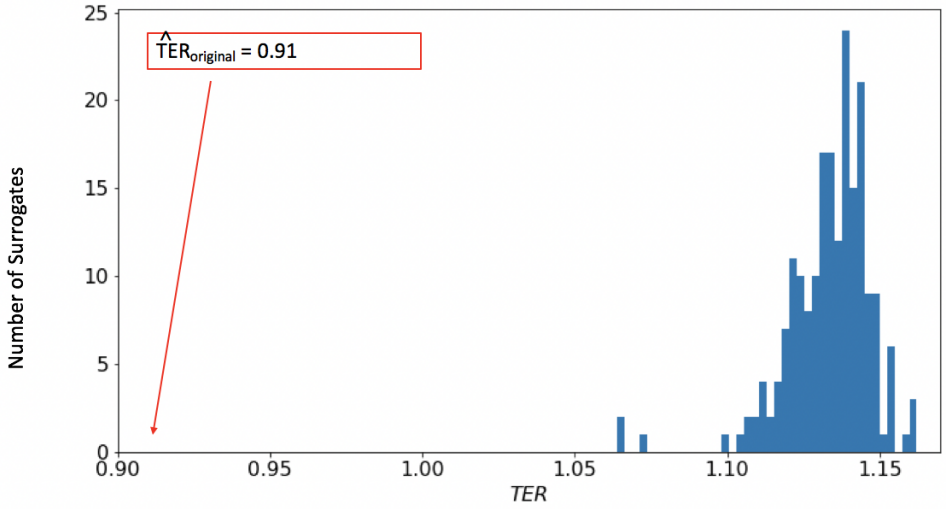
(a) $V_{bias} = -0.80\ V$



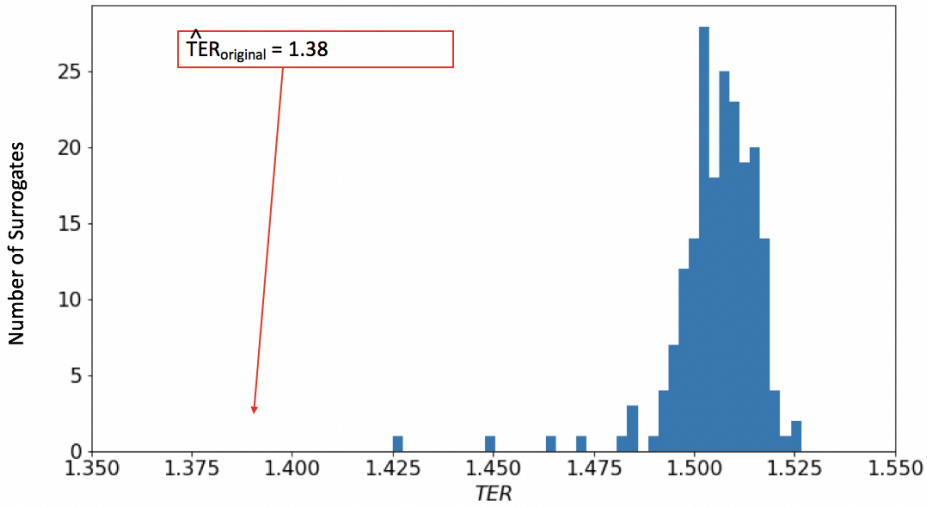
(b) $V_{bias} = -0.85\ V$



(c) $V_{bias} = -0.90\ V$



(d) $V_{bias} = -0.95\ V$



(e) $V_{bias} = -1.00\ V$

Figure 5.7: This figure shows histograms for the estimated total entropy rates associated with the interevent interval sequence surrogates. The estimated total entropy rate of the original entropy rate is labeled and shows whether it lies within the distribution of interevent intervals for the surrogates. The null hypothesis that the original interevent interval sequence comes from a process like the one used to construct the surrogates is rejected for the two highest magnitude applied bias voltage. Further experimental investigation should be conducted to ensure repeatability and a wider range of applied bias voltages should be explored. Preliminarily, these results are evidence that there may be an interesting nonlinear structure in the interevent intervals in the experimental data that may be physically linked with the proposed nanoprecipitation/dissolution mechanism in [66, 37].

In some cases where the null hypothesis was not rejected, additional hypothesis testing was conducted on interevent interval sequences constructed using other current threshold values. This is a check done to ensure the threshold we found ideal did not result in overlooking an otherwise present nonlinear structure. We present results for bias voltages -0.85 and -0.90 V at a threshold of -250 pA in appendix B.

The detection of evidence for nonlinear structure to the interevent interval sequences at some bias voltages and not others leads us to wonder if this is an emergent, bias-dependent property. As the results presented for this portion of our research are preliminary results, we caution against overinterpretation and instead suggest future studies to hopefully determine

whether this structure is linked to the formation/dissolution of nanoprecipitates. There is still significant parameter space to explore with respect to stronger bias voltages and different ionic solutions and concentrations. We discuss some possible next experimental steps in chapter 6.

Chapter 6

Future Directions and Conclusions

Both the normalized q-step specific entropy rate and interevent interval analyses have yielded promising early results and have set the stage for future research. In this chapter we will explore some possible future directions.

6.1 Future Directions for Normalized q-step Specific Entropy Rate

The normalized q-step specific entropy rate offers the ability to compare the future q steps ahead to the immediate future and to determine how different those two futures look. The next logical step would be to explore its utility to predict transitions. One possible route to explore would be to collect a set of data and use it as a training set to create the reconstructed state space at different values of q. A subsequent set of data may then be analyzed step by step to determine whether a particular block X_{t-p}^{t-1} in question is nearby a point in the training set's reconstructed state space associated with an impending transition.

6.2 Future Directions for Interevent Interval Analysis

The hypothesis testing conducted on the interevent interval sequences for the experimental nanopore data revealed evidence of a potentially interesting nonlinear structure that may be voltage dependent. The natural next steps to take would be to collect multiple new, longer data sets (with a greater number of events) and to extend the upper range of bias voltages to determine whether evidence of nonlinear structure is seen consistently and only for certain experimental parameters. If we can characterize the range of experimental parameters for which this effect is seen, we may take future steps to use these mathematical results to help illuminate the physical behavior of the pores.

Additionally, the difference between the interevent intervals for the simulation and experimental nanopore data suggest that there may be a need to revise the nanopore model to reflect the information dynamics of the experimental system. A future student focused on mathematical modeling may find this project an exciting endeavor.

6.3 Computational Suggestions for Future Work

We suggest future students interested in working on any component of this project apply to use a supercomputer for data analysis. Use of a supercomputer would substantially reduce the amount of time necessary to compute specific entropy rates, normalized q-step specific entropy rates, and total entropy rates. It would therefore allow for the expeditious exploration of a wide experimental parameter space.

6.4 Conclusions

As we conclude this writing it is helpful to summarize what we have learned and the utility of what we have learned. In our initial simulation studies of the application of local and specific entropy rates to nanopore simulation data have shown us that we can track transition events in nanopores and that the changes in local, specific, and q-step specific entropy rates during the transition processes reveals underlying changes in the information dynamics associated with those events. This is in contrast to systems where transitions may occur in the absence of underlying information changes, such as fully deterministic systems or systems where all transitions are alike. Although we can detect the transition events by simple visual inspection of the time series, we cannot determine changes to information status through visual inspection. Understanding that there are fundamental information-linked changes surrounding and during these transitions/current fluctuations is a new development and may help us explore new methods of characterizing transition behavior in dynamical systems.

We have also obtained preliminary results that point to a possibly interesting non-linear structure present in experimental nanopore interevent interval sequences at certain experimental parameters. It is possible that this non-linear structure may be linked to the proposed formation/dissolution of ionic nanoprecipitates inside of the nanopore that temporarily block/allow passage of current in the pores. These findings motivate future experimental study of nanopores to fully understand the phenomenon and revise mathematical models to include its contribution. We hope that a better understanding of this process mathematically will lead to advances in our understanding of these nanoscale objects and potentially nanometer scale physiological processes.

Bibliography

[1] S. Aminikhanghahi and D. J. Cook. A survey of methods for time series change point detection. *Knowledge and information systems*, 51(2):339–367, 2017.

