Characterisation and Estimation of Entropy Rate for Long Range Dependent Processes

Andrew Feutrill

March 11, 2023

Thesis submitted for the degree of
Doctor of Philosophy
in
Applied Mathematics
at The University of Adelaide
Faculty of Engineering, Computer and Mathematical Sciences
School of Mathematical Sciences

THE UNIVERSITY
of ADELAIDE

Contents

List of Tables

List of Figures

Signed Statement

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name, in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. In addition, I certify that no part of this work will, in the future, be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint-award of this degree.

I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Signed: ..... Date: .....

Acknowledgements

I would like to thank my principal supervisor, Professor Matthew Roughan, who has been a great mentor and support through the entire candidature. This would have not been possible without his foresight, guidance and shaping of many different ideas. In particular, thank you for your work the last year and a half through some very difficult times and uncertainty.

Thanks to Yuval for all of his feedback on the applied context of the work. Working on real data enabled us to identify current gaps in the mathematical literature, which we could then aim to understand. Also, having exposure to the cyber security environment and other problems has been extremely valuable.

A special thank you to Giang who assisted greatly at the end of my candidature. Her expertise and advice was crucial to ensuring I could get the final papers completed and the thesis written.

Thank you to Caitlin, Angus, Dennis and Phill whose friendship and support was crucial over the last few years. Whether it was working together on problems, going for a Banh Mi or a drink, I couldn't have done it without them. I wish to thank Adam and George, for interesting discussions and exposure to new problems, which was always a welcome distraction from my immediate work, but also exposure to different areas of mathematics, which could then inspire my own work.

I wish to thank Data to Decisions CRC for the initial support for this work and then CSIRO/Data61 for their support over the second half of the candidature. Thanks to the School of Mathematics for their support, in particular Finnur Larusson as head of school who has been great in very difficult times.

I wish to thank my Mother in Law, Rosie, who has been an amazing supporter since before I even enrolled, supporting and inspiring me to take this on. In addition, for all of the help with Polly, which we couldn't have done without. Thanks to Polly for being there with me in the late nights when she was very young and inspiring me to continue.

Finally, I wish to thank my wife Alice. She has been such a great sup-

porter and none of this would have been possible without her. Words cannot express how grateful I am for her being there and everything she has done for me and the family.

Dedication

To Alice and Polly

Abstract

Much of the theory of random processes has been developed with the assumption that distant time periods are weakly correlated. However, it has been discovered in many real-world phenomena that this assumption is not valid. These findings have resulted in extensive research interest into stochastic processes that have strong correlations that persist over long time periods. This phenomenon is called *long range dependence*.

This phenomena has been defined in the time and frequency domains by the slow decay of their autocorrelation function and the existence of a pole at the origin of the spectral density function, respectively. Information theory has proved very useful in statistics and probability theory. However, there has not been much research into the information theoretic properties and characterisations of this phenomena.This thesis characterises the phenomena of long range dependence, for discrete and continuous-valued stochastic processes in discrete time, by an information theoretic measure, the *entropy rate*.

The entropy rate measures the amount of information contained in a stochastic process on average, per random variable. Common characterisations of long range dependence in the time and frequency domains are given by the slow convergence to quantities of interest, such as the sample mean. We show that this type of behaviour is present in the entropy rate function, by showing that long range dependence also has slow convergence of the conditional entropy to the entropy rate, due to some entropic quantities diverging to infinity. As an extension we show for classes of Gaussian processes and Markov chains that long range dependence by an infinite amount of shared information between the past and future of a stochastic process.

The slow convergence has the impact of making accurate estimation of the differential entropy rate on data from long range dependent processes difficult, to the extent that existing techniques either are not accurate or are computationally intensive. We introduce a new estimation technique, that is able to balance these two concerns and make quick and accurate estimates of the differential entropy rate from continuous-valued data. We

develop and utilise a connection between the differential entropy rate and the Shannon entropy rate of its quantised process as the basis of the estimation technique. This allows us to draw on the extensive research into Shannon entropy rate estimation on discrete-valued data, and we show that properties for the differential entropy rate estimator can be inherited from the choice of Shannon entropy rate estimator.

Chapter 1

Introduction

In this thesis, we are concerned with understanding the behaviour of information theoretic based measures when applied to *long range dependent* (LRD) processes, *i.e.,* processes that have strong correlations with the past. The field of information theory is centred around the concept of *Shannon entropy*, which is a measure of the uncertainty of a random variable. An extension of this concept to stochastic processes produces a measure called the *entropy rate*. We characterise LRD by the convergence of the entropy rate, and then develop estimation techniques for the entropy rate which are robust to the influence of LRD.

Traditional models in probability theory have been extremely successful in modelling a wide range of real-world phenomena, such as arrivals of phone calls to exchanges, epidemics, and financial markets. These models typically rely on the assumption that the phenomena have weak correlations between distant time intervals, That is, the correlations between two points decays quickly as the distance between the time points increases. For example, the number of calls that occurred 10 hours ago has low correlation with the number of calls in the next hour. However for real-world processes, there exist many phenomena where this assumption is not valid and events that occur a long time in the past have a large impact on the present and future values.

The initial research into LRD was driven by the hydrologist Harold Edwin Hurst into flooding on the Nile river [89]. He was aiming to calculate the capacity of a reservoir on the Nile river and his analysis led to some unexpected conclusions. He discovered, by analysing empirical data, that a quantity called the rescaled range, measuring the variability of time series data over the entire time period, grows at a much faster rate than expected, assuming weak temporal correlations. Given the stochastic models at the time, it was expected that the rescaled range would increase in proportion to

$n^{\frac{1}{2}}$. However, Hurst discovered empirically that the rescaled range grows at the rate of $\sim n^{0.72}$, indicating that the strength of correlations of the present value with time periods in the distant past were much stronger than expected. This phenomena is called *long range dependence* (LRD), also called long memory. Another class which has strong negative correlations is called *constrained short range dependent* (CSRD), which has some similar properties to LRD processes. Models that do not exhibit these phenomena are called *short range dependent* (SRD), or short memory.

The discovery of LRD led to the development of stochastic models that could explain and induce much longer term correlations. Mandelbrot and Van Ness [127] introduced a class of models called fractional Brownian motion, and its increment process fractional Gaussian noise (FGN), which are LRD extensions of Brownian motion and Gaussian noise. These models were developed to model telecommunications systems, after an empirical discovery that errors tended to cluster together, rather than the typical assumption that errors were independent of each other.

A second class of LRD model is a linear time series model, autoregressive fractionally integrated moving average processes (ARFIMA). These models extend autoregressive integrated moving average (ARIMA) processes, utilising a fractional exponent of the differencing operator inducing regression on the infinite past. They were developed independently by Granger and Joyeux [77] and Hosking [87], to model complex phenomena in economics and hydrology.

Fractional Gaussian noise and ARFIMA processes have many properties in common, such as slow convergence of estimators, power-law decay of the autocorrelation function and the appearance of local trends which can be mistaken for non-stationarity.

An important parameter that is used to measure the strength of correlations, and therefore the degree of LRD in a process, is named the *Hurst* parameter, $\mathcal{H}$. This provides a characterisation, for SRD processes $\mathcal{H} = 0.5$, LRD processes have $\mathcal{H} > 0.5$ and CSRD have $\mathcal{H} < 0.5$. We show that for the two common LRD models, FGN and ARFIMA, that the amount of uncertainty in the processes decreases as $\mathcal{H} \to 0$ or 1 with the increase in the strength of positive and negative correlations.

The entropy rate has been used as a measure of the intrinsic uncertainty of a stochastic process and therefore a measure of the complexity of a stochastic process. An LRD process shares much information between the past and future. Hence, there is less uncertainty in LRD processes as compared with SRD processes, and thus we expect a lower entropy rate for LRD processes.

A constant theme of this work is the slow convergence rate of quantities of interest and in particular the entropy rate for LRD processes. An existing

example is the slow convergence of the sample mean to the expected value, given a large number of observations, n. For many processes the variance of the sample mean converges at the rate of n^{-1}, however in some cases of LRD processes the convergence rate is $n^{2\mathcal{H}-2}$. For the class of Gaussian processes we show in this thesis that the convergence rate of the conditional entropy to the entropy rate for LRD and CSRD processes is slower than for processes with SRD.

We show that some related measures, the excess entropy and mutual information between past and future are infinite for LRD/CSRD Gaussian processes, similar to the sum of the autocorrelation function in the time domain. This supports an alternate perspective on persistent correlations, as stochastic processes whose entire past and future share infinite information.

Similarly, we analyse the convergence in the case of discrete space LRD Markov chains, which are characterised by an infinite second moment of the return time random variable. In this case we show a similar result, that the convergence rate of the conditional entropy to the entropy rate for LRD processes is slower than the short memory case and the convergence rate is a function of the Hurst parameter. These results reinforce the idea that LRD is characterised by slower convergence to quantities, and demonstrates that this behaviour extends to other information theoretic quantities, such as the mutual information between past and future.

Estimation of the entropy rate is used in real-world applications to classify the uncertainty of a stochastic process. However, slow convergence of the entropy rate raises some questions: *"How do entropy rate estimators behave for LRD processes?"* and *"Can we develop computationally efficient estimators that are robust to the influence of LRD?"*.

For continuous valued processes there are a variety of estimation techniques and measures: sample entropy, approximate entropy, permutation entropy and specific entropy. We illustrate that the first three measures are unable to capture the underlying uncertainty of LRD processes. Specific entropy, however, provides robust estimation of the entropy rate, but it does so at a much higher computational complexity. Hence, we conclude that current estimation techniques have some issues in accurately estimating entropy rates with low time complexity. We improve this situation and develop an entropy rate estimation technique, NPD Entropy, that is robust to the influence of LRD and lower computational complexity.

NPD Entropy leverages non-parametric estimation techniques developed for the Shannon entropy rate, *i.e.*, on discrete state spaces to calculate differential entropy rate estimates. Shannon entropy rate estimators are often based on limit theorems involving expressions of the entropy rate and vary in their properties and performance on data. We prove a connection between

the differential entropy rate and the Shannon entropy rate of the quantised data. Then NPD Entropy estimates the Shannon entropy rate of a quantised version of the data, and then converts the estimate to a differential entropy rate estimate. The flexibility of this approach allows the choice of a suitable estimation technique, since many properties are inherited from the Shannon entropy rate estimator.

This thesis provides a new perspective on LRD, through the lens of information theoretic quantities. We demonstrate that the typical behaviour of LRD carries over into the information domain, characterised by slow convergence to important quantities, such as the entropy rate and the mutual information between past and future. However, given this difficult setting we can still estimate the differential entropy rate quickly and accurately.

1.1 Key Contributions

This thesis investigates and characterises the relationship between the entropy rate and its behaviour for LRD stochastic processes. We develop and apply entropy rate estimation techniques to data derived from LRD processes, that are robust to the influence of the LRD behaviour. The contributions are in the information theory of Gaussian processes, Markov chains and robust estimation of the differential entropy rate. We summarise the key contributions of this thesis as follows.

1. Proving for Gaussian process that the excess entropy and mutual information between past and future are equivalent. For LRD and CSRD processes we prove that the mutual information between past and future is infinite for many classes of processes, supporting an alternate definition of LRD stationary processes. Then proving that the convergence rates of the conditional entropy to the entropy rate is at a slower rate for stationary LRD/CSRD Gaussian processes, than SRD Gaussian processes.

2. Proving equivalence of the convergence of the conditional entropy to the entropy rate for Markov chains, to the convergence rate of the n-step probability transitions to the stationary distribution. Leading to the result for Markov chains, that the mutual information between past and future is infinite. Then proving that the convergence rate is dependent on the value of the Hurst parameter.

3. A comprehensive review of the current state of the art in entropy rate estimation for discrete and continuous-valued processes.

4. Introducing a new estimation technique for the differential entropy rate, NPD-Entropy, which exploits a connection between Shannon and differential entropy rate, and utilises a Shannon entropy rate estimator on a quantised version of the process. Proving that NPD-Entropy is able to inherit favourable statistical properties from the choice of the Shannon entropy rate estimator used in the implementation.

1.2 Outline of Thesis

Chapter 2 provides the background material for understanding the thesis. Most of the definitions and results are standard but are included for completeness and because in some cases there are small variations in the literature.

Chapter 3 analyses entropy rate of LRD Gaussian processes to provide an information theoretic characterisation of Gaussian processes. We begin with analysing two very commonly used LRD stochastic models, Fractional Gaussian Noise and ARFIMA(p,d,q). We define expressions for their entropy rates, and illustrate the influence of the strength of the correlations of the past and the total variance of the process, by analysing the entropy rate as a function of the Hurst parameter. We prove, for Gaussian processes, that the mutual information between past and future and the excess entropy are equivalent, showing an identical link in differential entropy that was shown for Shannon entropy by Crutchfield and Feldman [44], that is that they are equal for Gaussian processes. Subsequently we show they are infinite for large classes of LRD and CSRD processes.

Chapter 4 analyses the entropy rate of LRD Markov chains to provide an information theoretic characterisation. We analyse the convergence rate of the conditional entropy to the entropy rate, and show that it is equivalent to the convergence of n-step transition probabilities to the stationary distribution. This is a very well studied problem in the theory of Markov chains, called Markov chain mixing. We investigate the finiteness of the mutual information between past and future, and show that this is infinite for all LRD Markov chains, and in the case of power-law tails of the return time random variable that this forms a boundary between LRD and SRD processes, with SRD processes being finite. Utilising the connection with the Markov chain mixing problem, and the rich theory that has been developed, we analyse the convergence rate of the n-step probability transitions to the stationary distribution for LRD processes. We show that this is closely related to the finiteness of fractional moments of the return time distribution. The we find the rate of convergence for LRD Markov chains with power-law tailed com-

plementary cumulative distribution functions of the return time distribution, and show that the rate of convergence is related to the Hurst parameter.

Chapter 5 presents a review of entropy rate estimation for both Shannon and differential entropy rate, as a comprehensive review on entropy rate estimation is not known to the authors. We present and discuss the a wide variety of techniques for the parametric estimation for Gaussian, Markov, Hidden Markov and renewal processes. Then we investigate the state-of-the-art for non-parametric estimation, which make estimates without any assumptions on the process that generated the data. Given fewer techniques developed for continuous state spaces we focus in-depth on the estimation techniques, Approximate Entropy, Sample Entropy, Permutation Entropy and Specific Entropy. We conclude that there is a gap in the research for differential entropy estimators that are robust to LRD.

Chapter 6 focuses on the estimation of the differential entropy rate for LRD processes, and in particular developing efficient techniques that are robust to the influence of LRD. Since no differential entropy rate estimator has been developed that is accurate with low computational complexity. We generate simulated data from Fractional Gaussian Noise and ARFIMA(0,d,0) processes, and show that the current techniques either do not capture the complexity of LRD processes, or have high computational complexity. We make a link between the differential entropy rate and the Shannon entropy rate of a quantised version of a continuous valued process, showing that they differ by the logarithm of the size of the bins of the quantisation. Using this, we define an estimation technique, NPD-Entropy, which makes estimates of the Shannon entropy rate of a quantised version of the process, and adds the logarithm of the size of the bins for quantisation. This balances the two concerns, accuracy and time complexity, and provides estimates which are able to reflect the uncertainty of LRD processes and have lower computational complexity than known techniques. We then analyse theoretical properties of NPD-Entropy, and show that we are able to inherit useful statistical properties of estimators, such as consistency and bias, from the Shannon entropy rate estimator used.

In Chapter 7, we conclude the thesis by providing a summary of the results and discuss the potential extensions to this current work.

1.3 Publication List

Parts of this thesis have been published or submitted to journals.

1. **Feutrill, Andrew**, and Matthew Roughan. "A Review of Shannon and Differential Entropy Rate Estimation" Entropy 23.8 (2021): 1046.

2. **Feutrill, Andrew**, and Matthew Roughan. "Differential Entropy Rate Characterisations of Long Range Dependent Processes", `https://arxiv.org/abs/2102.05306`.

3. **Feutrill, Andrew**, and Matthew Roughan. "NPD Entropy: A Non-Parametric Differential Entropy Rate Estimator", `https://arxiv.org/abs/2105.11580`.

4. **Feutrill, Andrew**, and Matthew Roughan. "Convergence of Conditional Entropy for Long Range Dependent Markov Chains", `https://arxiv.org/abs/2110.14881`.

Chapter 2

Background

This thesis examines the concept of entropy and the entropy rate as a measure of uncertainty for stochastic processes. In particular, we will be using these concepts as a way of gaining additional insight into the concept of Long Range Dependence (LRD) when applied to continuous and discrete-valued stochastic processes in discrete time.

In this chapter, we begin by defining the concept of information entropy, Shannon and differential, *i.e.,* for discrete and continuous-valued random variables. Then we extend this concept to stochastic processes to define the *entropy rate*, the limit of the average new information per random variable in stochastic process. These are the information theoretic measures that will be used throughout the thesis and applied to analyse the behaviour of LRD processes.

We define LRD stochastic processes, and discuss their properties which have made traditional analysis difficult. We introduce two LRD continuous-valued examples, Fractional Gaussian Noise (FGN) and ARFIMA(p,d,q), which have been studied since their discovery and definition respectively by Mandelbrot and Van Ness [127] and independently by Hosking [87] and Granger and Joyeux [77]. We will illustrate that LRD is characterised by slow convergence to various quantities, such as the sample mean, and divergent sums of second order properties, such as the autocovariance function. The behaviour of information theoretic measures, when applied to LRD stochastic processes, is the theme of this thesis.