[2] G. Benettin, L. Galgani, A. Giorgilli, and J.-M. Strelcyn. Lyapunov characteristic exponents for smooth dynamical systems and for hamiltonian systems; a method for computing all of them. part 1: Theory. *Meccanica*, 15(1):9–20, Mar 1980.

[3] G. Biau and L. Devroye. *Lectures on the nearest neighbor method*. Springer, 2015.

[4] P. Bloomfield. *Fourier analysis of time series: an introduction*. John Wiley & Sons, 2004.

[5] S. F. Buchsbaum, N. Mitchell, H. Martin, M. Wiggin, A. Marziali, P. V. Coveney, Z. Siwy, and S. Howorka. Disentangling steric and electrostatic factors in nanoscale transport through confined space. *Nano Letters*, 13(8):3890–3896, 2013. PMID: 23819625.

[6] S. F. Buchsbaum, G. Nguyen, S. Howorka, and Z. S. Siwy. Dna-modified polymer pores allow ph- and voltage-gated control of channel flux. *Journal of the American Chemical Society*, 136(28):9902–9905, 2014. PMID: 24992159.

[7] S. Caires and J. A. Ferreira. On the non-parametric prediction of conditionally stationary sequences. *Statistical inference for stochastic processes*, 8(2):151–184, 2005.

[8] M. L. Cartwright and J. E. Littlewood. On non-linear differential equations of the second order: I. the equation y¨- k (1-y2) y?+ y= bλk cos (λl+ α), k large. *Journal of the London Mathematical Society*, 1(3):180–189, 1945.

[9] R. Castro and T. Sauer. Correlation dimension of attractors through interspike intervals. *Phys. Rev. E*, 55:287–290, Jan 1997.

[10] R. Castro and T. Sauer. Reconstructing chaotic dynamics through spike filters. *Phys. Rev. E*, 59:2911–2917, Mar 1999.

[11] F. S. Collins and V. A. McKusick. Implications of the human genome project for medical science. *Jama*, 285(5):540–544, 2001.

[12] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

[13] D. Darmon. Normalized q-step specific entropy rate. unpublished personal communication.

[14] D. Darmon. Specific differential entropy rate estimation for continuous-valued time series. *Entropy*, 18(5):190, 2016.

[15] D. Darmon. Information-theoretic model selection for optimal prediction of stochastic dynamical systems from data. *Physical Review E*, 97(3):032206, 2018.

[16] D. Darmon. Specific information dynamics with python (sidpy), version 0.1. `https://github.com/ddarmon/sidpy`, 2018.

[17] D. Darmon and P. E. Rapp. Specific transfer entropy and other state-dependent transfer entropies for continuous-state input-output systems. *Physical Review E*, 96(2):022121, 2017.

[18] W. J. Dondorp and G. M. De Wert. The thousand-dollar genome: an ethical exploration. *European Journal of Human Genetics*, 21(S1):S6, 2013.

[19] J. Fan and Q. Yao. *Nonlinear time series: nonparametric and parametric methods*. Springer Science & Business Media, 2008.

[20] D. Gabor. Theory of communication. part 1: The analysis of information. *Journal of the Institution of Electrical Engineers-Part III: Radio and Communication Engineering*, 93(26):429–441, 1946.

[21] J. Gao and Z. Zheng. Direct dynamical test for deterministic chaos. *EPL (Europhysics Letters)*, 25(7):485, 1994.

[22] R. Gencay and W. D. Dechert. The identification of spurious lyapunov exponents in jacobian algorithms. *Studies in Nonlinear Dynamics & Econometrics*, 1(3), 1996.

[23] C. Gilpin, D. Darmon, Z. Siwy, and C. Martens. Information dynamics of a nonlinear stochastic nanopore system. *Entropy*, 20(4):221, 2018.

[24] P. Grassberger and I. Procaccia. Characterization of strange attractors. *Physical review letters*, 50(5):346, 1983.

[25] P. Grassberger and I. Procaccia. Measuring the strangeness of strange attractors. In *The Theory of Chaotic Attractors*, pages 170–189. Springer, 2004.

[26] K. Gröchenig. Time-frequency analysis and the uncertainty principle. In *Foundations of Time-Frequency Analysis*, pages 21–36. Springer, 2001.

[27] A. Grossmann and J. Morlet. Decomposition of hardy functions into square integrable wavelets of constant shape. *SIAM journal on mathematical analysis*, 15(4):723–736, 1984.

[28] R. V. Hartley. Transmission of information. *Bell Labs Technical Journal*, 7(3):535–563, 1928.

[29] S. Howorka and Z. Siwy. Nanopore analytics: sensing of single molecules. *Chem. Soc. Rev.*, 38:2360–2384, 2009.

[30] S. Howorka and Z. Siwy. Nanopore analytics: sensing of single molecules. *Chemical Society Reviews*, 38(8):2360–2384, 2009.

[31] S. Howorka and Z. Siwy. Nanopores as protein sensors. 30:506–7, 06 2012.

[32] S. Howorka and Z. Siwy. Nanopores and nanochannels: From gene sequencing to genome mapping. *ACS Nano*, 10(11):9768–9771, 2016. PMID: 27934066.

[33] S. Howorka and Z. Siwy. Nanopores and nanochannels: from gene sequencing to genome mapping. *ACS nano*, 10(11):9768–9771, 2016.

[34] H. E. Hurst. Long term storage capacity of reservoirs. *ASCE Transactions*, 116(776):770–808, 1951.

[35] B. Hyland, Z. S. Siwy, and C. C. Martens. Nanopore current oscillations: Nonlinear dynamics on the nanoscale. *The journal of physical chemistry letters*, 6(10):1800–1806, 2015.

[36] L. Innes, D. Gutierrez, W. Mann, S. F. Buchsbaum, and Z. S. Siwy. Presence of electrolyte promotes wetting and hydrophobic gating in nanopores with residual surface charges. *Analyst*, 140:4804–4812, 2015.

[37] L. Innes, M. R. Powell, I. Vlassiouk, C. Martens, and Z. S. Siwy. Precipitation-induced voltage-dependent ion current fluctuations in conical nanopores. *The Journal of Physical Chemistry C*, 114(18):8126–8134, 2010.

[38] R. Izbicki and A. B. Lee. Nonparametric conditional density estimation in a high-dimensional regression setting. *Journal of Computational and Graphical Statistics*, 25(4):1297–1316, 2016.

[39] M. M. J. Effects of random reinforcement sequences1. *Journal of the Experimental Analysis of Behavior*, 22(2):301–310.

[40] S. K. Kannam, S. C. Kim, P. R. Rogers, N. Gunn, J. Wagner, S. Harrer, and M. T. Downton. Sensing of protein molecules through nanopores: a molecular dynamics study. *Nanotechnology*, 25(15):155502, 2014.

[41] H. Kantz and T. Schreiber. *Nonlinear Time Series Analysis*, volume 7. Cambridge University Press, 2004.

[42] A. N. Kolmogorov. A new metric invariant of transient dynamical systems and automorphisms in lebesgue spaces. In *Dokl. Akad. Nauk SSSR (NS)*, volume 119, page 2, 1958.

[43] M. Kolmogorov, E. Kennedy, Z. Dong, G. Timp, and P. A. Pevzner. Single-molecule protein identification by sub-nanopore sensors. *PLOS Computational Biology*, 13:1–14, 05 2017.

[44] L. Kozachenko and N. N. Leonenko. Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii*, 23(2):9–16, 1987.

[45] D. Left. Kernel Density Estimate plot of kernel density esitmator, 2010.

[46] Z. Lincheng and L. Zhijun. Strong consistency of the kernel estimators of conditional density function. *Acta Mathematica Sinica*, 1(4):314–318, 1985.

[47] J. T. Lizier. Jidt: An information-theoretic toolkit for studying the dynamics of complex systems. *arXiv preprint arXiv:1408.3270*, 2014.

[48] J. T. Lizier. Measuring the dynamics of information processing on a local scale in time and space. *Directed Information Measures in Neuroscience, Springer Berlin Heidelberg, Understanding Complex Systems*, pages 161–193, 2014.

[49] J. T. Lizier, M. Prokopenko, and A. Y. Zomaya. Local information transfer as a spatiotemporal filter for complex systems. *Physical Review E*, 77(2):026110, 2008.

[50] D. Lombardi and S. Pant. Nonparametric k-nearest-neighbor entropy estimator. *Physical Review E*, 93(1):013310, 2016.

[51] E. N. Lorenz. Deterministic nonperiodic flow. *Journal of the atmospheric sciences*, 20(2):130–141, 1963.

[52] S. C. Malpas. Neural influences on cardiovascular variability: possibilities and pitfalls. *American Journal of Physiology-Heart and Circulatory Physiology*, 282(1):H6–H20, 2002.

[53] S. L. Marple and S. L. Marple. *Digital spectral analysis: with applications*, volume 5. Prentice-Hall Englewood Cliffs, NJ, 1987.

[54] C. R. Martin and Z. S. Siwy. Learning nature's way: Biosensing with synthetic nanopores. *Science*, 317(5836):331–332, 2007.

[55] S. L. McMurran and J. J. Tattersall. The mathematical collaboration of ml cartwright and je littlewood. *The American mathematical monthly*, 103(10):833–845, 1996.

[56] M. M. Mohammad and L. Movileanu. Protein sensing with engineered protein nanopores. *Nanopore-Based Technology*, pages 21–37, 2012.

[57] J. A. Nelder and R. Mead. A simplex method for function minimization. *The computer journal*, 7(4):308–313, 1965.

[58] G. Nguyen, S. Howorka, and Z. S. Siwy. Dna strands attached inside single conical nanopores: ionic pore characteristics and insight into dna biophysics. *The Journal of membrane biology*, 239(1-2):105–113, 2011.

[59] H. Nyquist. Certain factors affecting telegraph speed. *Transactions of the American Institute of Electrical Engineers*, 43:412–422, 1924.

[60] A. Ostruszka, P. Pakoński, W. Słomczyński, and K. Życzkowski. Dynamical entropy for systems with stochastic perturbation. *Physical Review E*, 62(2):2018, 2000.

[61] T. W. Parks and C. S. Burrus. *Digital filter design*. Wiley-Interscience, 1987.

[62] G. Pérez-Mitta, A. G. Albesa, C. Trautmann, M. E. Toimil-Molares, and O. Azzaroni. Bioinspired integrated nanosystems based on solid-state nanopores:"iontronic" transduction of biological, chemical and physical stimuli. *Chemical science*, 8(2):890–913, 2017.

[63] T. S. Plett, W. Cai, M. Le Thai, I. V. Vlassiouk, R. M. Penner, and Z. S. Siwy. Solid-state ionic diodes demonstrated in conical nanopores. *The Journal of Physical Chemistry C*, 121(11):6170–6176, 2017.

[64] H. Poincaré. *New methods of celestial mechanics*, volume 13. Springer Science & Business Media, 1992.

[65] M. R. Powell, M. Sullivan, I. Vlassiouk, D. Constantin, O. Sudre, C. C. Martens, R. S. Eisenberg, and Z. S. Siwy. Nanoprecipitation-assisted ion current oscillations. *Nature Nanotechnology*, 3(1):51, 2008.

[66] M. R. Powell, M. Sullivan, I. Vlassiouk, D. Constantin, O. Sudre, C. C. Martens, R. S. Eisenberg, and Z. S. Siwy. Nanoprecipitation-assisted ion current oscillations. *Nature Nanotechnology*, 3(1):51, 2008.

[67] Y. Qiu, P. Hinkle, C. Yang, H. E. Bakker, M. Schiel, H. Wang, D. Melnikov, M. Gracheva, M. E. Toimil-Molares, A. Imhof, and Z. S. Siwy. Pores with longitudinal irregularities distinguish objects by shape. *ACS Nano*, 9(4):4390–4397, 2015. PMID: 25787224.

[68] G. A. Reyes del Paso, W. Langewitz, L. J. Mulder, A. Roon, and S. Duschek. The utility of low frequency heart rate variability as an index of sympathetic cardiac tone: a review with emphasis on a reanalysis of previous studies. *Psychophysiology*, 50(5):477–487, 2013.

[69] A. Rößler. Runge–kutta methods for the strong approximation of solutions of stochastic differential equations. *SIAM Journal on Numerical Analysis*, 48(3):922–952, 2010.

[70] M. P. Sawicki, G. Samara, M. Hurwitz, and E. Passaro. Human genome project. *The American journal of surgery*, 165(2):258–264, 1993.

[71] M. Schiel and Z. S. Siwy. Diffusion and trapping of single particles in pores with combined pressure and dynamic voltage. *The Journal of Physical Chemistry C*, 118(33):19214–19223, 2014.

[72] R. B. Schoch, J. Han, and P. Renaud. Transport phenomena in nanofluidics. *Reviews of modern physics*, 80(3):839, 2008.

[73] T. Schreiber and A. Schmitz. Improved surrogate data for nonlinearity tests. *Physical Review Letters*, 77(4):635, 1996.

[74] D. W. Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

[75] I. W. Selesnick and C. S. Burrus. Generalized digital butterworth filter design. *IEEE Transactions on signal processing*, 46(6):1688–1694, 1998.

[76] C. E. Shannon. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, 5(1):3–55, 2001.

[77] B. W. Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.

[78] J. S. Simonoff. *Smoothing methods in statistics*. Springer Science & Business Media, 2012.

[79] R. Simpson. Intro. electronics for scientists and engineers 2nd edit, 1987.

[80] Y. G. Sinai. On the notion of entropy of a dynamical system. In *Dokl. Akad. Nauk. SSSR*, volume 124, page 768, 1959.

[81] H. Singh, N. Misra, V. Hnizdo, A. Fedorowicz, and E. Demchuk. Nearest neighbor estimates of entropy. *American journal of mathematical and management sciences*, 23(3-4):301–321, 2003.

[82] Z. S. Siwy and M. Davenport. Nanopores: Graphene opens up to dna. *Nature nanotechnology*, 5(10):697, 2010.

[83] Z. S. Siwy, M. R. Powell, A. Petrov, E. Kalman, C. Trautmann, and R. S. Eisenberg. Calcium-induced voltage gating in single conical nanopores. *Nano letters*, 6(8):1729–1734, 2006.

[84] R. G. Stockwell, L. Mansinha, and R. Lowe. Localization of the complex spectrum: the s transform. *IEEE transactions on signal processing*, 44(4):998–1001, 1996.

[85] J. V. Stone. *Information theory: A tutorial introduction*. Sebtel Press, 2015.

[86] S. P. Strong, R. Koberle, R. R. de Ruyter van Steveninck, and W. Bialek. Entropy and information in neural spike trains. *Phys. Rev. Lett.*, 80:197–200, Jan 1998.

[87] M. Tarnopolski. Correlations between hurst exponent and maximal lyapunov exponent for some low-dimensional discrete conservative dynamical systems. *arXiv preprint arXiv:1501.03766*, 2015.

[88] J. Theiler. Spurious dimension from correlation algorithms applied to limited time-series data. *Physical review A*, 34(3):2427, 1986.

[89] Y. Tian, L. Wen, X. Hou, G. Hou, and L. Jiang. Bioinspired ion-transport properties of solid-state single nanochannels and their applications in sensing. *ChemPhysChem*, 13(10):2455–2470, 2012.

[90] H. Tong. *Non-linear time series: a dynamical system approach*. Oxford University Press, 1990.

[91] B. Van der Pol and J. Van Der Mark. Frequency demultiplication. *Nature*, 120(3019):363, 1927.

[92] B. M. Venkatesan and R. Bashir. Nanopore sensors for nucleic acid analysis. *Nature nanotechnology*, 6(10):615, 2011.

[93] M. Wanunu. Nanopores: A journey towards dna sequencing. *Physics of life reviews*, 9(2):125–158, 2012.

[94] L. Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

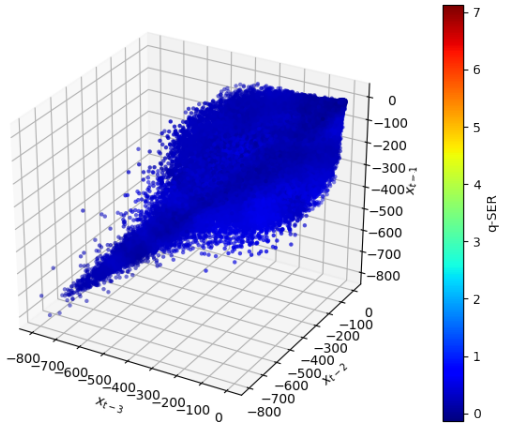
[95] R. W. Yeung. A new outlook on shannon's information measures. *IEEE transactions on information theory*, 37(3):466–474, 1991.