2.1 Entropy

Information entropy is a concept that was first introduced by Claude Shannon [154], post World War 2 where it was a building block of coding and

communication theory and spawned the field of information theory. We will begin this section by summarising the key concepts and results that will be utilised for both discrete and continuous random variables, and introduce the entropy rate, which will be used throughout this thesis as a measure of complexity or uncertainty of a stochastic process.

2.1.1 Shannon Entropy

Shannon entropy is a concept that was introduced by Claude Shannon [154] as a measure of uncertainty of a random variable. The motivation was to develop a robust measure of uncertainty, $H(p_1, \ldots, p_n)$ for finite set of probabilities $p_1, \ldots, p_n$, later generalised to countably infinite state spaces. The uncertainty measure is based on the following 3 properties [154]:

1. The function H should be continuous as a function of any individual probability p_i.

2. If all the p_i's are equal, *i.e.,* follow the uniform distribution, $p_i = \frac{1}{n}$, then the function should be monotonically increasing function of n. In other words, if there are more uniform choices there is more uncertainty.

3. If the uncertainty function, H, is decomposed into successive uncertainty functions, then the total H should be the weighted sum of H for both of the choices.

To interpret the last property, we provide the following example from Shannon [154]. Given a probability distribution, $p_1 = \frac{1}{2}, p_2 = \frac{1}{3}, p_3 = \frac{1}{6}$, at the top level can split this distribution into two uncertainty functions. We decompose this distribution into the uncertainty of p_1 with a probability $\frac{1}{2}$, or remaining two with probability $\frac{1}{2}$. Then the remaining probabilities, p_2 and p_3, are renormalised with probabilities $\frac{2}{3}$ and $\frac{1}{3}$. Then the decomposition of the function H is

$$H\left(\frac{1}{2}, \frac{1}{3}, \frac{1}{6}\right) = H\left(\frac{1}{2}, \frac{1}{2}\right) + \frac{1}{2} H\left(\frac{2}{3}, \frac{1}{3}\right).$$

These three properties lead to the following result, which characterises the only possible form of the uncertainty function, which is then used as the definition of information entropy.

Theorem 2.1.1 (Theorem 2 [154])**.** *The only function, H, that satisfies the probabilities given above is*

$$H = -K \sum_{i=1}^{n} p_i \log p_i,$$

where K is a positive constant. When $p_i = 0$, we define $0 \log 0 = 0$ due to a limiting argument and probability zero events do not influence the calculation.

Remark. *The proof is given in Appendix 2 of Shannon [154]. The constant K amounts to a choice of units, and we use the standard of choosing $K = 1$ and using $\log_2$ for discrete random variables to have the units of bits.*

Theorem 2.1.1 defines the Shannon entropy of a discrete random variable, and extend to a possibly infinite state space. For completeness we include the following definition of Shannon entropy.

Definition 2.1.1. *For a discrete random variable, X, with support on Ω, with a probability mass function $p(x)$, the Shannon entropy, $H(X)$ is defined as*

$$H(X) = -\sum_{x \in \Omega} p(x) \log p(x).$$

The units of entropy are dependent upon the choice of the base of the logarithm. It's common to use $\log_2$ in the case of a discrete random variable, and in this case the units are *bits*. Another common unit is the *nat* obtained from using the natural logarithm in Definition 2.1.1. The usage of nats is more common for a notion of entropy we will define later for continuous random variables. In this thesis we will use $\log_2$ for discrete random variables and the natural log for continuous random variables.

For example, the Bernoulli distributed random variable, $X \sim B(p)$, has two possible outcomes, 0 and 1, and a probability mass function of

$$\mathbb{P}(X = x) = \begin{cases} p, & \text{if } x = 1 \\ q = 1 - p, & \text{if } x = 0. \end{cases}$$

Then the entropy of the random variable, in bits, is

$$\begin{aligned} H(X) &= -\sum_{x \in \Omega} p(x) \log p(x), \\ &= -p \log p - (1 - p) \log(1 - p). \end{aligned}$$

We can see that the entropy of the random variable is completely characterised by the value of p. We have plotted the entropy of a Bernoulli random variable as a function of the probability, p, in Figure 2.1. To align this with the original inspiration we expect that this random variable should have the highest entropy, that is entropy is maximised when $p = \frac{1}{2}$.

A property of note here is that the uniform distribution over a finite number of elements represents the distribution with the highest entropy.

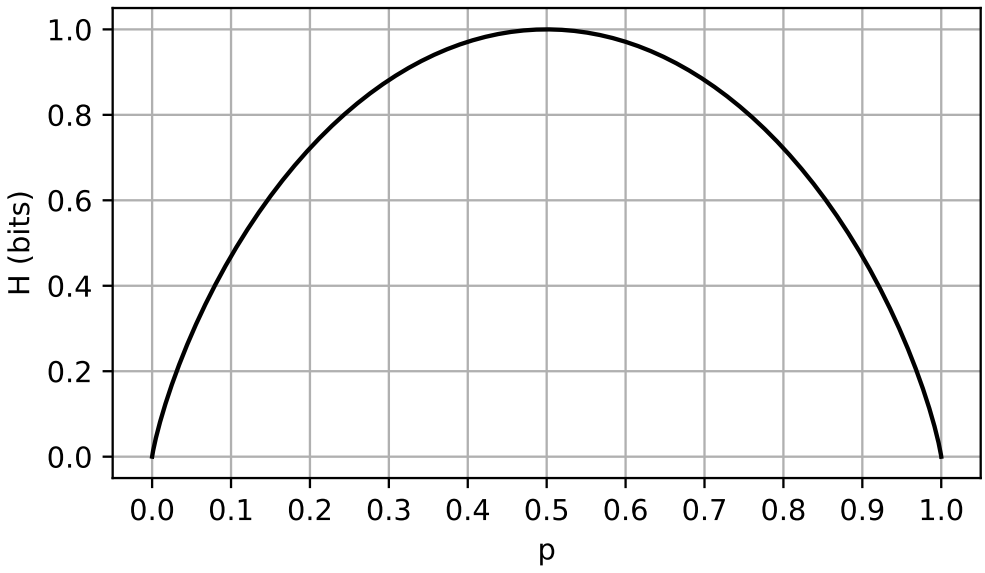

Figure 2.1: Entropy of a Bernoulli random variable as function of the probability, p. We can see that the function has the properties as defined by Shannon, and it is maximised when $p = \frac{1}{2}$.

Philosophically this aligns with the "Principle of Indifference" [103, pg. 45], attributed to Laplace, which states that given no other information all probabilities should be assigned to be equal, and the "Principle of Maximum Entropy" by Jaynes [96], which states that the distribution that best represents the current knowledge of system is the distribution which has the highest entropy. This principle inspired the maximum entropy approaches that were discussed earlier, with the Bernoulli random variable example.

To maximise the entropy function, we find inflection points and therefore we differentiate with respect to p and then solve for $\frac{dH}{dp} = 0$.

$$\begin{aligned}\frac{dH}{dp} &= -\log_2 p - p\left(\frac{1}{p\log(2)}\right) + \log_2(1-p) - (1-p)\left(\frac{-1}{(1-p)\log(2)}\right), \\ &= \log_2\left(\frac{1-p}{p}\right).\end{aligned}$$

Then we set this to 0 and solve for p, which gives

$$\log_2\left(\frac{1-p}{p}\right) = 0,$$

$$\iff \frac{1-p}{p} = 1,$$
$$\implies 2p = 1, \implies p = \frac{1}{2}.$$

It has a maximum at $p = \frac{1}{2}$, which aligns with our intuition for an uncertainty measure, since at this point it is the most difficult to predict the outcome. This is shown in Figure 2.1, where the maximum occurs at $p = \frac{1}{2}$ and decreases in either direction as p tends to 0 or 1. This concept is called maximum entropy estimation, and a large body of work in statistical inference of both models and parameters [109, pg. 36] [41, pg. 409], and optimisation where it is used as a dual to maximum likelihood approaches [15].

The result is that every distribution's uncertainty can be measured by its divergence from the uniform distribution. We make this idea more rigorous, for finite state spaces, with the following result.

Theorem 2.1.2 (Theorem 2.6.4 [41])**.** *The Shannon entropy, $H(X)$ is bounded by*

$$H(X) \leq \log N,$$

where N is the number of states in the state space of X. Equality occurs if and only if X is uniformly distributed.

We now extend the definition of Shannon entropy for a collection of multiple random variables, called the joint entropy. This is needed to define the concept of entropy we use for stochastic processes, entropy rate.

Definition 2.1.2. *For a collection of discrete random variables, $X_1, ..., X_n$, with support on, $\Omega_1, ..., \Omega_n$ and joint probability mass function $p(x_1, ..., x_n) = p(\mathbf{x})$, we define the joint entropy of the collection of random variables as,*

$$H(X_1, ..., X_n) = -\sum_{x_1 \in \Omega_1} ... \sum_{x_n \in \Omega_n} p(\mathbf{x}) \log p(\mathbf{x}).$$

Another important notion that we use is conditional entropy, $H(Y|X)$, which is the expected value of the entropy of a random variable Y, averaged over the knowledge of conditioning random variable X. Intuitively, the average amount of information of a random variable, Y, given we have knowledge of a value of a random variable, X.

Definition 2.1.3. *For random variables, X and Y with a joint probability mass function $(X, Y) \sim p(x, y)$, the conditional entropy is defined as*

$$H(Y|X) = -\sum_{x \in \Omega} \sum_{y \in \Omega} p(x, y) \log p(y|x).$$

If X and Y are independent random variables, then $H(Y|X) = H(Y)$.

The joint entropy and conditional entropy are complementary concepts, as the joint entropy of two random variables, X and Y, is equal to the entropy of a random variable $H(X)$ plus the conditional entropy of the other conditioned on the first, $H(Y|X)$. This is summarised in the following theorem [41, pg. 17].

Theorem 2.1.3 (Chain rule of entropy)**.**

$$H(X, Y) = H(X) + H(Y|X).$$

This result can be extended to an arbitrary collection of random variables, and we can calculate the joint entropy for a collection of random variables as the sum of conditional entropies. This is a very useful characterisation which is used in the proofs and calculation of other results.

Theorem 2.1.4. *Let $X_1, \ldots, X_n$ be a collection of random variables, then*

$$H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|H_{i-1}, \ldots, X_1).$$

Note that when $i = 1$ the contribution to the sum is $H(X_1)$.

A variety of entropic measures are defined using the conditional entropy, as we often want to quantify the uncertainty of one random variable knowing the value or distribution of another, or even more generally from a collection of observed random variables. A common uncertainty measure is called the relative entropy, commonly known as the Kullback-Leibler Divergence.

Definition 2.1.4. *The relative entropy, $D(p||q)$, between two probability mass functions $p(x)$ and $q(x)$ is defined as*

$$D(p||q) = \sum_{x \in \Omega} p(x) \log \frac{p(x)}{q(x)}.$$

This can be thought of as a measure of the difference between two probability mass functions, which are defined over the same state space. This quantity has some important properties, such as being always positive, by the Information Inequality [41, Theorem 2.6.3, pg. 28], and is equal to zero if and only if $p = q$. Relative entropy can be infinite if there exists any $x \in \Omega$ such that $p(x) > 0$, when $q(x) = 0$, since for this term $D(p||q) = p(x) \log \frac{p(x)}{q(x)} \to \infty$. Although relative entropy appears similar to a distance metric it is not symmetric, *i.e.,* $D(p||q) \neq D(q||p)$ in general.

Relative entropy is related to a metric called the Fisher Information metric, which is defined as

$$g_{ij}(\theta) = \mathbb{E}\left[\frac{\partial p(x,\theta)}{\partial \theta_j}\frac{\partial p(x,\theta)}{\partial \theta_k}\right],$$

for a set of coordinates $\theta = (\theta_1, \ldots, \theta_n)$, with a probability mass (or density) function $p(x, \theta)$. With the relationship given by

$$D(p(\theta + \delta)||p(\theta)) \approx \frac{\delta^2}{2} g(\theta)$$

[109, pg. 26]. The field of information geometry introduced by Amari [3, 4], analyses this metric to give insight into probability and statistics.

An extension of the relative entropy is the mutual information, which quantifies the amount of information that is shared between two random variables.

Definition 2.1.5. *The mutual information, $I(X;Y)$ between two random variables X and Y with joint probability mass function $p(x, y)$, marginal probability mass functions $p(x)$ and $p(y)$ respectively, is defined as*

$$\begin{aligned} I(X;Y) &= \sum_{x\in\Omega}\sum_{y\in\Omega} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}, \\ &= D(p(x,y)||p(x)p(y)). \end{aligned}$$

The mutual information is a useful concept in information theory, as it quantifies the difference between the joint mass function, $p(x, y)$, and the product of marginal mass functions $p(x)p(y)$. The mutual information has similar properties to the relative entropy, it is positive, and is zero if and only if $p(x, y) = p(x)p(y)$. Therefore, the mutual information quantifies how far the random variables X and Y are from independence.

To finalise this section we will present a theorem that summarises the links between mutual information and the joint and conditional entropy of two random variables.

Theorem 2.1.5. *The following equalities exist between mutual information $I(X;Y)$ and entropy H,*

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y), \\ I(X;Y) &= H(Y) - H(Y|X), \\ I(X;Y) &= H(X) + H(Y) - H(X,Y), \\ I(X;Y) &= I(Y;X), \\ I(X;X) &= H(X). \end{aligned}$$

This theorem summarises some important points, that mutual information is a symmetric uncertainty measure, that it quantifies the difference between the entropy of random variables, given knowledge of another, and that the mutual information of itself is the Shannon entropy, which justifies another name for Shannon entropy, the self information [41].

2.1.2 Differential Entropy

We will be considering entropic measures of stochastic processes composed of either discrete and continuous random variables in discrete time, so we will extend the concept of entropy to continuous-valued random variables. In Shannon's original work [154], he extended the definition to continuous random variables by considering the definition of Shannon entropy as the expected value of the information content, *i.e.*, $H(X) = -\mathbb{E}[\log(p(X))]$. There are many definitions and results of interest that are direct analogues of Shannon entropy. We have placed many of these definitions and results in Appendix B, but present the results that are directly used in future sections of the thesis.

Using this approach for the entropy of continuous random variables, with density $f(x)$, we define differential entropy.

Definition 2.1.6. *The differential entropy, $h(X)$ of a random variable, X, with support, Ω, and probability density function, $f(x)$, is,*

$$h(X) = \int_{\Omega} f(x) \log f(x) dx.$$

Differential entropy has some important properties which differ from the intuition we have developed for Shannon entropy. For example, differential entropy can be negative, or even diverge to $-\infty$. We see an example of this behaviour by considering the Dirac delta function, $\delta(x)$, *i.e.*, the unit impulse, defined by the properties $\delta(x) = 0$ for $x \neq 0$ and $\int_{-\infty}^{\infty} \delta(x)\, dx = 1$. The Dirac delta can be thought of, in terms of probability, as a completely determined point in time, that is, a function possessing no uncertainty. It can be constructed as the limit of rectangular pulses of constant area 1 as their width decreases, equivalent to the density of a uniform random variable, and hence we can calculate the differential entropy of the Dirac delta function as

$$\begin{aligned} h(X) &= -\int_{-a}^{a} \frac{1}{2a} \log\left(\frac{1}{2a}\right) dx, \\ &= \log(2a), \end{aligned}$$

which tends to $-\infty$ as $a \to 0$.

The intuition for $h(X) = -\infty$ from Cover and Thomas [41, pg. 248] is that the number of bits on average required to describe a continuous random variable, X to n-bit accuracy is $h(X) + n$, when using $\log_2$ for calculation of the differential entropy. Meaning $h(X) = -\infty$, can be read as requiring $n - \infty$ bits, so we can describe the random variable arbitrarily accurately without any using any bits of information.

Finally in this section we will discuss and prove a result that links the differential entropy of a random variable and the Shannon entropy of its quantisation. This is an important connection and we will extend this connection to stochastic processes to form the basis of an estimation technique in Chapter 5

Theorem 2.1.6 ([41, pg. 247])**.** *For a Riemann integrable density, $f(x)$, of a random variable, X, with an associated quantised random variable,*

$$X^{\Delta} = x_i \text{ if } i\Delta \leq X < (i+1)\Delta,$$

for a partition of the range of X into bins of size Δ. Then as $\Delta \to 0$,

$$H\left(X^{\Delta}\right) + \log \Delta \to h(X).$$

Proof. The probability that $X^{\Delta} = x_i$ is given by

$$\begin{aligned} p_i &= \int_{i\Delta}^{(i+1)\Delta} f(x)dx, \\ &= \Delta f(x_i). \end{aligned}$$

Then we can calculate the Shannon entropy of the quantised random variable, X^{Δ}, as

$$\begin{aligned} H(X^{\Delta}) &= -\sum_{i=-\infty}^{\infty} p_i \log p_i, \\ &= -\sum_{i=-\infty}^{\infty} \Delta f(x_i) \log \left(\Delta f(x_i)\right), \\ &= -\sum_{i=-\infty}^{\infty} \Delta f(x_i) \log f(x_i) - \log \Delta \sum_{i=-\infty}^{\infty} \Delta f(x_i), \\ &= -\sum_{i=-\infty}^{\infty} \Delta f(x_i) \log f(x_i) - \log \Delta. \end{aligned}$$

Where the last term simplifies since $\sum \Delta f(x_i) = \sum p_i = 1$. Now by the Riemann integrability of $f(x)$, this implies that $f(x) \log f(x)$ is Riemann integrable. Taking the limit as $\Delta \to 0$, gives that the first term above becomes $-\int f(x) \log f(x) dx = h(X)$, and the result follows. □

Note that the limit as $\Delta \to 0$ means that we are taking the limit of the expression as the window size decreases to zero. This applies throughout the thesis.

2.1.3 Entropy Rate

In this section we define an entropic measure that can be applied to stochastic processes, the entropy rate. This is defined as the asymptotic value of the average information per sampled random variable in a stochastic process.

Throughout this thesis we will use the entropy rate, both Shannon and differential, as a measure of complexity or randomness contained in a stochastic process. We leave the definition and discussion of differential entropy rate to Appendix B, however it is a natural analogue replacing joint Shannon entropy with joint differential entropy.

We will define and discuss the entropy rate in terms of the Shannon entropy, applying to random variables on discrete state spaces.

Definition 2.1.7. *For a discrete-valued, discrete-time stochastic process, $\chi = \{X_i\}_{i \in \mathbb{N}}$, the entropy rate, is defined as,*

$$H(\chi) = \lim_{n \to \infty} \frac{1}{n} H(X_1, ..., X_n),$$

where the limit exists.

The entropy rate does not exist for many stochastic processes, as it requires the existence of a limit.

We will not be giving definitions and background for stochastic processes, a thorough treatment can be found in Cinlar [34], however for the purpose of discussion we will need to define some properties. In general, the processes we will be analysing in this thesis will have the property of stationarity. Intuitively, this corresponds to the time invariance of the marginal distribution, and simplifies the analysis.

Definition 2.1.8. *For a stochastic process, $\chi = \{X_i\}_{i \in \mathbb{N}}$, with a cumulative distribution function at times, $t_1, \ldots, t_n$, of, $F_X(t_1, \ldots, t_n)$. We say that the process is stationary if $F_X(t_1 + \tau, \ldots, t_n + \tau) = F_X(t_1, \ldots, t_n), \forall \tau > 0, t_1, \ldots, t_n \in \mathbb{R}$, and $\forall n \in \mathbb{N}$.*

This property is helpful for finding the entropy rate of a stochastic process, as it implies that the distribution doesn't vary with time shifts, which is sufficient for the existence of the limit [41, pg. 77]. Stationarity allows another characterisation of the entropy rate of a stochastic process, as the limit of the conditional entropy of the process. We will provide a proof, since we use characterisation of the entropy rate for stationary processes to analyse the convergence rates of the conditional entropy to the entropy rate for both Gaussian processes and Markov chains. In the following proof, one of the implications is that the entropy rate must exist for stationary processes, which intuitively makes sense as the distribution will not change asymptotically, and hence a limit can be achieved.

Theorem 2.1.7 (Theorem 4.2.1 [41])**.** *For a stationary stochastic process, the entropy rate exists and is equal to*

$$H(\chi) = \lim_{n\to\infty} H(X_n|X_{n-1},\dots,X_1).$$

Proof. We begin by showing that the conditional entropy is non-increasing as n increases.

$$\begin{aligned} H(X_{n+1}|X_n,\dots,X_2,X_1) &\leq H(X_{n+1}|X_n,\dots,X_2), \\ &\leq H(X_n|X_{n-1},\dots,X_1). \end{aligned}$$

The first inequality follows since conditioning cannot increase entropy, and the second follows from the stationarity of the process. Now if we take the limit of $H(X_{n+1}|X_n,\dots,X_2,X_1)$ as $n\to\infty$, this limit exists since we have a decreasing sequence of non-negative numbers, by the monotone convergence theorem [149, Theorem 11.28]. By the chain rule of entropy, Theorem 2.1.3, we have

$$\frac{H(X_1,\dots,X_n)}{n} = \frac{1}{n}\sum_{i=1}^{n} H(X_i|X_{i-1},\dots,X_1).$$

Since we have shown that the conditional entropy has a limit as $n\to\infty$, this implies that the average converges to the same limit, as convergence of the sequence implies convergence of the Cesaro mean to the same limit [189, pg. 76]. Therefore by taking the limits of both sides we have

$$\begin{aligned} H(\chi) &= \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} H(X_i|X_{i-1},\dots,X_1), \\ &= \lim_{n\to\infty} H(X_n|X_{n-1},\dots,X_1). \end{aligned}$$

□

We will briefly give an example of calculating the entropy rate for Markov chains, which have the property that

$$\mathbb{P}(X_n = x_n | X_{n-1} = x_{n-1}, \ldots, X_1 = x_1) = \mathbb{P}(X_n = x_n | X_{n-1} = x_{n-1}).$$

We will discuss Markov chains in greater detail in Section 4, however this property is enough to illustrate the behaviour of interest. We call the probabilities, $p_{ij} = \mathbb{P}(X_n = j | X_{n-1} = i)$, the transition probabilities of the Markov chain and the stationary distribution is, $\pi_i = \mathbb{P}(X_n = i)$, when it exists. Hence for a stationary Markov chain

$$H(X_n | X_{n-1}, \ldots, X_1) = H(X_n | X_{n-1}).$$

Then we calculate the entropy rate for a stationary Markov chain as

$$\begin{aligned} H(\chi) &= \lim_{n\to\infty} H\left(X_n | X_{n-1}, \ldots, X_1\right), \\ &= \lim_{n\to\infty} H\left(X_n | X_{n-1}\right), \\ &= H(X_2 | X_1), \\ &= -\sum_{i\in\Omega}\sum_{j\in\Omega} \pi_i p_{ij} \log p_{ij}. \end{aligned}$$

Where we have used the definition of conditional entropy and the Markov property to simplify the calculation. In Chapter 4 we will look to analyse the entropy rate of a Markov chain that converges to stationarity, but starts in an initial state given an arbitrary initial distribution.

To quantify the difference between the past and future of a stochastic process, we will expand the definition of mutual information, Definition 2.1.5, to define a quantity called the mutual information between past and future. This quantifies the shared information between the infinite past $\{X_s, s < 0\}$ and infinite future $\{X_s, s \geq 0\}$. We will start by defining this in full generality, as this forms the basis of some interesting quantities relating to the entropy rate.

Definition 2.1.9. *The mutual information between past and future for lag, τ is defined as*

$$I^{(\tau)} = I(\{X_s, s < t\}, \{X_s, s \geq t + \tau\}).$$

Remark. *The definition of mutual information between past and future is identical in the case of differential entropy, substituting mutual information for differential entropy.*

This is tied to the concept of information regularity of a process [90, Theorem IV.6], which has been defined as stochastic processes for which $I^{(\tau)} \to 0$ as $\tau \to \infty$ [119]. Then we define the mutual information between past and future as the mutual information between past and future for 0 lags, $I_{\text{p-f}}$,

$$I_{\text{p-f}} = I(\{X_s, s < t\}, \{X_s, s \geq t\}).$$

Some connections between this concept and the finiteness of sums incorporating the partial autocorrelation function will be explored in Chapter 3.

The next quantity we will define is the excess entropy, which has been formulated as the infinite sum of the differences between the conditional entropy and the entropy rate, as we observe more random variables, and therefore gives characterisation of the rate of convergence of the conditional entropy to the entropy rate.

Definition 2.1.10. *The excess entropy, E, of a stochastic process, $\chi = \{X_i\}_{i \in \mathbb{N}}$, is defined as*

$$E = \sum_{n=1}^{\infty} \left(H_e(n) - H(\chi) \right),$$

where

$$\begin{aligned} H_e(n) &= H(X_1, \ldots, X_n) - H(X_1, \ldots, X_{n-1}), \\ &= H(X_n | X_{n-1}, \ldots, X_1). \end{aligned}$$

Remark. *The definition of excess entropy is identical in the case of continuous valued processes, by replacing the differential entropy in place of Shannon entropy.*

This quantity is discussed extensively in Crutchfield and Feldman [44], who give a few different characterisations. However, the key intuition is the excess or additional entropy that is accumulated in the convergence of the conditional entropy to the entropy rate. Hence, it is a measure of the rate of convergence of the conditional entropy to the entropy rate. We present a useful characterisation of the excess entropy as a limit.

Theorem 2.1.8 (Proposition 7 [44])**.** *The excess entropy can be written as*

$$E = \lim_{n \to \infty} \left[H(X_1, \ldots, X_n) - nH(\chi) \right].$$

The theorem shows that the excess entropy is the limit of the difference between the joint entropies and n copies of the entropy rate. Next we will present a result that links the mutual information between past and future and the excess entropy, showing that the two concepts are in fact equivalent. The amount of shared information between the past and future of a process is equal the accumulated information when converging from the conditional entropy to the entropy rate.

Theorem 2.1.9 (Proposition 8 [44])**.** *The excess entropy is equal to the mutual information between past and future*

$$E = I_{\text{p-f}},$$

when the limit exists.

Remark. *Note that this statement only applies in the discrete case, and the proof given by Crutchfield and Feldman is not fully rigorous. We will prove this more rigorously in Chapter 3.*

This connection between mutual information and excess entropy will be used later in this thesis, as we want to characterise the amount of shared information between past and future, and the excess entropy provides an alternate way to analyse the quantity.

2.2 Long Range Dependence

Long range dependence (LRD) refers to a process where influence of past values persists over long time periods. In mathematical terms, intuitively we think of this as the slow decay of the autocovariance function, which we define below. We will introduce some common definitions and characterisations of LRD in this section, and then discuss some common models which exhibit this type of behaviour. Some further discussion of LRD, in particular discussion of the influence of self-similarity and non-stationarity in included in Appendix C.

Definition 2.2.1. *The autocovariance function of a stationary stochastic process, $\{X_i\}_{i\in\mathbb{Z}^+}$, is defined as*

$$\gamma(k) = \mathbb{E}\left[(X_n - \mu)\,(X_{n+k} - \mu)\right].$$

We define the autocorrelation coefficient for k lags as

$$\rho(k) = \frac{\gamma(k)}{\gamma(0)} = \frac{\gamma(k)}{\sigma^2} = Corr(X_{k+n}, X_n),$$

where σ^2 is the variance of a random variable of the process.

Remark. *Note that this is easily confused with the autocorrelation function of a stochastic process, which is defined as*

$$R(k) = \mathbb{E}\left[X_n X_{n+k}\right].$$

Throughout this thesis we will be using the autocorrelation coefficient, $\rho(k)$ and not the autocorrelation function, $R(k)$, when analysing stochastic processes.

Another quantity of interest that measures the dependence structure of a stochastic process is the partial autocorrelation function.

Definition 2.2.2. *The partial autocorrelation function of a stationary stochastic process, $\{X_i\}_{i\in\mathbb{Z}^+}$, is defined as*

$$\begin{aligned}&\alpha(1) = \rho(1),\\ &\alpha(n) = Corr(X_{k+n} - P_{X_n,\ldots,X_{k+n-1}}X_{k+n}, X_n - P_{X_n,\ldots,X_{k+n-1}}X_n), \forall n \geq 2,\end{aligned}$$

where $P_{X_n,\ldots,X_{k+n-1}}$ is the linear projection onto the space spanned by the intermediate observations.

The partial autocorrelation function, more intuitively, is the autocorrelation coefficient between two observations when removing the linear dependence of the observations between them.

A related quantity of a stochastic process is the spectral density, which represents the distribution of frequencies within the process. We will use a theorem from the time series literature to form a definition and show the relationship between the spectral density and the autocorrelation function.

Theorem 2.2.1 (Corollary 4.3.1 (i) [23])**.** *A complex-valued function, $\gamma(k)$, is the autocovariance function of a discrete-time stationary stochastic process, $\{X_i\}_{i\in\mathbb{Z}}$ if and only if $\forall k \in \mathbb{Z}$, we have*

$$\gamma(k) = \int_{-\pi}^{\pi} e^{ik\lambda} dF(\lambda),$$

where F is a right continuous, non-decreasing, bounded function on the real interval $[-\pi, \pi]$ with $F(-\pi) = 0$.

We call the function, F, from this theorem the spectral distribution and if it is differentiable everywhere then we define the spectral density as the function as follows.

Definition 2.2.3. *The spectral density of an autocovariance function, $\gamma(k)$, is defined $\forall \lambda \in [-\pi, \pi]$, as*

$$f(\lambda) = \frac{1}{2\pi} \sum_{n=-\infty}^{\infty} e^{-in\lambda} \gamma(n).$$

Note that the spectral density is non-negative for all $\lambda \in [-\pi, \pi]$.

Theorem 2.2.1 gives the autocovariance in terms of the spectral density as

$$\gamma(k) = \int_{-\pi}^{\pi} e^{ik\lambda} f(\lambda) d\lambda.$$

We can see that autocovariance and the spectral density are related through the Fourier Transform [23, pg. 117], and are equivalent via the Kolmogorov Isomorphism Theorem [19], and hence these are both characterisations of the correlation structure of a stochastic process. Analysis via the spectral density allows for a larger number of techniques to study stochastic processes, and we will use spectral approaches to investigate LRD in this thesis.

We now define LRD in two equivalent ways, via the autocorrelation and spectral density. The following statement defines the concept of LRD in terms of its autocorrelation function, which is the most common approach to defining the phenomenon.

Definition 2.2.4. *Let $\{X_n\}_{n\in\mathbb{N}}$ be a stationary process. If there exists $\alpha \in (0, 1)$, and $c_\gamma > 0$, such that the auto-covariance $\gamma(k)$ satisfies*

$$\lim_{k\to\infty} \frac{\gamma(k)}{c_\gamma k^{-\alpha}} = 1,$$

then we say that the process is long range dependent.

The definition of LRD in the frequency domain considers the limit of the spectral density near the origin.

Definition 2.2.5. *Let $\{X_n\}_{n\in\mathbb{N}}$ be a stationary process. If there exists $\beta \in (0, 1)$, and $c_f > 0$, such that the spectral density $f(\lambda)$ satisfies*

$$\lim_{\lambda\to 0} \frac{f(\lambda)}{c_f |\lambda|^{-\beta}} = 1,$$

then we say that the process is long range dependent.

This perspective tells us that these processes are dominated by the low frequencies, which corresponds to long wavelengths and hence the long range correlations. Note that both the definitions result in an asymptotic power-law of their respective functions, and asymptotic power-laws are a constant theme in the LRD literature.

In Hurst's original work, he was aiming to estimate the size of a dam for the Nile River and was analysing a time series of the Nile River annual minima [150]. He calculated a quantity named the rescaled range, $\frac{R}{S}(X_1, \ldots, X_n)$, which operates on n observations, $x_1, \ldots, x_n$. This is defined as

$$\frac{R}{S}(X_1, \ldots, X_n) = \frac{\max_{0 \leq i \leq n} \left(S_i - \frac{i}{n} S_n\right) - \min_{0 \leq i \leq n} \left(S_i - \frac{i}{n} S_n\right)}{\sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(X_i - \frac{S_n}{n}\right)}}.$$

In this definition $S_n = X_1 + \ldots + X_n$ is the partial sum sequence, and therefore $\frac{S_n}{n}$ is the sample mean. The numerator quantifies how far the partial sums deviate from the uniform growth, *i.e.,* the first term is measuring the maximum difference in the index between the observed value and the scaled partial sum. The denominator is the sample standard deviation and provides a normalisation.

A typical assumption that was previously used in reservoir management was that the yearly outflow could be well modelled by

$$X_i = \mu + \epsilon_i,$$

where μ is the observed average and ϵ_i is i.i.d. white noise such that, $\mathbb{E}[\epsilon_i] = 0$ and finite variance [129]. An argument in Samorodnitsky [150, pg. 177], shows that under assumptions of short range correlations that $\frac{R}{S}(X_1, \ldots, X_n)$ grows at the rate of $n^{\frac{1}{2}}$. This highlights the unexpected outcome of Hurst's empirical analysis of the Nile river minima, given the start of the art theory at the time.

First we will introduce Gaussian processes, since many of the processes studied in this thesis are of this type. Then we define a short range dependent Gaussian process to show why the assumption of growth of the rescaled range was expected to be at the rate of $n^{\frac{1}{2}}$.

Definition 2.2.6. *A stochastic process is called a Gaussian process if and only if every finite collection of random variables has a multivariate Gaussian distribution. That is for every $t_1, \ldots, t_k \in \mathbb{R}$,*

$$(X_{t_1}, \ldots, X_{t_k}) \sim \mathcal{N}(\mu, \mathbf{\Sigma}),$$

where μ is the vector of expected values and Σ is the covariance matrix.

An important Gaussian process is Brownian Motion, which has applications across many areas of probability theory.

Definition 2.2.7. *We define Brownian Motion,* $B = \{B(t)\}_{t\geq 0}$*, by the following properties:*

1. $B(0) = 0$.
2. *B has independent increments,* i.e., $\forall t > 0$*, the increments* $B(t+u) - B(t), \forall u \geq 0$*, are independent of* $B(s), \forall s \leq t$.
3. *B has Gaussian increments,* i.e., $B(t+u) - B(t) \sim \mathcal{N}(0, u)$*, is normally distributed.*
4. *The sample paths of B are continuous almost surely.*

Note that the process is non-stationary since the covariance function isn't time invariant. Brownian Motion is not long range dependent, however it will be used throughout this thesis because Brownian Motion, and its increment process Gaussian Noise, form the basis of processes used in this thesis, Fractional Brownian Motion and Fractional Gaussian Noise.