[96] H. Zhang, Y. Tian, and L. Jiang. Fundamental studies and practical applications of bio-inspired smart solid-state nanopores and nanochannels. *Nano Today*, 11(1):61–81, 2016.

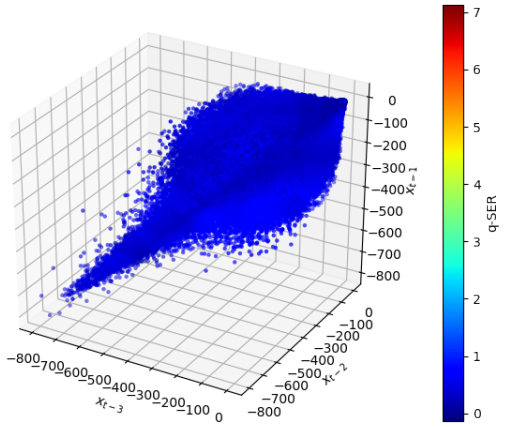
Appendix A

Additional Normalized q-step Specific Entropy Rate Results

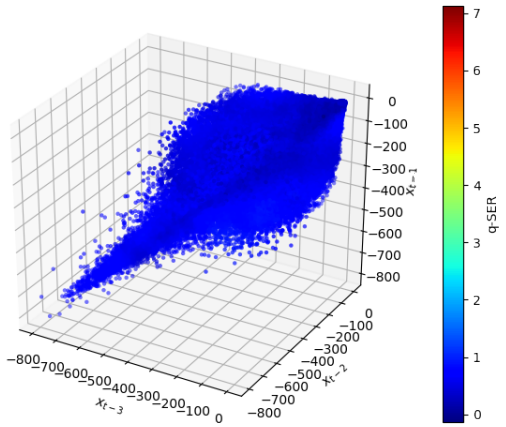
We show additional normalized q-step specific entropy rate results for our experimental nanopore with downsampling factors of 4 and 8. One notable difference between these normalized q step specific entropy rates and the factor of 2 downsampled normalized q-step specific entropy rate is the region which undergoes the initial increase with increasing q. For these higher rates of downsampling the lowest conductance states experience the initial increase in the normalized q-step specific entropy rate. in the factor 8 downsampled results we also see that this increase begins to fall off as we reach the highest measured values of q. If we think about these results, it is logical to expect that as we downsample the regions associated with the lowest conductance states would have a q-step future that looks very different than the corresponding 0-step future. When we have downsampling rates of 8 it also makes sense that for the highest values of q this effect would diminish as it is possible that q steps down it is more likely the system will be undergoing a subsequent transition to the negative conductance states.


q-SER
x_{t-1}
x_{t-2}
x_{t-3}

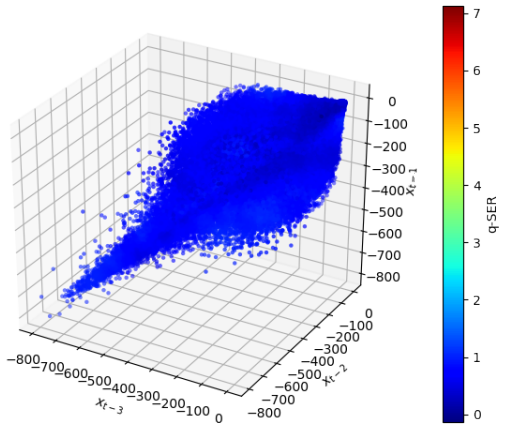
(a)


q-SER
x_{t-1}
x_{t-2}
x_{t-3}

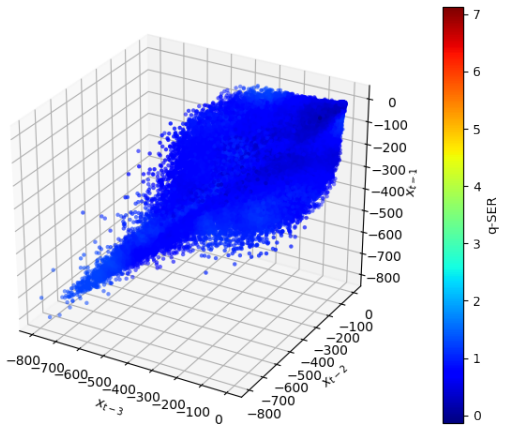
(b)


q-SER
x_{t-1}
x_{t-2}
x_{t-3}

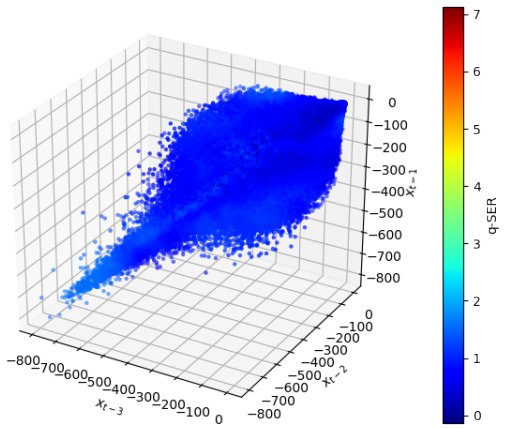
(c)


q-SER
x_{t-1}
x_{t-2}
x_{t-3}

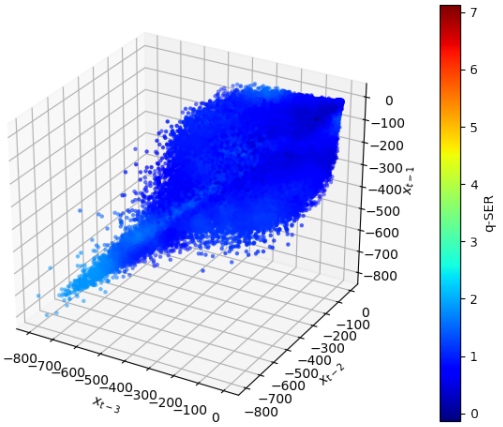
(d)


q-SER
x_{t-1}
x_{t-2}
x_{t-3}

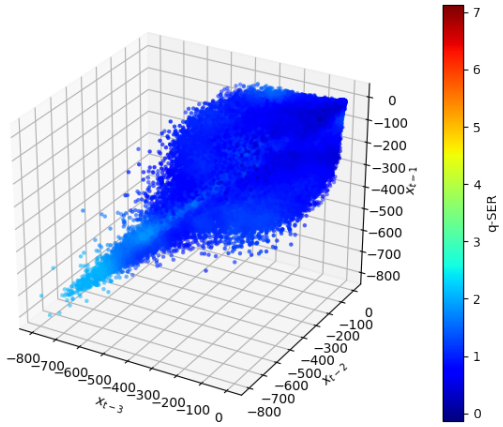
(e)


q-SER
x_{t-1}
x_{t-2}
x_{t-3}

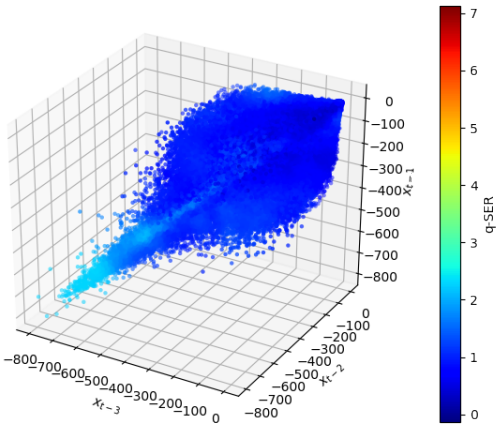
(f)


q-SER
x_{t-3}
x_{t-2}
x_{t-1}

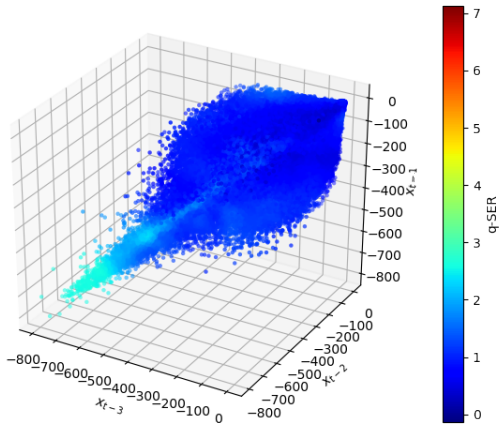
(g)