Hurst found in his investigation of the Nile flooding that the rescaled range grows at the rate of $n^{0.72}$, which differed from the expected value of $n^{\frac{1}{2}}$ for short range correlations [89]. For example, the rescaled range of Brownian motion grows at the rate of $n^{1/2}$ [151, pg. 176]. Hurst then found this phenomena in a range of physical time series, such as sunspots, rainfall and atmospheric pressure [80]. This lead to the definition Hurst parameter, $\mathcal{H}$, as the exponent of the rescaled range, *i.e.,* $n^{\mathcal{H}}$. However, we will define it equivalently with respect to the exponent α in Definition 2.2.4 as

$$\mathcal{H} = 1 - \alpha/2.$$

Some intuition for the higher value of $\mathcal{H}$ for long memory processes, is that the longer trends will cause higher deviations of the partial sums, causing higher than expected growth.

The following theorem shows that Definition 2.2.4 and Definition 2.2.5 are equivalent, with exponents in terms of the Hurst parameter.

Theorem 2.2.2 (Theorem 2.1 [14])**.**

1. *If Definition 2.2.4 holds with* $0 < \alpha = 2 - 2\mathcal{H} < 1$*. Then the spectral density,* $f(\lambda)$*, exists and*

$$\lim_{\lambda \to 0} \frac{f(\lambda)}{c_f |\lambda|^{1-2\mathcal{H}}} = 1,$$

where

$$c_f = \frac{c_\rho \sigma^2}{\pi} \frac{\Gamma(2\mathcal{H}-1)}{\sin(\pi - \pi\mathcal{H})}.$$

2. *If Definition 2.2.5 holds with* $0 < \beta = 2\mathcal{H} - 1 < 1$. *Then*

$$\lim_{k\to\infty} \frac{\gamma(k)}{c_\gamma k^{2\mathcal{H}-2}} = 1,$$

where

$$c_\rho = \frac{2c_f\Gamma(2-2\mathcal{H})}{\sin(\pi\mathcal{H} - \frac{1}{2}\pi)\sigma^2}.$$

Since exponents of LRD processes are in the range $\alpha \in (0,1)$ this implies that $\mathcal{H}$ must be in the range $(0.5,1)$. To summarise the behaviour of processes by their Hurst parameter: the region $\mathcal{H} > 1/2$ exhibits LRD, also known as persistent, in the time series literature, and $\mathcal{H} < 1/2$ for negatively correlated processes, also known as anti-persistent; and $\mathcal{H} = 1/2$ for short range correlated processes, *e.g.*, Brownian Noise. An important consequence of Definition 2.2.4 and Theorem 2.2.2, is on the summability of the correlation function, namely:

$$\sum_{k=-\infty}^{\infty} \gamma(k) \begin{array}{ll} < \infty, & \text{for } 0 < \mathcal{H} \leq \frac{1}{2}, \\ = \infty, & \text{for } \frac{1}{2} < \mathcal{H} < 1. \end{array}.$$

This is often presented as the definition of LRD in other works. Sometimes a weaker definition, $\sum_{k=-\infty}^{\infty} |\gamma(k)| = \infty$, is given as the definition of LRD [151, pg. 194].

Negatively correlated processes, $\mathcal{H} < 1/2$, have not received as much consideration as the SRD and LRD cases, due to fewer practical applications. They have many properties in common with short range dependent processes as they still have a summable autocorrelation function [14, 72]. However, in addition their structure enforces that $\sum_{k=-\infty}^{\infty} \gamma(k) = 0$ [14, pg. 52]. This is quite a strict and surprising property, and hence these processes have been called constrained short range dependent (CSRD) [72].

A commonly analysed self-similar process is a generalisation of Brownian Motion, called Fractional Brownian Motion (FBM). Its increment process, Fractional Gaussian Noise, is one of the fundamental processes studied in this thesis. We define FBM by its covariance function, since it is a Gaussian process the mean and covariance are sufficient to completely characterise the stochastic process [101, pg. 103].

Definition 2.2.8. *Fractional Brownian Motion, $B_{\mathcal{H}} = \{B_{\mathcal{H}}(t)\}_{t\geq 0}$, is defined by the following properties:*

1. $B_{\mathcal{H}}(0) = 0$.
2. $\mathbb{E}[B_{\mathcal{H}}(t)] = 0, \forall t > 0$.
3. *The covariance function is* $\mathbb{E}[B_{\mathcal{H}}(t)B_{\mathcal{H}}(s)] = \frac{1}{2}\left(t^{2H} + s^{2\mathcal{H}} - |t-s|^{2\mathcal{H}}\right)$.
4. *The sample paths of B are continuous almost surely.*

FBM is also a non-stationary process, due to the time dependence of the covariance function. Note that this process is generally not considered to be LRD, due to its non-stationarity [150]. In contrast to Brownian motion the non-overlapping increments of the process are not independent, due to the long memory. The increments of the process are stationary since we can see that for any $r > 0$,

$$\mathbb{E}[(B_{\mathcal{H}}(t+r) - B_{\mathcal{H}}(r))\,(B_{\mathcal{H}}(s+r) - B_{\mathcal{H}}(r))] = \mathbb{E}[B_{\mathcal{H}}(t)B_{\mathcal{H}}(s)],$$

by expanding and applying the covariance function. The increments are correlated as we can see for $t_1 < t_2 \leq t_3 < t_4$, that

$$\mathbb{E}[B_{\mathcal{H}}(t_4 - t_3)B_{\mathcal{H}}(t_2 - t_1)] = \frac{1}{2}\left((t_4 - t_3)^{2\mathcal{H}} + (t_2 - t_1)^{2\mathcal{H}} - |t_4 - t_3 - (t_2 - t_1)\,|^{2\mathcal{H}}\right).$$

Then we can conclude that the correlations of the increments are positive for $\mathcal{H} > \frac{1}{2}$, negatively for $\mathcal{H} < \frac{1}{2}$ and independent in the case $\mathcal{H} = \frac{1}{2}$, by the sign of the expression above [18, pg. 9].

Sample paths of fractional Brownian motion are shown in Figure 2.2 for a range of Hurst parameters, $\mathcal{H} = 0.2, 0.5, 0.8$. The figure shows the influence of the Hurst parameter with the appearance of long periods of upwards and downwards "trends". The negatively correlated process has a smaller range of movement and less smooth paths from the frequent direction shifts. For $\mathcal{H} = 0.5$ we have the regular Brownian motion discussed above, with no correlation between disjoint increments.

The phenomenon of LRD has large effects on even the most fundamental statistics that are used to analyse data. A common result in statistics concerns the size of the variance of the sample mean, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$, and how it converges to the true value with an increasing number of observations, n. This can be summarised in the following result.

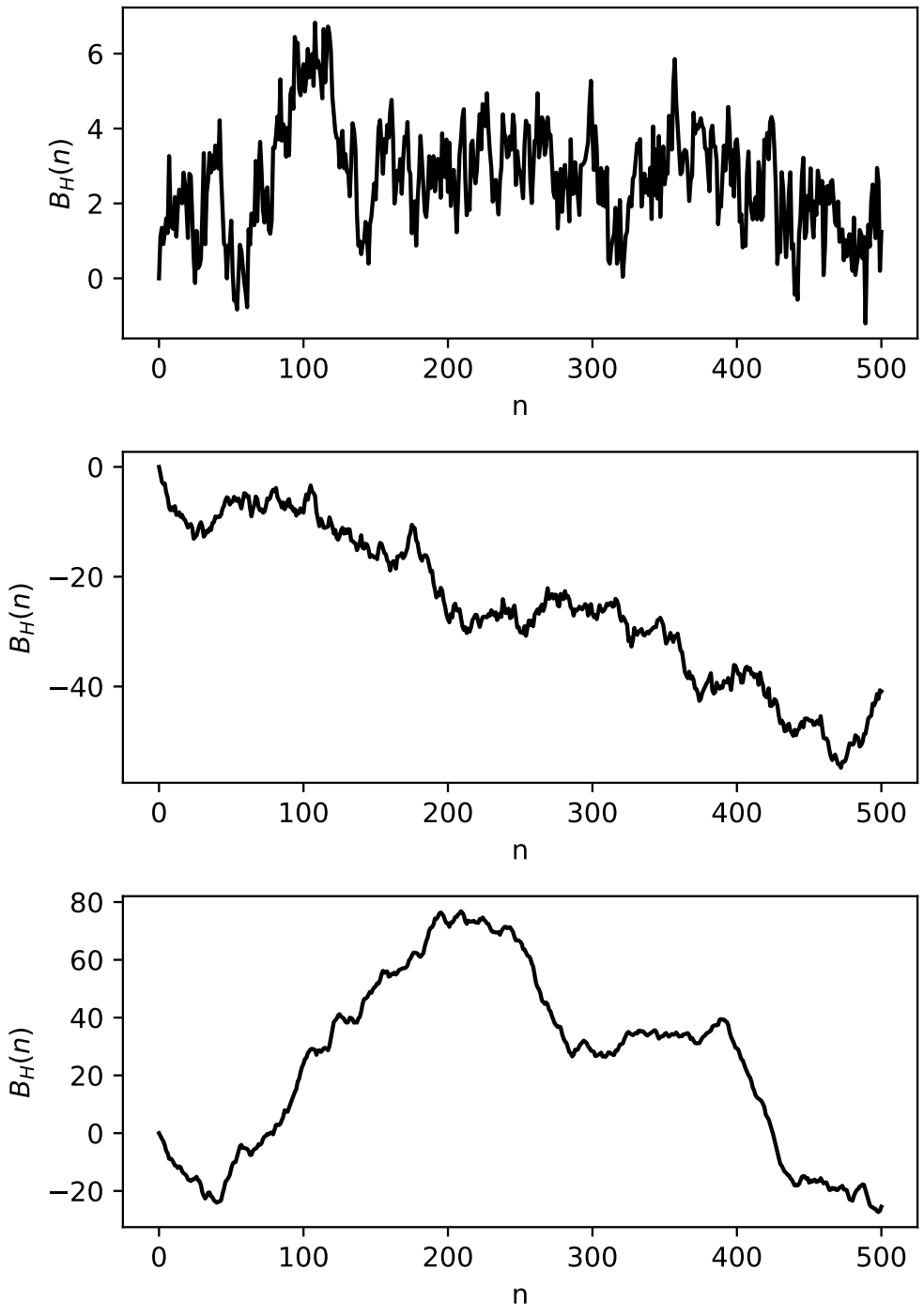

Figure 2.2: Sample paths of Fractional Brownian Motion with Hurst parameters, $\mathcal{H} = 0.2, 0.5, 0.8$. We see that the long range dependence results in smoother paths with lager local trends. In contrast, the negatively correlated paths occur over a much smaller range due to the lower likelihood of trends. These realisations were generated using the *fbm* Python package [66], using the Davies-Harte method [49].

Theorem 2.2.3 (1.1 [14]). *For a sample of n observations of an i.i.d. stochastic process, $\{X_i\}_{i\in\mathbb{Z}}$, with common mean, $\mathbb{E}[X_i] = \mu$ the common variance of the observations is $Var(X_i) = \sigma^2 < \infty$, then*

$$Var(\bar{X}_n) = \sigma^2 n^{-1}.$$

This is a foundational result in statistics, and is used to calculate the uncertainty in parameter estimates. This is linked to another fundamental result in statistics, the central limit theorem, which states that the distribution of the sample mean of n independent and identically distributed observations has a normal distribution in the limit, *i.e.,* for $\infty < x < \infty$,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n}{\sigma\sqrt{n}} \leq x\right) = \Phi(x),$$

where $\Phi(x)$ is the cumulative distribution function of a normally distributed random variable and σ is the standard deviation [146, pg. 161].

The rate of convergence of the variance of the sample mean generalises from i.i.d. data to short range dependent processes, such as ARMA processes and Gaussian Noise, with the variance becoming, $\sigma^2 c(\rho) n^{-1}$, *i.e.,* differing from the i.i.d. case by a term that is a function of the correlations [14, pg. 5]. Note that this does not change the rate of convergence of the variance of the sample mean [14, pg. 5]. However, for LRD processes this is not necessarily true, and any conclusion on the rate of convergence of the sample variance would be an underestimate, as the following result for Fractional Gaussian Noise, which we define in the next section shows [13].

Theorem 2.2.4. *For a sample of n observations of a Fractional Gaussian Noise process, $\{X_i\}_{i\in\mathbb{Z}}$, where the variance of the observations is $Var(X_i) = \sigma^2 < \infty$, then*

$$Var(\bar{X}_n) = \sigma^2 n^{2\mathcal{H}-2},$$

where $\mathcal{H}$ is the Hurst parameter.

Therefore, for processes exhibiting LRD we have to be extremely careful when we are applying common statistical techniques. However, the influence of LRD is more expansive than this simple example illustrates. We present, in Appendix C, some other behaviour that illustrates the quantitative differences in how LRD stochastic processes behave.

This type of result, a qualitative change in behaviour from short range to long range dependence gives an alternate perspective on LRD, as a phase transition between regimes with different properties, as described by Samorodnitsky [151]. We will discuss this perspective in Appendix C with some results

that illustrate this phase transition, such as the behaviour of partial sums and maxima. In this thesis we will illustrate that a phase transition also occurs in the entropic domain for classes of LRD and CSRD processes.

In the next sections we will introduce two common stationary LRD models, Fractional Gaussian Noise (FGN) and ARFIMA(p,d,q). These were developed as extensions to common probabilistic models, Gaussian Noise and ARMA/ARIMA classes of processes. They are parsimonious models that use very few parameters to induce the long memory properties we have discussed. These two models will form the bulk of the discussion and the examples throughout the thesis. They are both discrete-time models and in general we will not discuss continuous-time models, except where they form relevant examples to discuss certain properties. Later in Chapter 4, we will extend some results and characterisations to LRD Markov processes.

2.2.1 Fractional Gaussian Noise

We will begin by defining Fractional Gaussian Noise, which was introduced first by Mandelbrot and Van Ness [127] to describe and explain Hurst's empirical findings. We saw in the previous section that FBM is not a stationary process. However, its increments are stationary, and these increments are the definition of Fractional Gaussian Noise.

Definition 2.2.9. *We define discrete-time Fractional Gaussian Noise (FGN), $X(n)$, $\forall n \in \mathbb{Z}^+$ as*

$$X(n) = B^{\mathcal{H}}(n) - B^{\mathcal{H}}(n-1).$$

We calculate the autocovariance function of FGN as

$$\mathbb{E}\left[X_{n+k}X_n\right] = \mathbb{E}\left[\left(B^{\mathcal{H}}(n+k) - B^{\mathcal{H}}(n+k-1)\right)\left(B^{\mathcal{H}}(n) - B^{\mathcal{H}}(n-1)\right)\right],$$

since FBM has zero expectation, and hence FGN also has zero expectation. Applying the covariance function in Definition 2.2.8 to the individual terms and cancelling, we get

$$\gamma(k) := \mathbb{E}\left[X_{n+k}X_n\right] = \frac{\sigma^2}{2}\left((k+1)^{2\mathcal{H}} - 2k^{2\mathcal{H}} + (k-1)^{2\mathcal{H}}\right).$$

This shows that the covariance is only dependent on the time between observations, and since this is a Gaussian process with constant mean and covariance only dependent on the time between observations, then the process is stationary [23, pg. 13]. Then the autocorrelation function is

$$\rho(k) = \frac{1}{2}k^{2\mathcal{H}}g(k^{-1}),$$

$$\text{where } g(x) = (1+x)^{2\mathcal{H}} - 2 + (1-x)^{2\mathcal{H}}.$$

For $\mathcal{H} \neq \frac{1}{2}$ we take the Taylor series expansion at the origin and the first non-zero term is $2\mathcal{H}(2\mathcal{H}-1)x^2$. Taking the limit as $k \to \infty$ we have

$$\rho(k) \to \mathcal{H}(2\mathcal{H}-1)k^{2\mathcal{H}-2}.$$

Therefore, we can see by Definition 2.2.4 that the parameter range $\frac{1}{2} < \mathcal{H} < 1$, FGN is LRD. Another note is that in the case of $\mathcal{H} = \frac{1}{2}$ that the autocorrelation function is $\rho(k) = 0$, and therefore Gaussian Noise is a special case of FGN with $\mathcal{H} = \frac{1}{2}$.

Figure 2.3 shows some sample paths of FGN for a range of different Hurst parameters. These are derived from the same sample paths of FBM given in Figure 2.2. We can see that longer trends emerge when the Hurst parameter is greater than $1/2$, and the negative correlation of lower $\mathcal{H}$ values results in a process that rapidly shifts between positive and negative values.

The following theorem from Beran [14] derives the spectral density of FGN, based on initial work on self-similar processes [157], defined in Definition C.0.1.

Theorem 2.2.5 (Proposition 2.7 [14]). *The spectral density, $f(\lambda)$, of Fractional Gaussian Noise is given for $\lambda \in [-\pi, \pi]$ to be*

$$f(\lambda) = 2c_f(1-\cos\lambda)\sum_{j=-\infty}^{\infty}|2\pi j + \lambda|^{-2\mathcal{H}-1},$$

where $c_f = \frac{\sigma^2}{2\pi}\sin(\pi\mathcal{H})\Gamma(2\mathcal{H}+1)$ and σ^2 is the variance of a random variable of the process.

Similar to the approach of the covariance function, taking a Taylor series expansion around the origin yields the following corollary.

Corollary 2.2.5.1 (Corollary 2.1 [14]). *For FGN with Hurst parameter $\mathcal{H}$*

$$f(\lambda) \sim c_f|\lambda|^{1-2\mathcal{H}}, \text{ as } \lambda \to 0.$$

Therefore, we can see that FGN has the required properties in the spectral domain. This concludes our introduction to FGN, we will use this process throughout the thesis and draw on the properties discussed in this section.

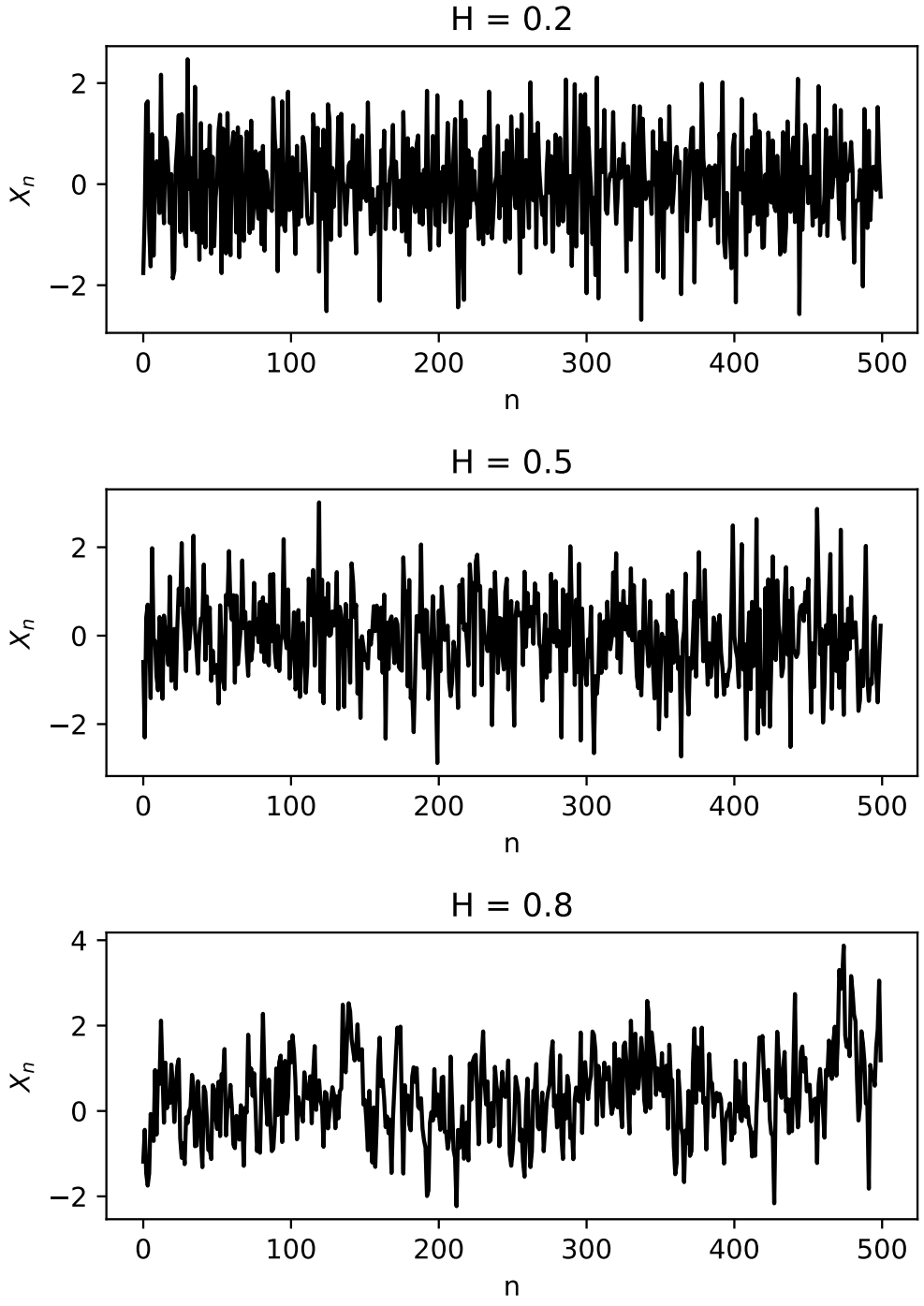

Figure 2.3: Sample paths of Fractional Gaussian Noise with Hurst parameters, $\mathcal{H} = 0.2, 0.5, 0.8$. These realisations were based on the paths of FBM from Figure 2.2, and generated using the *fbm* Python package [66], using the Davies-Harte method [49].. We see the the same properties exist for FGN, longer trends, and larger deviation from the origin.

2.2.2 ARFIMA(p,d,q)

In this section we will introduce the ARFIMA(p,d,q) class of models, which are generalisations of the Autoregressive Moving Average (ARMA) class of time series models. ARMA models, popularised by the influential book by Box and Jenkins [21], and their generalisation Autoregressive Integrated Moving Average (ARIMA) models often used to model non-stationary processes. ARFIMA models are a generalisation of ARIMA, using fractional exponents of the differencing operator, resulting in stationary class of positively and negatively correlated models. They are widely used since the linear structure is easy to analyse, can be applied to a broad range of data.

We begin by defining two polynomials that are required in the definition of the time series models discussed above, they are the autoregressive, $\phi(x)$, and moving average, $\psi(x)$, polynomials respectively. They are used as polynomials of the lag operator, L, where $LX_n = X_{n-1}$, with the polynomials defined as

$$\phi(x) = 1 - \sum_{j=1}^{p} \phi_j x^j, \text{for coefficients } \phi_j \text{ and } p \in \mathbb{Z}^+,$$

$$\psi(x) = 1 + \sum_{j=1}^{q} \psi_j x^j, \text{for coefficients } \psi_j \text{ and } q \in \mathbb{Z}^+.$$

Definition 2.2.10. *A stationary stochastic process $\{X_i\}_{i\in\mathbb{Z}}$, such that*

$$\phi(L)(1-L)^d X_i = \psi(L)\epsilon_i,$$

for some $-1/2 < d < 1/2$ and ϵ_i for $i \in \mathbb{Z}$, is an i.i.d. process with $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$, is called a ARFIMA$(p, d, q)$ process.

Remark. *Note that this process is a Gaussian process, since it's stationary and the noise is given by a normal random variable. The definition can be expanded to allow non-normally distributed noise, however we only use normally distributed noise in this thesis.*

The fractional integration in this case comes from the $(1-L)^d$ term which we expand using the generalised binomial theorem

$$(1-L)^d = \sum_{k=0}^{\infty} \binom{d}{k} (-1)^k L^k,$$

where the binomial coefficients are generalised by the Gamma function,

$$\binom{d}{k} = \frac{\Gamma(d+1)}{\Gamma(k+1)\Gamma(d-k+1)}.$$

These are an extremely flexible class of processes. The autoregressive and moving average polynomials allow any short term behaviour to be modelled while still retaining the asymptotic power-law decay of the covariance functions, and therefore LRD. An important special case of an ARFIMA(p,d,q) process is that of ARFIMA(0,d,0), where $\phi(x) = \psi(x) = 1$, with no lag on the noise, ϵ, and the all the auto-regressive terms on the previous values come from the differencing operator, $(1 - L)^d$. This is much easier to analyse since it doesn't involve double summation to obtain the coefficients, and many theoretical results have been derived in this case.

The exponent, d, has a relationship to the Hurst parameter,

$$\mathcal{H} = d + \frac{1}{2}.$$

In Figure 2.4 we have plotted ARFIMA(0,d,0) processes with exponent $d = -0.3, 0, 0.3$, which are equivalent to $\mathcal{H} = 0.2, 0.5, 0.8$. The process has very similar behaviour to FGN, where we see the emergence of long term trends in the noise for the positively correlated process, and strong oscillation of the negatively correlated process. An advantage of these models is their ability to generate this behaviour with a linear structure.

The zeros of the autoregressive and moving average polynomials impact the stationarity and the existence of some representations, via inversion of the moving average polynomial. If the complex roots of the autoregressive polynomial, $\phi(x)$, lie outside the unit circle then there is one unique stationary solution of the model [21, pg. 55]. This is called causal in the time series literature, and the implication is that the moving average representation, has coefficients with a convergent sum, *i.e.*, $X_n = \sum_{j=0}^{\infty} \theta_j \epsilon_{n-j}$ with $\sum_{j=0}^{\infty} |\theta_j| < \infty$ [23, pg. 85]. If the moving average polynomial has roots outside the unit circle, then we call the process invertible and there is a unique autoregressive representation. The implication being that the autoregressive representative sum has convergent coefficients, $\epsilon_n = \sum_{j=-\infty}^{\infty} \pi_j X_{n-j}$ with $\sum_{j=0}^{\infty} |\pi_j| < \infty$. In this thesis, we will be analysing ARFIMA processes that are both stationary and invertible, with all roots of both polynomials lying outside of the unit circle.

We derive the spectral density by considering the ARFIMA process as an ARMA process on a variable that has been through the linear filter, $(1-L)^d$. That is, we can represent the ARFIMA process, if stationary, as

$$X_n = \frac{\psi(L)}{\phi(L)} (1 - L)^{-d} \epsilon_n.$$

Which is an ARMA process

$$X_n = \frac{\psi(L)}{\phi(L)} Y_n,$$

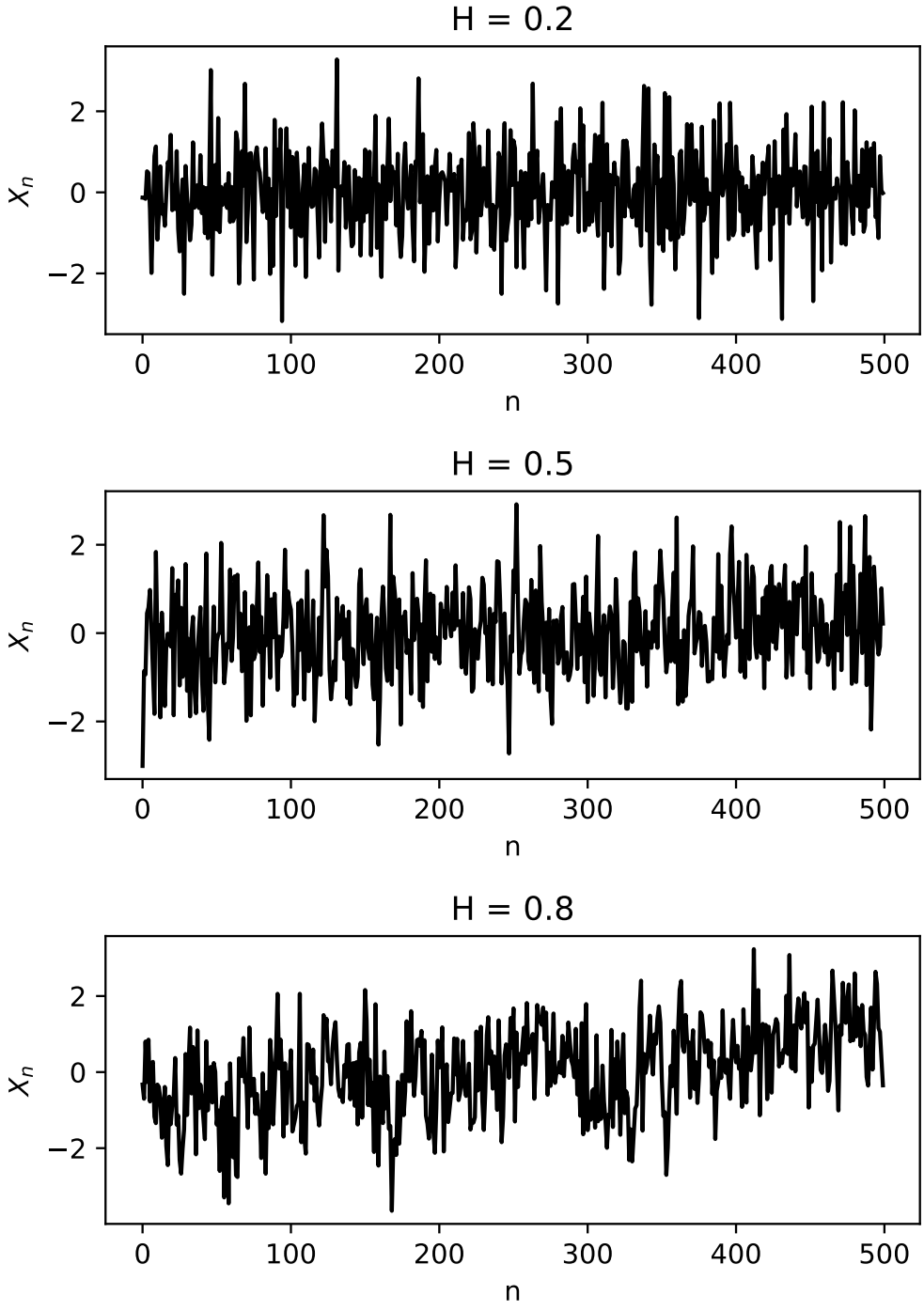

Figure 2.4: Sample paths of ARFIMA(0,d,0) with $d = -0.3, 0, 0.3$ with corresponding Hurst parameters, $\mathcal{H} = 0.2, 0.5, 0.8$. We see the the same properties exist as observed in FGN, although induced by a linear model

for $Y_n = (1-L)^{-d}\epsilon_n$. By Theorem 4.10.1 in Brockwell and Davis [23], the spectral density of a process, $Y = \sum_{k=-\infty}^{\infty} h_k X_k$, that has been through a linear filter, $h(x)$, has a resulting spectral density of $f_Y(\lambda) = |h(e^{i\lambda})|^2 f_X(\lambda)$. Therefore, the spectral density of an ARMA process is

$$f_{ARMA}(\lambda) = \frac{\sigma_\epsilon^2}{2\pi}\frac{|\psi(e^{i\lambda})|^2}{|\phi(e^{i\lambda})|^2},$$

where σ_ϵ^2 is the variance of the innovations, $\epsilon_n \sim \mathcal{N}(0, \sigma_\epsilon)$. Thus the spectral density of an ARFIMA(p,d,q) process is given by,

$$f(\lambda) = \frac{\sigma_\epsilon^2|\psi(e^{i\lambda})|^2}{2\pi|\phi(e^{i\lambda})|^2}|1-e^{i\lambda}|^{-2d}. \tag{2.1}$$

Since $|1-e^{i\lambda}| = \sqrt{2-2\cos\lambda} = 2\sin\left(\frac{|\lambda|}{2}\right)$, and noting the limit as $\lambda \to 0$ is

$$\lim_{\lambda\to 0} 2\sin\left(\frac{1}{2}|\lambda|\right) \to |\lambda|,$$

the spectral density of an ARFIMA process as $\lambda \to 0$ is

$$f(\lambda) \sim \frac{\sigma_\epsilon^2|\psi(1)|^2}{2\pi|\phi(1)|^2}|\lambda|^{-2d}.$$

By Definition 2.2.5, we can see that when $d > 0$ an ARFIMA process is LRD. Directly computing the covariance functions of ARFIMA(p,d,q) processes can be quite difficult in general, however we can show by Theorem 2.2.2 that as $k \to \infty$,

$$\rho(k) \sim c_\rho k^{2d-1}.$$

We present the special case of ARFIMA(0,d,0), from Beran [14, pg. 64], where the correlation function is given by

$$\rho(k) = \frac{\Gamma(1-d)\Gamma(k+d))}{\Gamma(d)\Gamma(k+1-d)}.$$

Due to an asymptotic result on the ratio of Gamma functions [167] showing that as $k \to \infty$,

$$\frac{\Gamma(k+a)}{\Gamma(k+b)} \sim k^{a-b},$$

this implies that

$$\rho(k) \sim \frac{\Gamma(1-d)}{\Gamma(d)} k^{2d-1}.$$

An example theorem characterising ARFIMA(0,d,0) processes, is given below. Dropping the additional complexity from the autoregressive and moving average polynomials, we can get exact autoregressive and moving average representations, with exact expressions for the coefficients.

Theorem 2.2.6 (Proposition 2.2 [14]). *Let X_n be a ARFIMA(0,d,0) process with $-\frac{1}{2} < d < \frac{1}{2}$. Then*
(i) the following infinite autoregressive representation holds:

$$\sum_{k=0}^{\infty} \pi_k X_{n-k} = \epsilon_n,$$

where $\epsilon_n (n = 1, 2, ...)$ are independent identically distributed random variables and

$$\pi_k = \frac{\Gamma(k-d)}{\Gamma(k+1)\Gamma(-d)}.$$

For $k \to \infty$ we have,

$$\pi_k \sim \frac{1}{\Gamma(-d)} k^{-d-1}.$$

(ii) The following infinite moving average representation holds:

$$X_n = \sum_{k=0}^{\infty} a_k \epsilon_{n-k}$$

where $\epsilon_n (n = 1, 2, ...)$ are independent identically distributed random variables and

$$a_k = \frac{\Gamma(k+d)}{\Gamma(k+1)\Gamma(d)}.$$

For $k \to \infty$ we have

$$a_k \sim \frac{1}{\Gamma(d)} k^{d-1}.$$

Note that the autoregressive and moving average polynomials are $\phi(x) = \psi(x) = 1, \forall x$, and therefore ARFIMA(0,d,0) is stationary and invertible.

Chapter 3

Differential Entropy Rate Characterisation of Long Range Dependent Gaussian Processes

We discussed in the last chapter that the entropy rate of discrete time stochastic processes has been studied as a measure of the average uncertainty. Most studies have focused on processes whose correlations decay quickly, and hence the dependence on past observations disappears rapidly. However, many real processes from a variety of contexts, *i.e.,* data networks [116, 178, 179], climate [171], hydrology [14, 89, 114], and economics [39, 180], have been shown to exhibit long range dependence.