q-SER
x_{t-3}
x_{t-2}
x_{t-1}

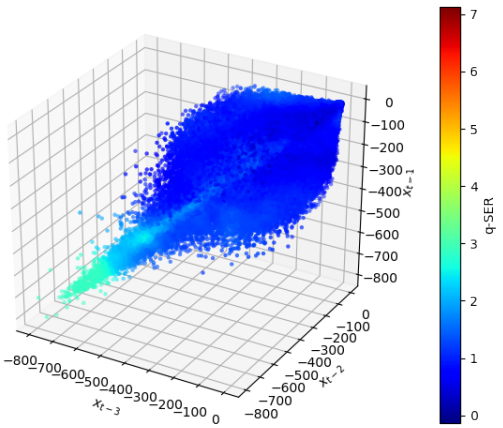
(h)


q-SER
x_{t-3}
x_{t-2}
x_{t-1}

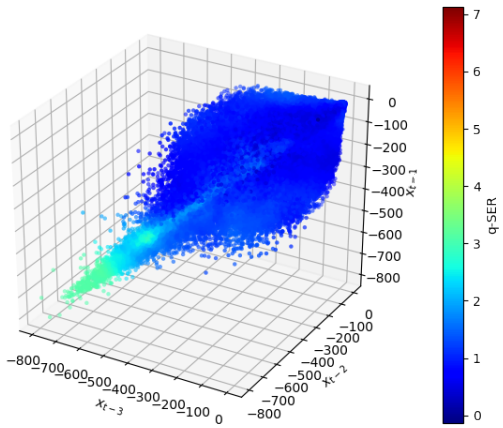
(i)


q-SER
x_{t-3}
x_{t-2}
x_{t-1}

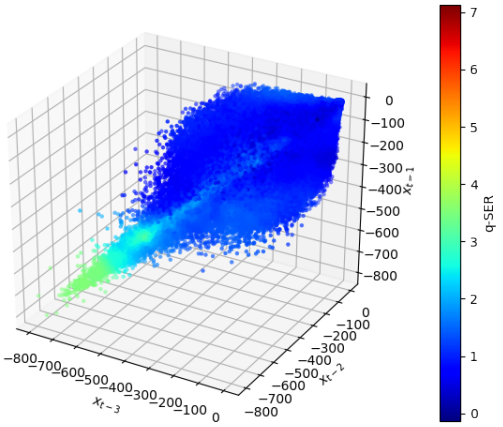
(j)


q-SER
x_{t-3}
x_{t-2}
x_{t-1}

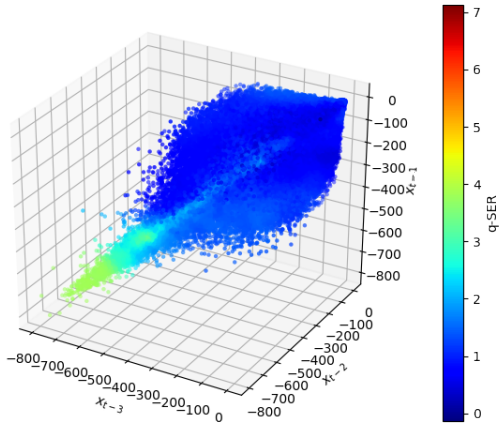
(k)


q-SER
x_{t-3}
x_{t-2}
x_{t-1}

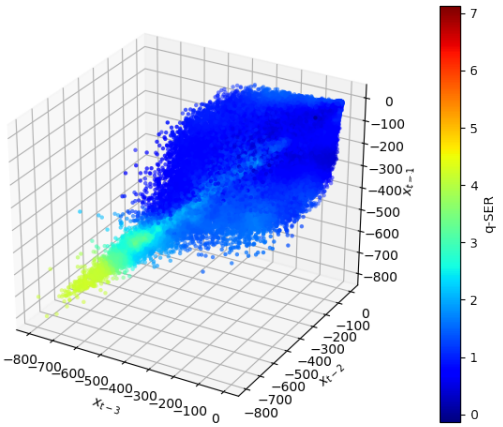
(l)


q-SER
x_{t-3}
x_{t-2}
x_{t-1}

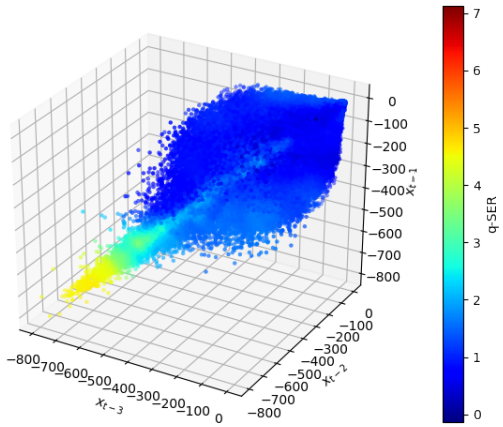
(m)


q-SER
x_{t-3}
x_{t-2}
x_{t-1}

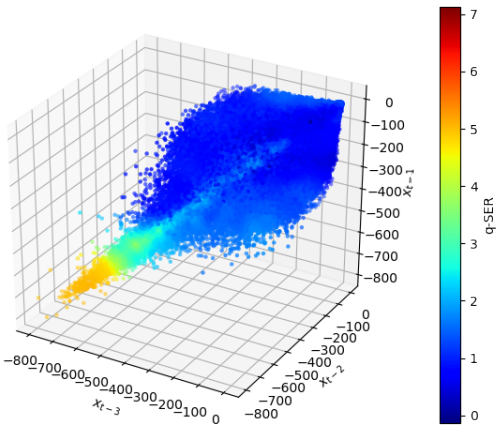
(n)


q-SER
x_{t-3}
x_{t-2}
x_{t-1}

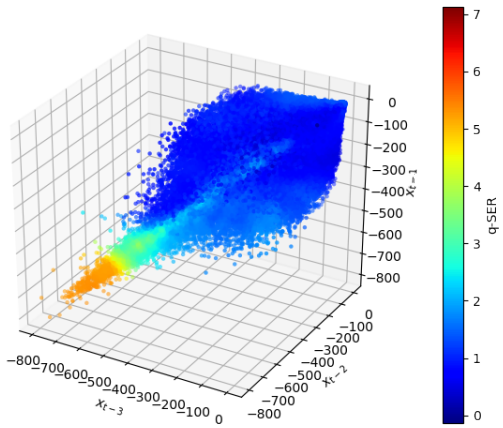
(o)


q-SER
x_{t-3}
x_{t-2}
x_{t-1}

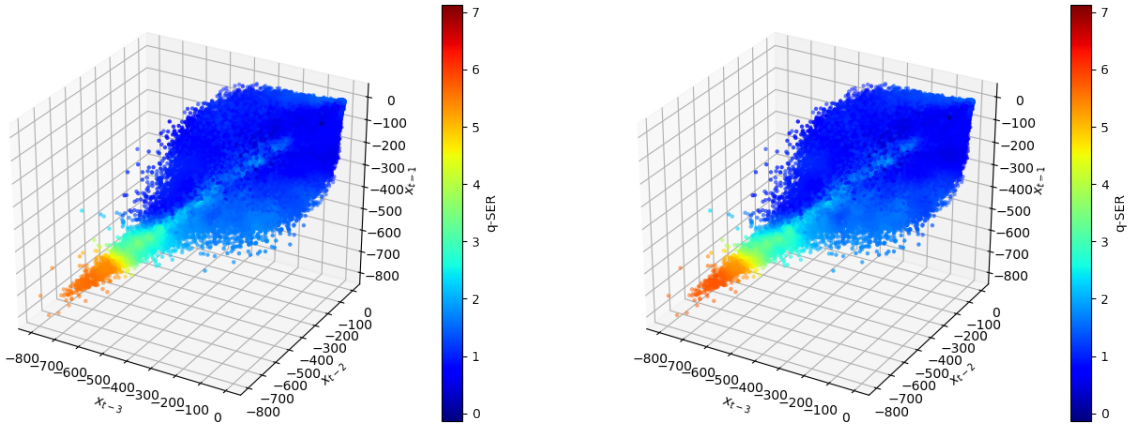
(p)


q-SER
x_{t-3}
x_{t-2}
x_{t-1}

(q)