Recent work has investigated information theoretic characterisation of long range and short range processes [30, 54, 120], using the finiteness of mutual information between past and future. In this chapter, we aim to clarify this characterisation and investigate its implications for Gaussian processes.

We calculate the differential entropy rate for the two most common stationary Gaussian Long Range Dependent (LRD) processes: Fractional Gaussian Noise (FGN) and the Auto-Regressive Fractionally-Integrated Moving Average (ARFIMA). We start by deriving the entropy rate for these processes, and show that they both have negative poles as the processes tend towards strong long-range correlations, but that their behaviour when anti-correlated is surprisingly different: FGN has a pole similar to that for positive correlations, but ARFIMA does not. This contradicts common intuition based on their similar spectral densities that FGN and ARFIMA(0,d,0) are close to equivalent.

We also investigate the links between the two information measures: excess entropy and the mutual information between past and future processes, and compare these to the differential entropy rate. We show that the dif-

ferential entropy rate definition for excess entropy is equivalent to the mutual information between past and future for continuous valued discrete time Gaussian processes, and hence that excess entropy is infinite for all LRD and CSRD Gaussian processes with power-law decaying autocovariance functions.

3.1 Entropy rate function for Fractional Gaussian Noise

We want to understand the effect of memory on the entropic properties of a stochastic process. We start with the entropy rate characterisation for Gaussian processes originally derived by Kolmogorov (see Ihara [92, pg. 76])

$$h(\chi) = \frac{1}{2}\log(2\pi e) + \frac{1}{4\pi}\int_{-\pi}^{\pi} \log(2\pi f(\lambda))\, d\lambda, \tag{3.1}$$

where $f(\lambda)$ is the spectral density, *i.e.,* the Fourier transform of the autocovariance function for a mean zero process. We use this characterisation due to the specific dependence on the spectral density, for Gaussian processes. This assists us in deriving understanding of LRD and SRD Gaussian processes, and their impact on entropy rates, due to the properties of spectral densities, i.e. pole at $\lambda = 0$ for LRD processes, continuous and positive for SRD processes, and zero at $\lambda = 0$ for CSRD processes.

We will begin by investigating the spectral density of Fractional Gaussian Noise (FGN), which is given by Theorem 2.2.5

$$f(\lambda) = 2c_f(1 - \cos\lambda) \sum_{j=-\infty}^{\infty} |2\pi j + \lambda|^{-2\mathcal{H}-1}, \tag{3.2}$$

where, $c_f = \frac{\sigma^2}{2\pi}\sin(\pi\mathcal{H})\Gamma(2\mathcal{H}+1)$, $\mathcal{H}$ is the Hurst parameter, and σ^2 is the variance of the process.

This spectral density is difficult to analyse as it has an infinite sum of absolute values. In particular, when we apply the entropy rate characterisation 3.1, as it involves taking a logarithm of a sum, making analytical calculation prohibitively difficult. However, we can still use this expression to derive some properties of the entropy rate of FGN processes.

3.1.1 Comparison of approximate and analytical spectral density for entropy rate calculation

Substituting the spectral density of FGN (3.2) into the second term in the entropy rate expression (3.1) we get

$$\begin{aligned}\int_{-\pi}^{\pi} \log\left(2\pi f(\lambda)\right) d\lambda &= 2\pi \log(4\pi c_f) + \int_{-\pi}^{\pi} \log(1-\cos\lambda) d\lambda \\ &\quad + \int_{-\pi}^{\pi} \log\left(\sum_{j=-\infty}^{\infty} |2\pi j + \lambda|^{-2\mathcal{H}-1}\right) d\lambda, \\ &= 2\pi \log(4\pi c_f) - 2\pi \log 2 \\ &\quad + \int_{-\pi}^{\pi} \log\left(\sum_{j=-\infty}^{\infty} |2\pi j + \lambda|^{-2\mathcal{H}-1}\right) d\lambda. \end{aligned} \tag{3.3}$$

Where the second term, $\int_{-\pi}^{\pi} \log(1-\cos\lambda)d\lambda = -2\pi\log 2$, is given in Jeffrey and Dai [97, pg. 274]. The last term is finite for all $\mathcal{H} \in (0,1)$, since the singularity that exists when $\lambda = j = 0$ in the absolute value term is integrable. This is important as we can then see that this does not affect the asymptotic behaviour of FGN processes. The resulting entropy rate is

$$\begin{aligned} h(\chi) = \frac{1}{2}\log(2\pi e\sigma^2) + \frac{1}{2}\log\left(\sin(\pi\mathcal{H})\Gamma(2\mathcal{H}+1)\right) \\ + \frac{1}{4\pi}\int_{-\pi}^{\pi} \log\left(\sum_{j=-\infty}^{\infty} |2\pi j + \lambda|^{-2\mathcal{H}-1}\right) d\lambda, \end{aligned} \tag{3.4}$$

from plugging the result of 3.3 into 3.1.

We calculate $h(\chi)$ using numerical integration via Python's SpiPy library [177]. We plot the differential entropy rate of Fractional Gaussian Noise as a function of the Hurst parameter, $\mathcal{H}$, in Figure 3.1. The plot shows the impact of the variance on entropy rate calculation, and hence that the entropy rate of Fractional Gaussian Noise has a large dependence on the variance, which isn't surprising given the characterisation of entropy rate for Gaussian processes via its spectral density. Each unit increase in variance has a smaller effect on the value of the differential entropy, due to the $\log(\sigma^2)$ term. In general, the entropy rate of a Gaussian process is proportional to the logarithm of the innovation variance [51]. So we expect that the entropy rate scales at this rate in Gaussian processes, and therefore that the impact of variance reduces as the size increases.

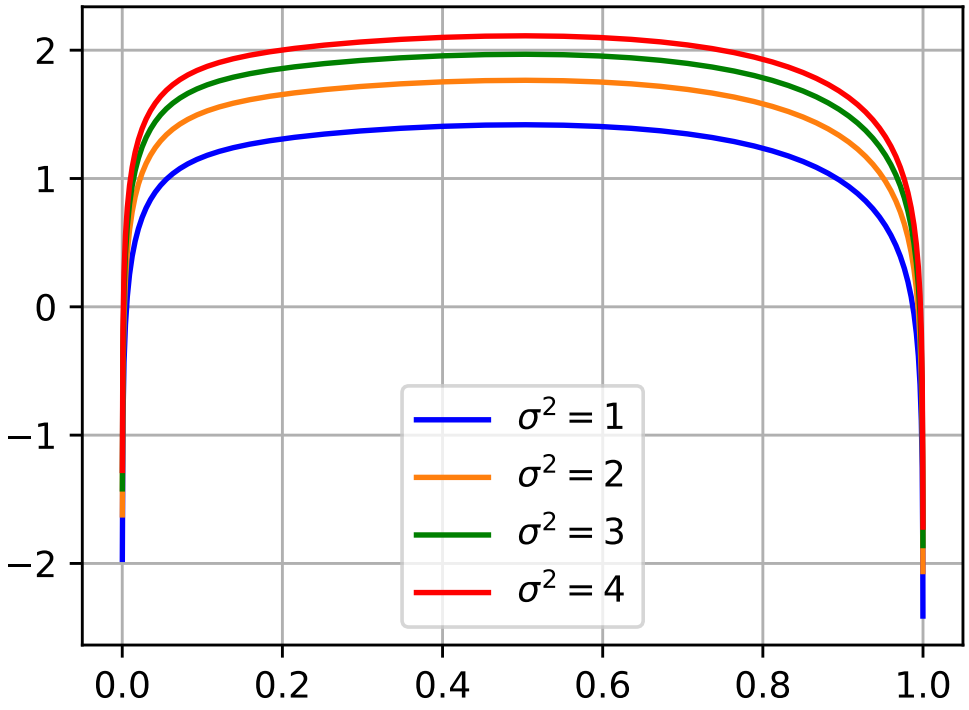

Figure 3.1: Entropy rate of Fractional Gaussian Noise as a function of the Hurst Parameter. The maximum is at $\mathcal{H} = 0.5$, where the process is white Gaussian noise. As $\mathcal{H} \to 0$ or 1, the function tends towards $-\infty$, as the strength of the negative or positive correlations increase. The impact of changing variance decreases as the variance increase, due to the $\log(\sigma^2)$ term.

Another approach is to consider the dependence between positive and negative correlations and the leading constant c_f for the limit of spectral density as $\lambda \to 0$, since these adjust for the degree of correlation. This approach is inspired by Veitch et al. [174]. We derive the entropy rate function similar to Equation 3.4, without expanding c_f. We substitute Equation 3.3 into the entropy rate expression which gives

$$
\begin{aligned}
h(\chi) &= \frac{1}{2}\log(2\pi e) \\
&\quad + \frac{1}{4\pi}\left(2\pi \log(4\pi c_f) - 2\pi \log 2 + \int_{-\pi}^{\pi} \log\left(\sum_{j=-\infty}^{\infty} |2\pi j + \lambda|^{-2\mathcal{H}-1}\right) d\lambda\right), \\
&= \frac{1}{2}\log(2\pi e) + \frac{1}{2}\log(4\pi c_f) - \frac{1}{2}\log 2
\end{aligned}
$$

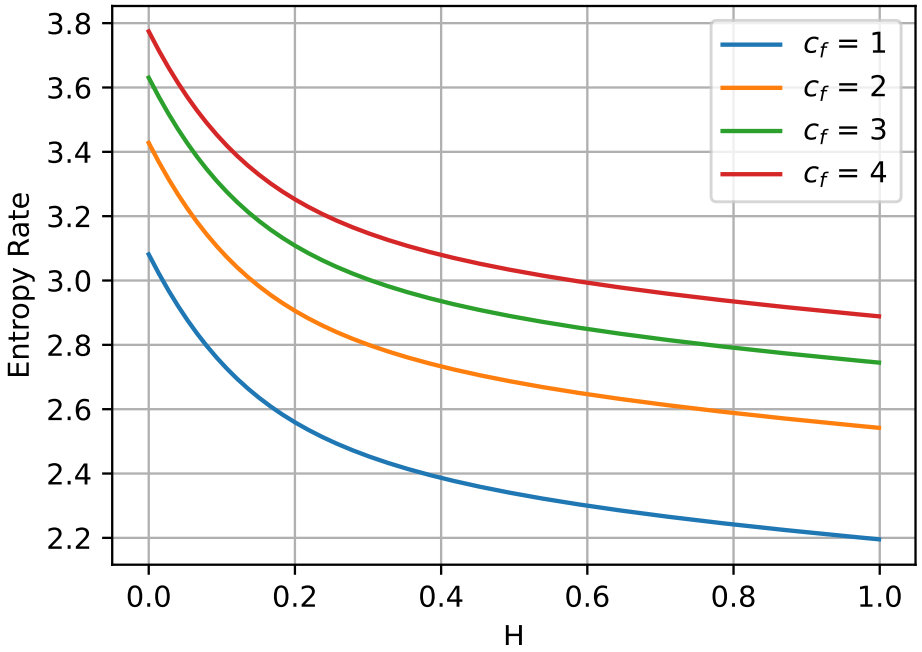

Figure 3.2: Entropy rate of Fractional Gaussian Noise as a function of the Hurst Parameter, however considering the impact of c_f. The impact of changing c_f decreases as the variance increase, due to the $\log(c_f)$ term.

$$+ \frac{1}{4\pi} \int_{-\pi}^{\pi} \log \left(\sum_{j=-\infty}^{\infty} |2\pi j + \lambda|^{-2\mathcal{H}-1} \right) d\lambda,$$

$$= \frac{1}{2} \log(4\pi^2 e c_f) + \frac{1}{4\pi} \int_{-\pi}^{\pi} \log \left(\sum_{j=-\infty}^{\infty} |2\pi j + \lambda|^{-2\mathcal{H}-1} \right) d\lambda. \tag{3.5}$$

We plot the differential entropy rate as a function of $\mathcal{H}$, for constant $c_f = 1, 2, 3, 4$, in Figure 3.2.

The spectral density expression is quite cumbersome to work with and an approximation is often used, which is accurate at low frequencies [14, pg. 53]. It is derived from a Taylor series expansion of the spectral density and is given by,

$$f(\lambda) \approx c_f |\lambda|^{1-2\mathcal{H}}.$$

We can obtain a closed form for the entropy rate if we substitute this ap-

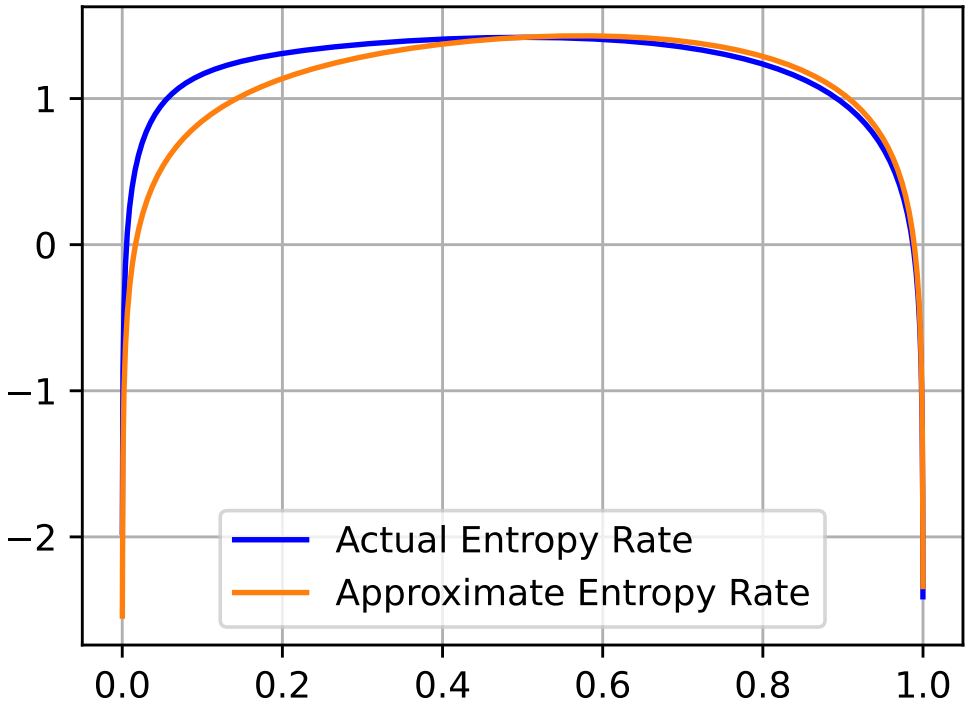

Figure 3.3: Comparison of the numerically integrated spectral density and the spectral density approximation. The approximation is relatively good for $\mathcal{H} \geq 1/2$ but an underestimate for $\mathcal{H} \leq 1/2$.

proximation into the integral in the entropy rate expression (3.1) to get

$$\begin{aligned}\int_{-\pi}^{\pi} \log(2\pi f(\lambda))\, d\lambda &= \int_{-\pi}^{\pi} \log\left(2\pi c_f\right) d\lambda + \int_{-\pi}^{\pi} \log\left(|\lambda|^{1-2\mathcal{H}}\right)\, d\lambda, \\ &= 2\pi \log\left(2\pi c_f\right) + 2\left(1-2\mathcal{H}\right) \int_{0}^{\pi} \log\left(\lambda\right)\, d\lambda, \\ &= 2\pi \log\left(2\pi c_f\right) + 2\left(1-2\mathcal{H}\right)\left(\pi \log \pi - \pi\right).\end{aligned}$$

Note that there is a singularity at the origin of the spectral density of LRD processes. However, the integral is still well defined and finite in this case. Therefore the entropy rate approximation is,

$$\tilde{h}(\chi) = \frac{1}{2}\log(2\pi e\sigma^2) + \frac{1}{2}\log\left(\sin(\pi\mathcal{H})\Gamma(2\mathcal{H}+1)\right) + \frac{1}{2}(1-2\mathcal{H})(\log\pi - 1),$$

which differs from the exact formulation only in the last term.

Figure 3.3 shows the entropy rate and its approximation. We can see that the entropy rate approximation is very good for the positively correlated cases $\mathcal{H} \geq 0.5$ and at the limits around $\mathcal{H} = 0$ or 1. However for moderately, negatively-correlated processes the approximation is a noticeable underestimate of the entropy rate.

3.1.2 Properties of Entropy rate for Fractional Gaussian Noise

Figure 3.3 shows some interesting properties

- The entropy rate function as a function of $\mathcal{H}$ is not symmetric. Negatively correlated processes seem to have higher uncertainty the same distance from $\mathcal{H} = 0.5$.
- The entropy rate asymptotically tends to $-\infty$ as $\mathcal{H} \to 0$ or 1.
- The maximum entropy rate occurs at 0.5. Indicating that the maximum entropy occurs for white Gaussian noise.

We explain how these properties emerge below.