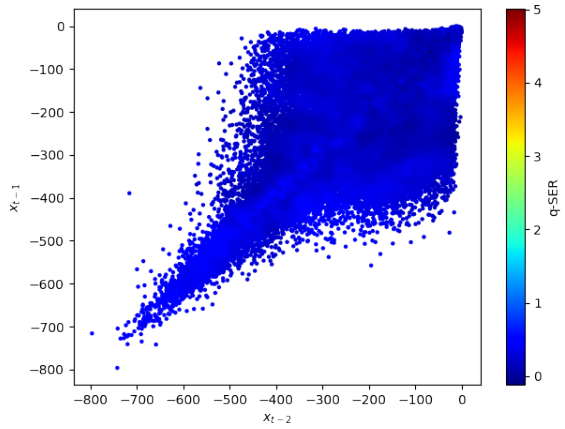

q-SER
x_{t-3}
x_{t-2}
x_{t-1}

(r)

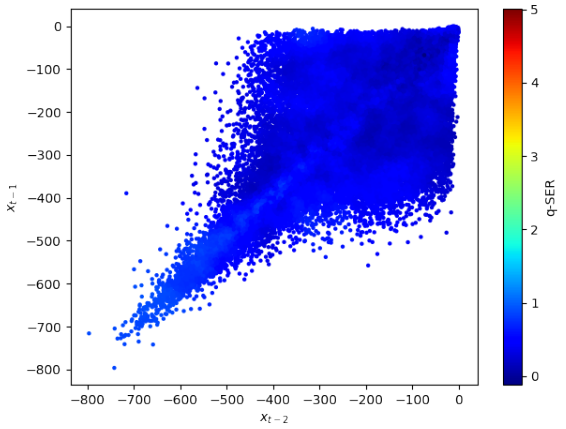


(s) (t)

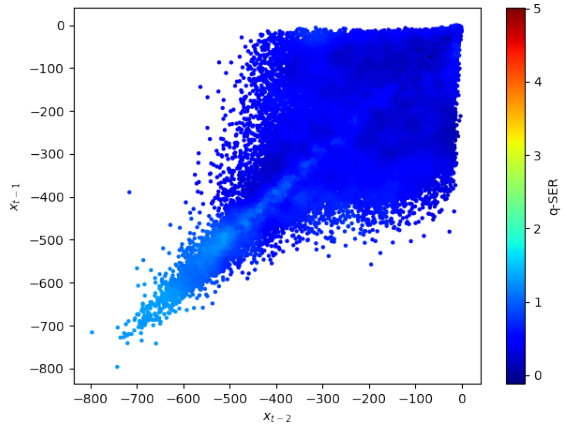
Figure A.1: This figure shows the evolution of the normalized q-step specific entropy rate with changing q. q changes from 1 (a) to 20 (t). The q step future looks increasingly more different than the zero step future with increasing q and the effect is most notable in the lowest conductance states. Downsampling rate = 4.


0
−100
−200
−300
−400
−500
−600
−700
−800
x_{t-1}
−800 −700 −600 −500 −400 −300 −200 −100 0
x_{t-2}
5
4
3
2
1
0
q-SER

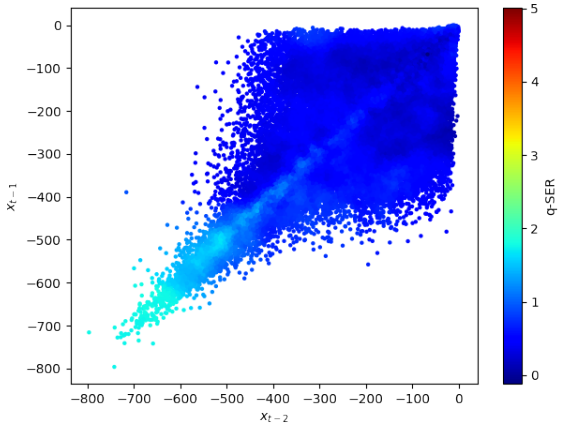
(a)


0
−100
−200
−300
−400
−500
−600
−700
−800
x_{t-1}
−800 −700 −600 −500 −400 −300 −200 −100 0
x_{t-2}
5
4
3
2
1
0
q-SER

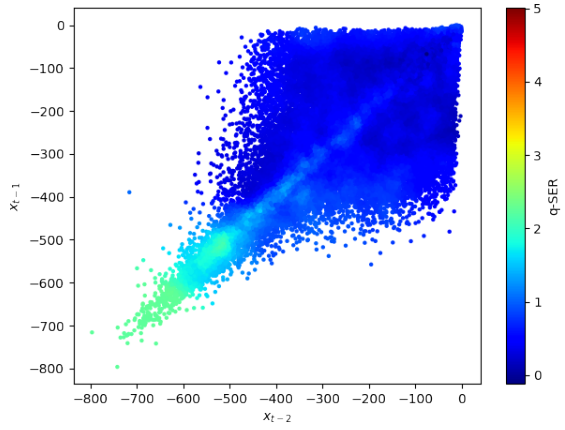
(b)


0
−100
−200
−300
−400
−500
−600
−700
−800
x_{t-1}
−800 −700 −600 −500 −400 −300 −200 −100 0
x_{t-2}
5
4
3
2
1
0
q-SER

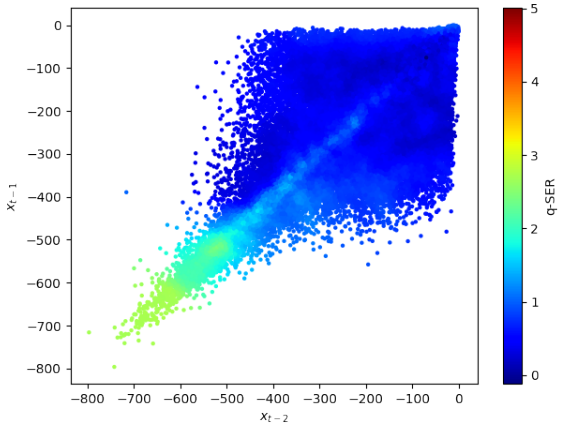
(c)


0
−100
−200
−300
−400
−500
−600
−700
−800
x_{t-1}
−800 −700 −600 −500 −400 −300 −200 −100 0
x_{t-2}
5
4
3
2
1
0
q-SER

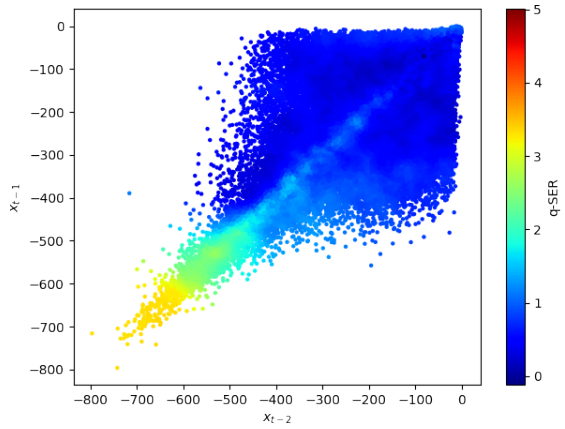
(d)


0
−100
−200
−300
−400
−500
−600
−700
−800
x_{t-1}
−800 −700 −600 −500 −400 −300 −200 −100 0
x_{t-2}
5
4
3
2
1
0
q-SER

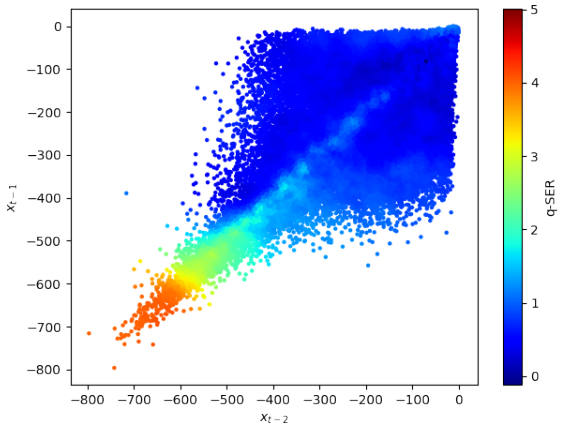
(e)


0
−100
−200
−300
−400
−500
−600
−700
−800
x_{t-1}
−800 −700 −600 −500 −400 −300 −200 −100 0
x_{t-2}
5
4
3
2
1
0
q-SER

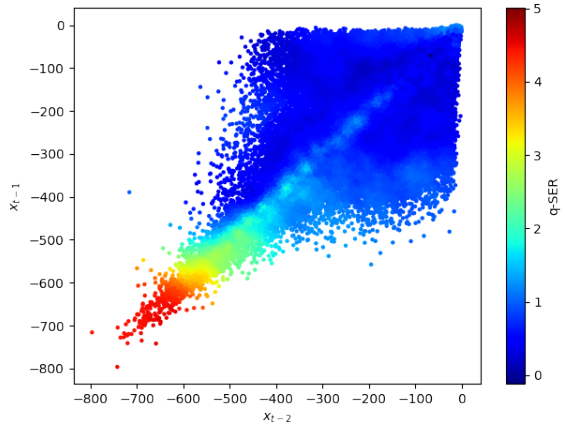
(f)


x_{t-1}
x_{t-2}
q-SER

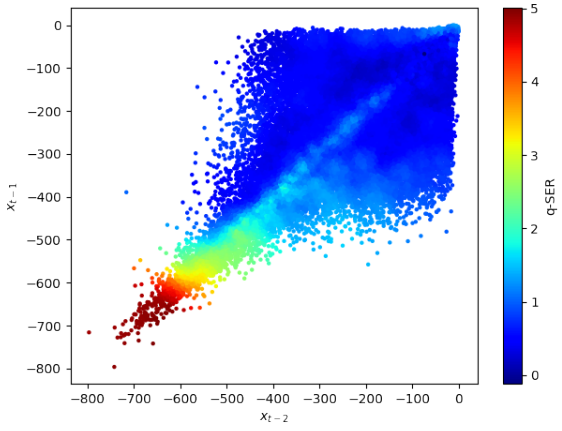
(g)


x_{t-1}
x_{t-2}
q-SER

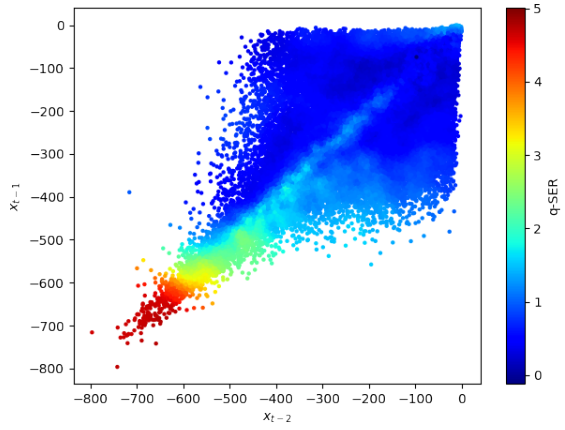
(h)


x_{t-1}
x_{t-2}
q-SER

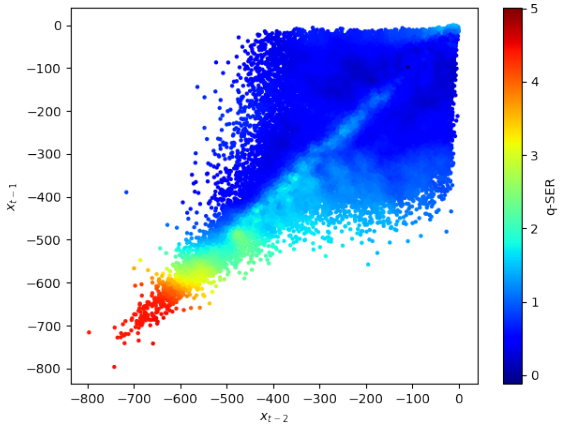
(i)


x_{t-1}
x_{t-2}
q-SER

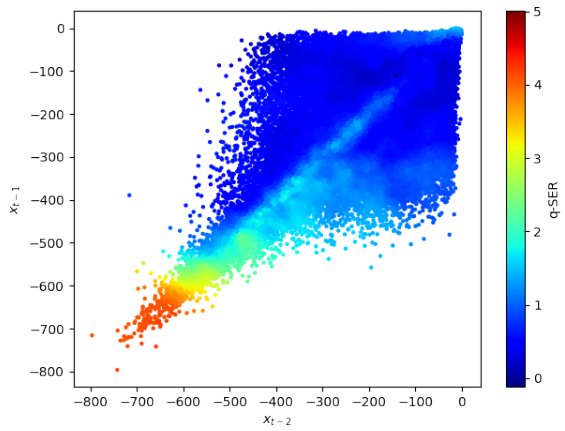
(j)


x_{t-1}
x_{t-2}
q-SER

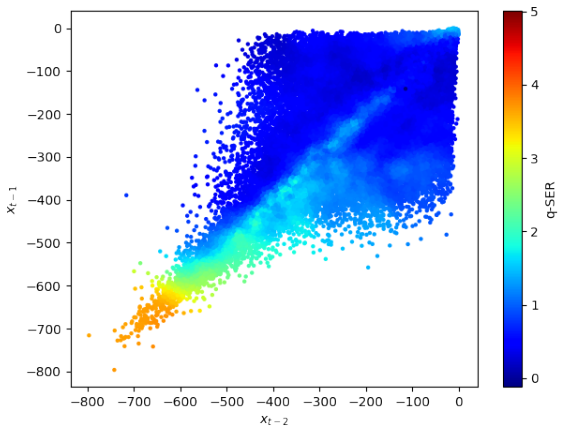
(k)


x_{t-1}
x_{t-2}
q-SER

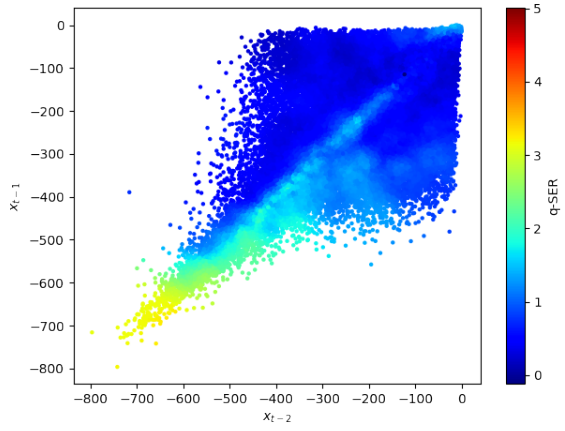
(l)


x_{t-1}
x_{t-2}
q-SER

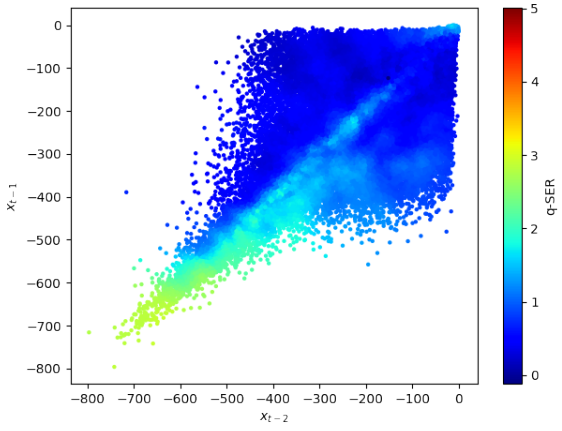
(m)


x_{t-1}
x_{t-2}
q-SER

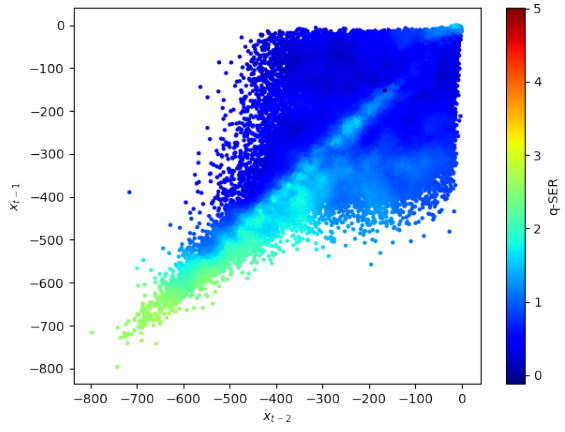
(n)