3.1.2.1 Asymptotic behaviour

Theorem 3.1.1. *The differential entropy rate of Fractional Gaussian Noise, $h(\chi) \to -\infty$ as $\mathcal{H} \to 0$ or 1.*

Proof. When $\mathcal{H} \to 0$ or 1, the term $c_f \to 0$, as the gamma function terms are non-zero, however the trigonometric terms tend to 0 as $\mathcal{H}$ tends to an integer value. Hence, asymptotically the entropy rate expression is dominated by $\log c_f \to -\infty$, as $c_f \to 0$, since for all $\mathcal{H}$ the integral term is finite. □

Remark. *Note that the approximation works well in the limits $\mathcal{H} \to 0$ or 1, and so the theorem describes the asymptotic behaviour of entropy rate well. Moreover, the theorem lines up with the intuition for an LRD process. As we move closer to either perfectly positively or negatively correlated, the process becomes "less uncertain",* i.e., *we have less entropy on average. When the uncertainty disappears, by viewing the entire past we can accurately infer the current value. It's important to reiterate that the differential entropy can be $-\infty$, which can be interpreted as least uncertainty for a process.*

3.1.2.2 Maximum

We want to understand the maximum of differential entropy rate, as a function of the Hurst parameter. This will provide an understanding of which parameter choices represent the highest uncertainty. We differentiate the entropy rate, with respect to $\mathcal{H}$ and then solve for $\mathcal{H}$ when the derivative equals zero. Here we need to apply this to the exact formula because the approximation distorts the location of the maximum. Therefore, dropping constant terms, we get

$$\begin{aligned}\frac{dh}{d\mathcal{H}} &= \frac{1}{2}\frac{d}{d\mathcal{H}}\log\left(\sigma^2\sin(\pi\mathcal{H})\Gamma(2\mathcal{H}+1)\right) \\ &\qquad + \frac{1}{4\pi}\frac{d}{d\mathcal{H}}\int_{-\pi}^{\pi}\log\left(\sum_{j=-\infty}^{\infty}|2\pi j+\lambda|^{-2\mathcal{H}-1}\right)d\lambda \\ &= \frac{1}{2}\frac{d}{d\mathcal{H}}\log\left(\sin(\pi\mathcal{H})\right) + \frac{1}{2}\frac{d}{d\mathcal{H}}\log\left(\Gamma(2\mathcal{H}+1)\right) \\ &\qquad - \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\sum_j \log(|2\pi j+\lambda|)|2\pi j+\lambda|^{-2\mathcal{H}-1}}{\sum_j |2\pi j+\lambda|^{-2\mathcal{H}-1}}d\lambda \\ &= \frac{\pi}{2}\cot(\pi\mathcal{H}) + \psi(2\mathcal{H}+1) - \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\sum_j \log(|2\pi j+\lambda|)|2\pi j+\lambda|^{-2\mathcal{H}-1}}{\sum_j |2\pi j+\lambda|^{-2\mathcal{H}-1}}d\lambda.\end{aligned}$$

where $\psi(z) = \Gamma'(z)/\Gamma(z)$ is the digamma function.

Then we set this expression to zero, and solve for $\mathcal{H}$. This is a transcendental equation with no closed form. We solve it numerically using Python's SciPy package [177], which yields $\mathcal{H} \approx 0.500$. Therefore we conjecture that the maximum entropy rate, using the exact spectral density, is at $\mathcal{H} = 0.5$, which aligns with the idea that a SRD FGN process has more uncertainty than any equivalent LRD FGN process.

Note that from the solution of the spectral density approximation is $\mathcal{H} \approx 0.516$. So although using the spectral density approximation is acceptable for many purposes, it can lead to false conclusions about the properties of the differential entropy rate.

3.2 Entropy rate function for ARFIMA(p,d,q)

We consider the differential entropy rate function of a related process to Fractional Gaussian Noise, which is ARFIMA(p,d,q), the fractional extension of the ARIMA (Autoregressive Integrated Moving Average) processes, by extending to non-integer differencing parameters, d [77, 87]. FGN and ARFIMA(0,d,0) are commonly used stationary LRD processes for modelling

real phenomena, and in particular FGN and ARFIMA(0,d,0) have very similar properties in the time and frequency domains, as seen in Sections 2.2.1 and 2.2.2 respectively. Additionally, these processes have been linked by limit of their autocorrelation coefficients, $\rho(k) := \gamma(k)/\sigma^2$, under aggregation and rescaling [72]. However, ARFIMA processes do differ from FGN in that you could change the rate of convergence to a fixed point, *i.e.*, alter the eventual limit under aggregation and rescaling, with the addition of additive noise [174], which implies that this class is less robust to the addition of noise. Hence, there may be some differences in behaviour when looking through an entropic lens.

We will express an entropy rate characterisation for ARMA processes in terms of its innovation process variance, from Ihara [92, pg. 78], and show that this can be extended to ARFIMA(0,d,0) and ARFIMA(p,d,q) processes. Then we will use the result to characterise the entropy rate of an ARFIMA(0,d,0) process in terms of its process variance.

Theorem 3.2.1 (From Ihara [92, pg. 78])**.** *The entropy rate of an ARMA(p,q) process is given by,* $h(\chi) = \frac{1}{2}\log(2\pi e\sigma_\epsilon^2)$.

This is an interesting result, since the entropy rate of an ARMA process is dependent upon the variance of the innovations, independent of the parameters of an ARMA process. We will investigate if the autoregressive and moving average parameters have any impact on the entropy rate of ARFIMA processes. Now, we state our extension of this result to ARFIMA(0,d,0) and present a proof based on Ihara's proof of Theorem 3.2.1

Theorem 3.2.2. *The entropy rate of a stationary ARFIMA(0,d,0) process is given by,* $h(\chi) = \frac{1}{2}\log(2\pi e\sigma_\epsilon^2)$.

Proof. First we calculate $\int_{-\pi}^{\pi}\log(2\pi f(\lambda))d\lambda$, using the spectral density of an ARFIMA process given in (2.1)

$$\begin{aligned}\int_{-\pi}^{\pi}\log(2\pi f(\lambda))d\lambda &= \int_{-\pi}^{\pi}\log(\sigma_\epsilon^2|1-e^{i\lambda}|^{1-2\mathcal{H}})d\lambda,\\ &= \int_{-\pi}^{\pi}\log(\sigma_\epsilon^2)d\lambda + (1-2\mathcal{H})\int_{-\pi}^{\pi}\log|1-e^{i\lambda}|d\lambda.\end{aligned}$$

Now we transform the elements in the last term using their trigonometric representation,

$$\begin{aligned}|1-e^{i\lambda}| &= |1-\cos(\lambda)-i\sin(\lambda)|,\\ &= \sqrt{(1-\cos(\lambda))^2+\sin^2(\lambda)},\end{aligned}$$

$$
\begin{aligned}
&= \sqrt{2 - 2\cos(\lambda)}, \\
&= \sqrt{4\sin^2\left(\frac{\lambda}{2}\right)}, \\
&= 2\left|\sin\left(\frac{\lambda}{2}\right)\right|.
\end{aligned}
$$

This makes the integral of the log spectral density,

$$
\begin{aligned}
\int_{-\pi}^{\pi} \log|1 - e^{i\lambda}| d\lambda &= 2\int_0^{\pi} \log\left(2\sin\left(\frac{\lambda}{2}\right)\right) d\lambda, \\
&= 2\int_0^{\pi} \log(2) d\lambda + 2\int_0^{\pi} \log\left(\sin\left(\frac{\lambda}{2}\right)\right) d\lambda,
\end{aligned}
$$

We substitute $y = \lambda/2$,

$$
\begin{aligned}
\int_{-\pi}^{\pi} \log|1 - e^{i\lambda}| d\lambda &= 2\pi\log(2) + 2\int_0^{\frac{\pi}{2}} \log(\sin y) 2dy, \\
&= 2\pi\log(2) + 4\left(-\frac{\pi}{2}\log(2)\right), \\
&= 0.
\end{aligned}
$$

Where the equality $\int_0^{\frac{\pi}{2}} \log(\sin y) dy = -\frac{\pi}{2}\log(2)$ is given by [108].
So the last term of the spectral density vanishes, and

$$
\begin{aligned}
\int_{-\pi}^{\pi} \log(2\pi f(\lambda)) d\lambda &= \int_{-\pi}^{\pi} \log(\sigma_\epsilon^2) d\lambda + (1 - 2\mathcal{H}) \int_{-\pi}^{\pi} \log|1 - e^{i\lambda}| d\lambda, \\
&= 2\pi\log(\sigma_\epsilon^2).
\end{aligned}
$$

Using Kolmogorov's entropy rate expression, the entropy rate is therefore,

$$
\begin{aligned}
h(\chi) &= \frac{1}{2}\log(2\pi e) + \frac{1}{4\pi}\left(2\pi\log(\sigma_\epsilon^2)\right), \\
&= \frac{1}{2}\log(2\pi e\sigma_\epsilon^2).
\end{aligned}
$$

□

Remark. *This can be shown also using the infinite autoregressive expression in Theorem 2.2.6, $X_n = \epsilon_n - \sum_{k=1}^{\infty} \pi_k X_{n-k}$, and substituting into the conditional entropy rate for stationary processes,*

$$
h(\chi) = \lim_{n\to\infty} h(X_n | X_{n-1}, ..., X_0).
$$

Then we can remove the conditioning from the entropy rate calculation,

$$h(\chi) = \lim_{n\to\infty} h\left(\epsilon_n - \sum_{k=1}^{\infty} \pi_k X_{n-k} \middle| X_{n-1}, ..., X_0\right) = \lim_{n\to\infty} h(\epsilon_n).$$

Which then implies that $h(\chi) = \frac{1}{2}\log(2\pi e\sigma_\epsilon^2)$, i.e., *the entropy rate of the process depends only on the entropy introduced at each step by the innovations. Therefore, we conclude that the entropy rate of an ARFIMA(0,d,0) process is constant with respect to the innovation variance, but potentially has dependence on* $\mathcal{H}$ *when considering its process variance.*

We can generalise to ARFIMA(p,d,q) processes by adding an additional condition, the invertibility of the moving average polynomial.

Theorem 3.2.3. *The entropy rate of a stationary ARFIMA(p,d,q) process with invertible moving average polynomial is given by,* $h(\chi) = \frac{1}{2}\log(2\pi e\sigma_\epsilon^2)$.

Proof. Since ARFIMA(p,d,q) processes are stationary and invertible, this implies that the polynomials $\phi(x)$ and $\psi(x)$, have roots outside of the unit circle, i.e. each root $z \in \mathbb{C}$ is such that $|z| > 1$. By the Fundamental Theorem of Algebra, both the autoregressive and moving average polynomials can be factored into affine factors. As the constant terms are 1, this implies the polynomials can be factored as $\phi(x) = \prod_{i=1}^{p}(1 - a_i x)$ and $\psi(x) = \prod_{i=i}^{q}(1 - b_i x)$, where $|a_i|, |b_i| < 1, \forall i$. Recall from equation (2.1)that the spectral density is given by,

$$f(\lambda) = \frac{\sigma_\epsilon^2}{2\pi}|1 - e^{i\lambda}|^{-2d}\frac{|\psi(e^{i\lambda})|^2}{|\phi(e^{i\lambda})|^2}.$$

Hence,

$$\begin{aligned}
&\log((2\pi f(\lambda))) \\
&\quad = \log(\sigma_\epsilon^2) - 2d\log|1 - e^{i\lambda}| + 2\log\left|\prod_{j=1}^{q}(1 - a_j e^{i\lambda})\right| - 2\log\left|\prod_{j=1}^{p}(1 - b_j e^{i\lambda})\right|, \\
&\quad = \log(\sigma_\epsilon^2) - 2d\log|1 - e^{i\lambda}| + \sum_{j=1}^{q} 2\log|1 - a_j e^{i\lambda}| - \sum_{j=1}^{p} 2\log|1 - b_j e^{i\lambda}|.
\end{aligned}$$

Now we calculate the integral of the log spectral density,

$$\int_{-\pi}^{\pi} \log(2\pi f(\lambda))d\lambda = \int_{-\pi}^{\pi} \log(\sigma_\epsilon^2)d\lambda - \int_{-\pi}^{\pi} 2d\log|1 - e^{i\lambda}|d\lambda$$

$$+\sum_{j=1}^{q} 2\int_{-\pi}^{\pi} \log|1-a_j e^{i\lambda}|d\lambda - \sum_{j=1}^{p} 2\int_{-\pi}^{\pi} \log|1-b_j e^{i\lambda}|d\lambda,$$
$$= 2\pi \log(\sigma_\epsilon^2).$$

Where the third equality is given as all the integrals of $\log|1-ae^{i\lambda}|$ over $[-\pi,\pi]$ vanish for $|a|\leq 1$ [158].

We substitute this expression into Kolmogorov's entropy rate expression for Gaussian processes.

$$h(\chi) = \frac{1}{2}\log(2\pi e) + \frac{1}{4\pi}(2\pi\log(\sigma_\epsilon^2)),$$
$$= \frac{1}{2}\log(2\pi e\sigma_\epsilon^2).$$

□

This result leads to the following corollary, which can finalise the discussion of the differential entropy rate in terms of innovation variance for the classes of AR, MA, ARMA processes. This is relevant as the definition in terms of the innovation variance is the perspective that is commonly used in the time series literature, when modelling real world processes.

Corollary 3.2.3.1. *The differential entropy rate of stationary AR(p), invertible MA(q) and, stationary and invertible ARMA(p,q) processes is* $h(\chi) = \frac{1}{2}\log(2\pi e\sigma_\epsilon^2)$.

Hence, for these models the entropy rate can be calculated in terms of the variance of its innovations. However we want to compare the entropy rates, as a function of their Hurst parameter, between ARFIMA(0,d,0) and FGN, so we want to fix the variance of process itself, σ^2. We will use the autocovariance function of ARFIMA(0,d,0), from Beran [14, pg. 63],

$$\gamma(k) = \sigma_\epsilon^2 \frac{(-1)^k\Gamma(1-2d)}{\Gamma(k-d+1)\Gamma(1-k-d)}.$$

Note that $\gamma(0)=\sigma^2$,

$$\sigma^2 = \gamma(0) = \sigma_\epsilon^2 \frac{\Gamma(1-2d)}{\Gamma(1-d)^2},$$

and hence,

$$\sigma_\epsilon^2 = \sigma^2 \frac{\Gamma(1-d)^2}{\Gamma(1-2d)}.$$

This leads to the following characterisation of ARFIMA(0,d,0) processes in terms of the Hurst parameter, $\mathcal{H}$, noting that $d = \mathcal{H} - 1/2$.

Corollary 3.2.3.2. *The entropy rate of an ARFIMA(0,d,0) process for a fixed process variance, σ^2, is given by,*

$$h(\chi) = \frac{1}{2}\log(2\pi e\sigma^2) + \log\left(\Gamma\left(\frac{3}{2} - \mathcal{H}\right)\right) - \frac{1}{2}\log\left(\Gamma(2-2\mathcal{H})\right). \tag{3.6}$$

Proof. By Theorem 3.2.2 and from the characterisation of σ_ϵ^2 above,

$$\begin{aligned} h(\chi) &= \frac{1}{2}\log\left(2\pi e\sigma^2\frac{\Gamma(1-d)^2}{\Gamma(1-2d)}\right), \\ &= \frac{1}{2}\log(2\pi e\sigma^2) + \log(\Gamma(1-d)) - \frac{1}{2}\log(\Gamma(1-2d)), \\ &= \frac{1}{2}\log(2\pi e\sigma^2) + \log\left(\Gamma\left(\frac{3}{2} - \mathcal{H}\right)\right) - \frac{1}{2}\log(\Gamma(2-2\mathcal{H})). \end{aligned}$$

□

Remark. *The same approach can be used for more general ARFIMA(p,d,q) processes. However in this case, there is no general closed form for the autocovariance function, so the variance must be calculated for each process and then substituted for the innovation variance. Interestingly, this result indicates that the effect of the changing the process variance is balanced by the effect of the change in the Hurst parameter, with respect to the innovation variance. This results in the constant differential entropy rate when considered in terms of its innovation variance.*

We show the plot of the ARFIMA(0,d,0) entropy rate as a function of the Hurst parameter, $\mathcal{H}$, with process variance, $\sigma^2 = 1, 2, 3, 4$, in Figure 3.4. The plot shows some interesting behaviour, particularly when compared to the FGN entropy rate function in Figure 3.6. Some of these observed properties are:

- The entropy rate is not symmetric, much less so than FGN. The positively correlated side has a dramatic drop, however the negatively correlated side stays relatively high. In order words, there is a demonstrable difference between FGN and ARFIMA(0,d,0) in the behaviour as CSRD processes.
- The entropy rate asymptotically tends to $-\infty$ as $\mathcal{H} \to 1$ only.
- The maximum entropy rate occurs at the same point as FGN, $\mathcal{H} = 0.5$, indicating that the maximum entropy occurs for white Gaussian noise.