x_{t-1}
x_{t-2}
q-SER

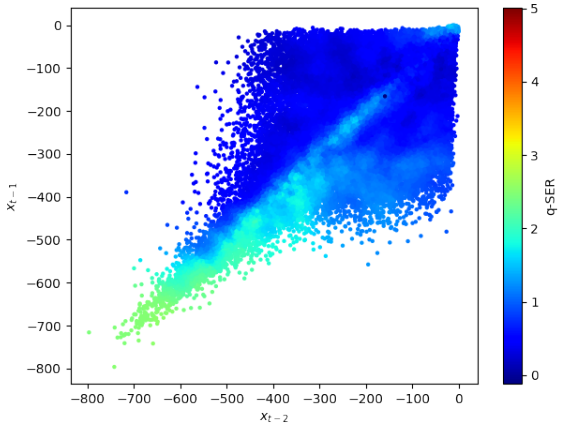
(o)


x_{t-1}
x_{t-2}
q-SER

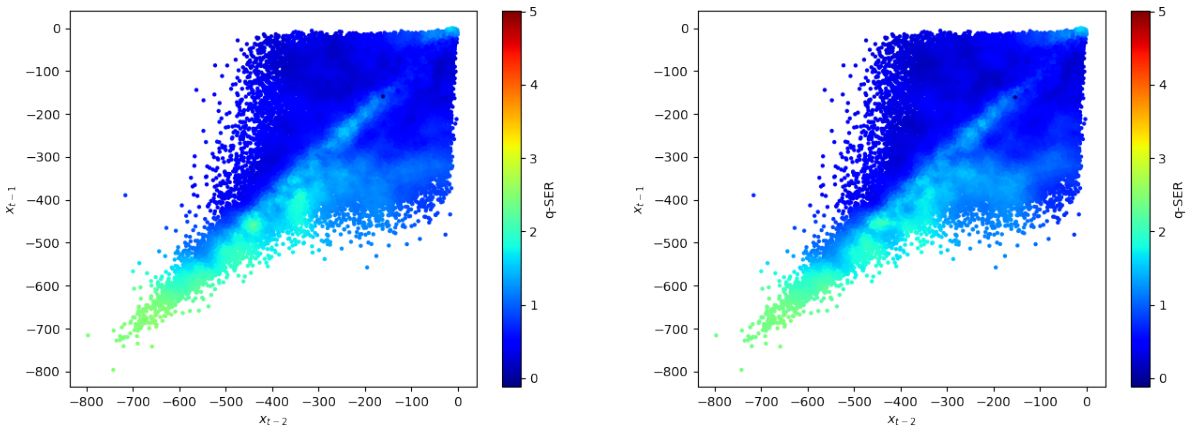
(p)


x_{t-1}
x_{t-2}
q-SER

(q)


x_{t-1}
x_{t-2}
q-SER

(r)



(s) (t)

Figure A.2: This figure shows the evolution of the normalized q-step specific entropy rate with changing q. q changes from 1 (a) to 20 (t). The q step future looks increasingly more different than the zero step future with increasing q and the effect is most notable in the lowest conductance states. This effect diminishes at the highest measured q values. Downsampling rate = 8.

Appendix B

Additional Surrogate Analysis

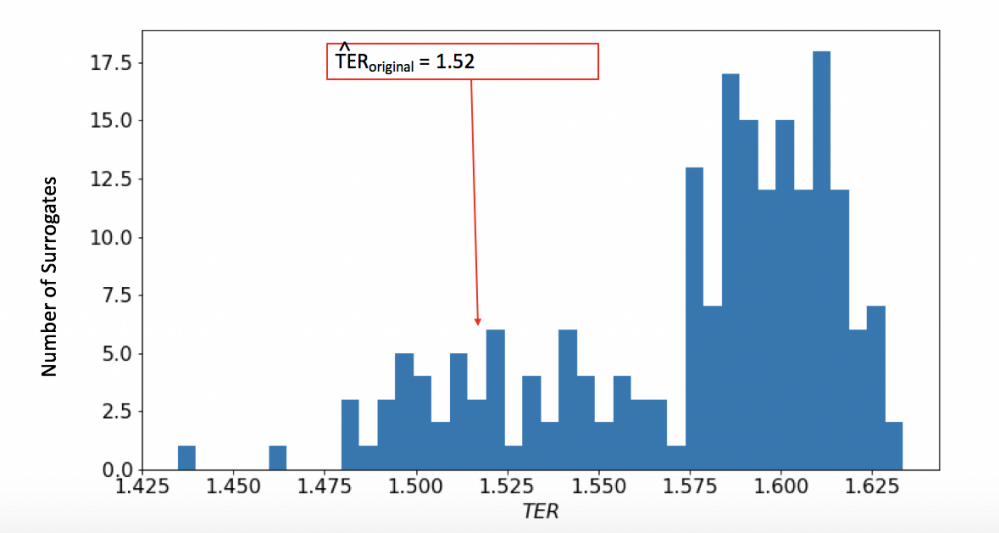
We started the process of performing additional surrogate analysis using different thresholds for identification of events of possible interest. We do not know a priori whether partial pore openings play an important role in the underlying dynamics of the experimental system. We have begun testing additional threshold values, prioritizing these tests for data where initial hypothesis testing with other current thresholds did not yield evidence of an underlying nonlinear structure in the interevent interval sequences.

We present results for the following conditions:

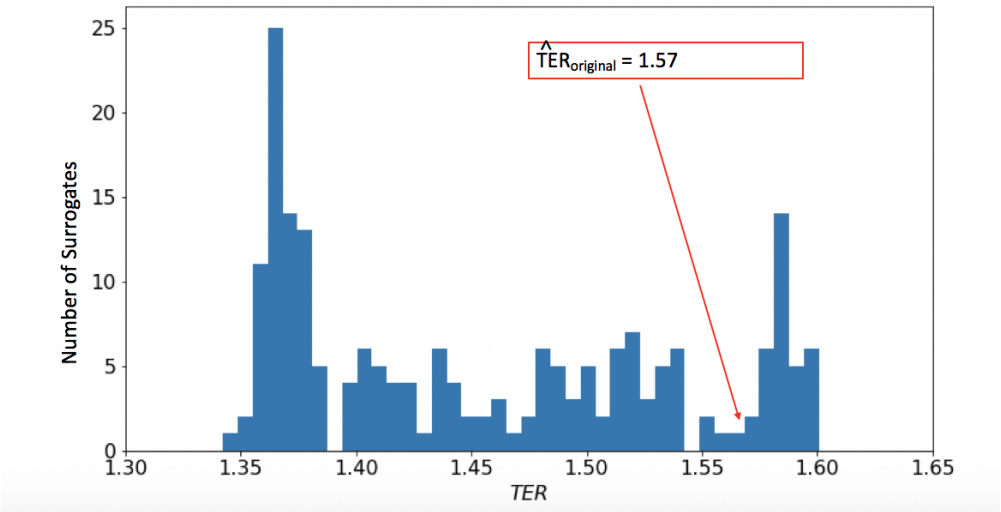
$Bias\ (V)$	$Threshold\ (pA)$	$P\ Value$
-0.85	-250	0.15
-0.90	-250	0.83

Table B.1: The table shows the results of additional hypothesis testing for experimental nanopore interevent intervals. The null hypothesis is not rejected in either test.

The histograms for total entropy rate are shown below with the value of the total entropy rate for the original interevent interval sequences explicitly marked.



(a) $V_{bias} = -0.85\ V$



(b) $V_{bias} = -0.90\ V$

Figure B.1: This figure shows histograms for the estimated total entropy rates associated with the interevent interval sequence surrogates. The estimated total entropy rate of the original entropy rate is labeled and shows whether it lies within the distribution of interevent intervals for the surrogates. The null hypothesis is not rejected for either set of experimental conditions. Further experimental investigation should be conducted to ensure repeatability and a wider range of applied bias voltages should be explored.