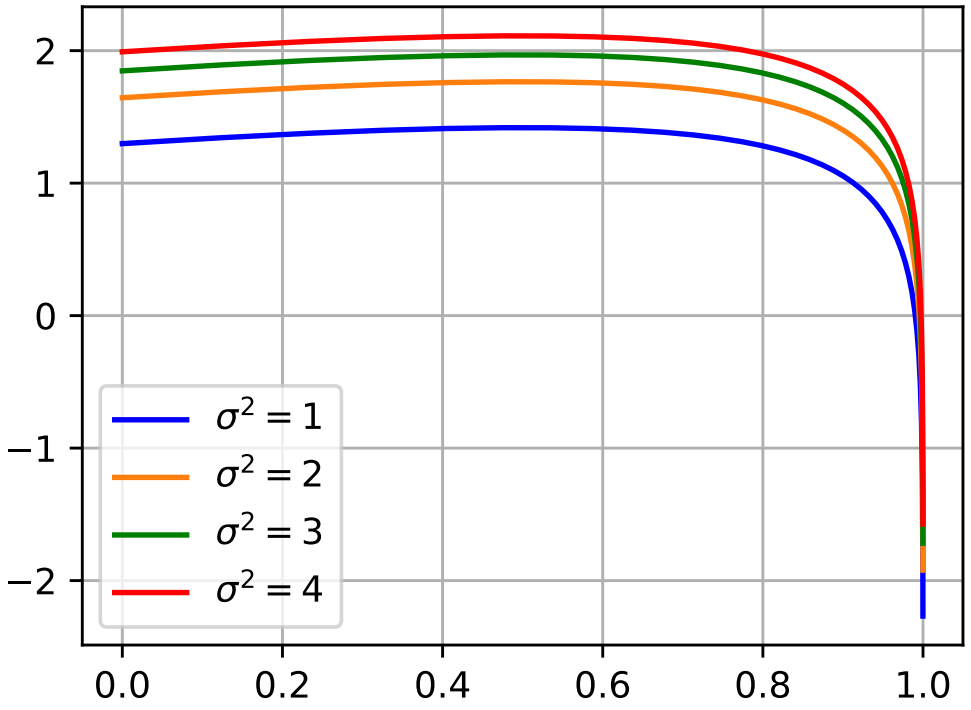

Figure 3.4: The entropy rate of ARFIMA(0,d,0) as a function of the Hurst parameter, $\mathcal{H}$, for variance, $\sigma^2 = 1, 2, 3, 4$. On the positively correlated side, $\mathcal{H} > 0.5$, we see a similar asymptotic behaviour to FGN. However, for negatively correlated processes, the amount of entropy in the process, stays quite high. We see the maximum of the function at $\mathcal{H} = 0.5$, which intuitively shows that the highest uncertainty occurs for the white Gaussian noise process.

Note that we have examined the dependence between the degree of positive or negative correlations and the process variance. This was chosen due to the dependence observed in Gaussian processes between the entropy rate and variance. As in Section 3.1 we also derive the entropy rate function without expanding c_f and plot for constant $c_f = 1, 2, 3, 4$.

The spectral density for an ARFIMA(p,d,q) process is,

$$f(\lambda) = f_{ARMA}(\lambda)|1 - e^{i\lambda}|^{-2d},$$

where

$$f_{ARMA}(\lambda) = \frac{\sigma_\epsilon^2}{2\pi} \frac{|\psi(e^{i\lambda})|^2}{|\phi(e^{i\lambda})|^2}.$$

Noting that in the limit as $\lambda \to 0$, that $f(\lambda) \sim f_{ARMA}(0)|\lambda|^{-2d}$.

Taking c_f for ARFIMA(p,d,q) processes as the coefficient in the asymptotic limit as $\lambda \to 0$, we have

$$c_f = f_{ARMA}(0) = \frac{\sigma_\epsilon^2}{2\pi} \frac{|\psi(1)|^2}{|\phi(1)|^2}.$$

In the case of ARFIMA(0,d,0) we have that the autoregressive and moving average polynomials are constant and equal to one, i.e. $\psi(x) = \phi(x) = 1$, we get that

$$c_f = \frac{\sigma_\epsilon^2}{2\pi}.$$

Substituting this into the entropy rate expression we get

$$\begin{aligned} h(\chi) &= \frac{1}{2}\log(2\pi e) + \frac{1}{4\pi}\int_{-\pi}^{\pi} \log\left(2\pi c_f |1 - e^{i\lambda}|^{-2d}\right) d\lambda, \\ &= \frac{1}{2}\log(2\pi e) + \frac{1}{2}\pi \log(4\pi c_f) - \frac{1}{2}\pi \log 2 + \frac{1}{4\pi}\left(2\pi \log(2\pi c_f)\right), \\ &= \frac{1}{2}\log(4\pi^2 e c_f). \end{aligned}$$

Note that this differs from Equation 3.5 only by the last term. Therefore the differences between FGN and ARFIMA(0,d,0) are solely due to the integral term in FGN's spectral density. Additionally, the entropy rate is constant with respect to $\mathcal{H}$ for c_f. This is due to the result of Theorem 3.2.2, as in the case of ARFIMA(0,d,0), the entropy rate is constant with respect to the innovation variance and c_f, for ARFIMA(0,d,0), scales the innovation variance. The differential entropy rate function is plotted as a function of $\mathcal{H}$, for a range of different c_f's is given in Figure 3.5.

Similar to the previous section, we will prove the asymptotics of the entropy rate function, and show that the maximum occurs at $\mathcal{H} = 0.5$.

Corollary 3.2.3.3. *The differential entropy rate of ARFIMA(0,d,0), $h(\chi)$, tends to negative infinity as $H \to 1$, for a fixed variance $\sigma^2 \in \mathbb{R}$.*

Proof. As $\mathcal{H} \to 1$, the term $\Gamma(\frac{3}{2} - \mathcal{H})$ is bounded away from zero, as well as the $\frac{1}{2}\log(2\pi e\sigma^2)$, for a fixed variance $0 < \sigma^2 < \infty$. Now, as $\mathcal{H} \to 1$, the term $\Gamma(2 - 2\mathcal{H}) \to \Gamma(0)$. There exists a singularity for the gamma function at 0, which diverges to infinity. Which implies that the term $-\frac{1}{2}\log(\Gamma(2 - 2\mathcal{H})) \to -\infty$, since $\Gamma(x) \to \infty$, as $x \to 0$. This implies that $h(\chi) \to -\infty$, as $\mathcal{H} \to 1$. □

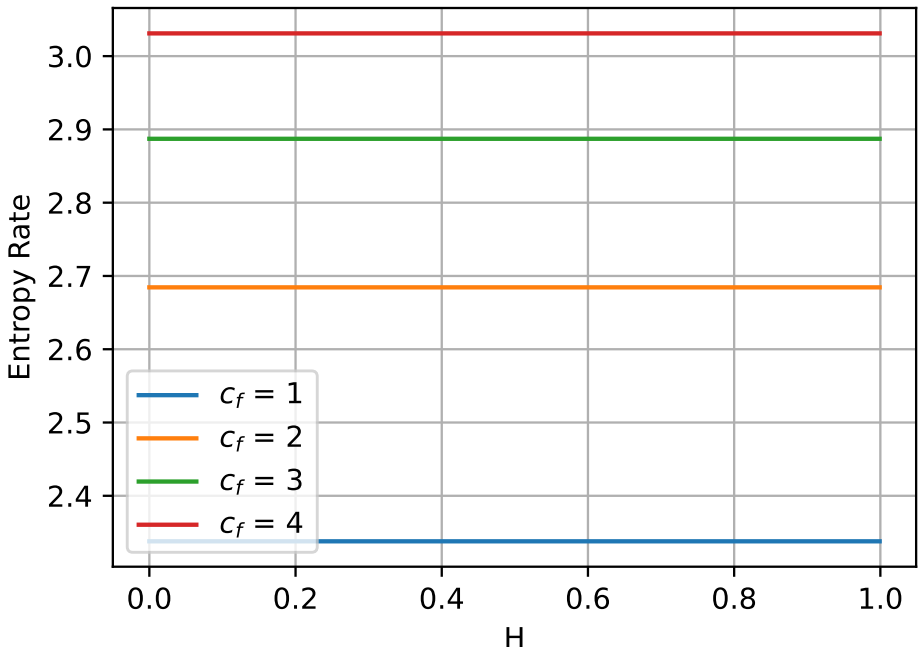

Figure 3.5: The entropy rate of ARFIMA(0,d,0) as a function of the Hurst parameter, $\mathcal{H}$, for c_f, $\sigma^2 = 1, 2, 3, 4$. Due to the relationship with the innovation variance the entropy rate function for ARFIMA(0,d,0) processes is constant for $\mathcal{H}$.

Remark. *Note that the value of the entropy rate function for an ARFIMA(0,d,0) process as* $\mathcal{H} \to 0$*, when* $\sigma^2 = 1$*, is*

$$h(\chi) = \frac{1}{2}\log(2\pi e\sigma^2) + \log\left(\Gamma\left(\frac{3}{2}\right)\right) - \frac{1}{2}\log(\Gamma(2)) \approx 1.298.$$

Which doesn't drop that far from its maximum of approximately 1.419*, particularly in comparison to the asymptotic value of* $-\infty$ *as* $\mathcal{H} \to 1$*.*

To complete this section of the analysis, we will consider the maximum of the entropy rate function of ARFIMA(0,d,0), and conclude which Hurst parameter has the highest uncertainty, in the sense of maximum differential entropy rate.

Theorem 3.2.4. *The differential entropy rate of ARFIMA(0,d,0) as a function of* $\mathcal{H}$ *attains its maximum at* $\mathcal{H} = 1/2$*.*

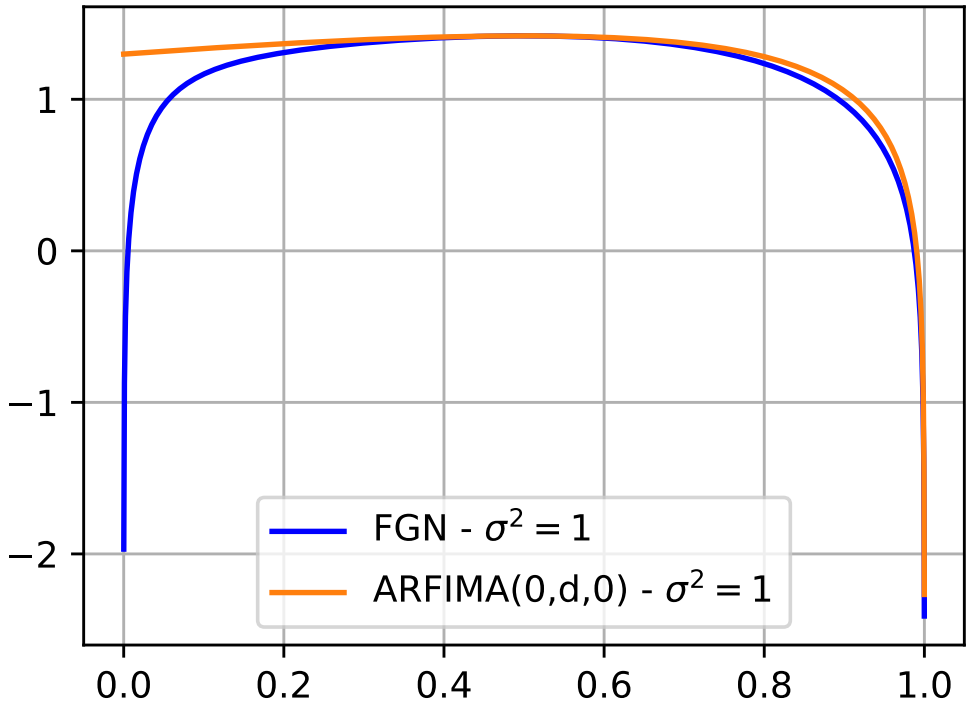

Figure 3.6: Comparison of the entropy rate as function of the Hurst parameter, for both ARFIMA(0,d,0) and FGN processes, with variance 1. It appears that the ARFIMA(0,d,0) process has an entropy rate which is greater than or equal to FGN for all values of $\mathcal{H}$. The negatively correlated portion falls away quickly as $\mathcal{H} \to 0$ for FGN but stays relatively high for the ARFIMA(0,d,0) process. The maximum of the functions coincide at $\mathcal{H} = 0.5$.

Proof. We differentiate the entropy rate function with respect to $\mathcal{H}$, and we get

$$\begin{aligned}
\frac{dh(\chi)}{d\mathcal{H}} &= \frac{d}{d\mathcal{H}}\left(\frac{1}{2}\log(2\pi e\sigma^2) + \log\left(\Gamma(\frac{3}{2}-\mathcal{H})\right) - \frac{1}{2}\log\left(\Gamma(2-2\mathcal{H})\right)\right), \\
&= \frac{d}{d\mathcal{H}}\left(\log(\Gamma(\frac{3}{2}-\mathcal{H}))\right) - \frac{d}{d\mathcal{H}}\left(\log(\Gamma(2-2\mathcal{H}))\right), \\
&= \frac{\Gamma(\frac{3}{2}-\mathcal{H})\psi\left(\frac{3}{2}-\mathcal{H}\right)}{\Gamma(\frac{3}{2}-\mathcal{H})} - \frac{\Gamma(2-2\mathcal{H})\psi(2-2\mathcal{H})}{\Gamma(2-2\mathcal{H})}, \\
&= \psi\left(\frac{3}{2}-\mathcal{H}\right) - \psi(2-2\mathcal{H}),
\end{aligned}$$

where $\psi(x)$ is the digamma function.

Then we set $\frac{dh(\chi)}{d\mathcal{H}} = 0$, and we have $\psi\left(\frac{3}{2} - \mathcal{H}\right) = \psi(2 - 2\mathcal{H})$. Since $\psi\left(\frac{3}{2} - \mathcal{H}\right)$ and $\psi(2 - 2\mathcal{H})$ are monotonically decreasing functions for $\mathcal{H} \in [0, 1]$ with the latter having a higher rate of decrease, this implies that the digamma functions intersect in at most one point. Since $\frac{3}{2} - \mathcal{H} = 2 - 2\mathcal{H}$ only when $\mathcal{H} = 1/2$, this implies that $h(\chi)$ achieves a unique maximum at this point. □

This aligns with our intuition, that the highest uncertainty occurs for this model when it is uncorrelated and equal to white Gaussian noise, as it simplifies to $X_n = \epsilon_n$, identical to FGN processes. This explains why the maxima coincide for the two processes, given the same process variance, although ARFIMA(0,d,0) appears to have a higher differential entropy across the entire parameter range, when not at $\mathcal{H} = 0.5$. This is observation is consistent with previous results in this area such as Burg's Theorem [32], that the AR and ARMA class of processes are the maximum entropy models given appropriate constraints on the covariances and impulse responses [68, 91]. Further research is required to understand whether the ARFIMA class has maximum entropy for processes with power-law decaying autocorrelation.

We have shown in this section that the behaviour for the ARFIMA(0,d,0) model differs from that of FGN in the behaviour of their CSRD processes. This is a surprising discovery and warrants further investigation. Both models, however, have much less uncertainty as the strength of the positive correlations increases, as well as a maximum uncertainty occurring for uncorrelated processes. Hence, we may be able to characterise the behaviour of LRD processes on the entropy rate as tending to $-\infty$ as the strength of correlations increases.

In remainder of the chapter we look at other information theoretic measures as way to characterise the behaviour of SRD and LRD processes.

3.3 Mutual Information and Excess Entropy for Long Range Dependent Gaussian Processes

In this section we continue analysing of the differential entropy rate for stochastic models that exhibit LRD. We investigate the links between the amount of entropy that is accumulated during the convergence of the conditional entropy to the entropy rate and the amount of information that is shared between the past and future of a stochastic process, the *excess entropy*. Then we will classify when this quantity converges and diverges for the range of the Hurst parameter for ARFIMA and FGN processes.

We extend the standard notion of mutual information for continuous-valued random variables that we defined in Definition B.1.4 to the special case of mutual information between past and future, $I_{\text{p-f}}$, which will measure the amount of information about the infinite future, given knowledge of the infinite past. We extend the definition of mutual information between past and future for continuous-valued random variables in the same way as the discrete-valued case from Definition 2.1.9, which in this case is

$$I_{\text{p-f}} = I(\{X_s, s < 0\}, \{X_s, s \geq 0\}),$$

where X_s is a continuous-valued random variables for all $s \in \mathbb{Z}$.

Exactly as in Theorem 2.1.5, alternative characterisations for the mutual information also apply in the case of differential entropy. One we will use in the analysis is $I(X;Y) = h(X) - h(X|Y) = h(Y) - h(Y|X)$ [41, pg.251].

There exist many processes that have infinite excess entropy but are not long range dependent. Some examples are given, including deterministic processes, in Crutchfield and Feldman [44].

Crutchfield and Feldman [44] analysed a quantity named excess entropy, $\sum_{n=1}^{\infty} (H(X_n|X_{n-1}, \ldots, X_0) - H(\chi))$ from Definition 2.1.10, for the Shannon entropy H and corresponding entropy rate $H(\chi)$, which has been shown to be equivalent to the mutual information between past and future. This has been named, with implicit interpretation, as stored information [155], effective measure complexity [78, 121], predictive information [133]. Importantly, it has been used to measure the convergence rate of the conditional entropy, based on past observations, to the entropy rate. We aim to extend this result to differential entropy, and then the question of classification of LRD processes via the amount of shared information can be made by the convergence rate to the entropy rate. We extend Definition 2.1.10 to the case of differential entropy.

Definition 3.3.1. *The differential* excess *entropy, E, of a stochastic process, $\{X_i\}_{i\in\mathbb{N}}$, is defined as,*

$$\begin{aligned} E &= \sum_{n=1}^{\infty} \left(h_e(n) - h(\chi)\right), \\ &= \lim_{n\to\infty} \left[h(X_n, \ldots, X_0) - nh(\chi)\right]. \end{aligned} \tag{3.7}$$

where

$$\begin{aligned} h_e(n) &= h(X_1, .., X_n) - h(X_1, ..., X_{n-1}), \\ &= h(X_n|X_{n-1}, ..., X_1). \end{aligned}$$

We have the tools available to make an explicit link between the mutual information between past and future and the excess entropy of a continuous-valued, discrete-time stochastic process. This is an exact analogue of Proposition 8 from Crutchfield and Feldman [44] and has been stated utilising a different approach by Ding and Xiang [54].

Theorem 3.3.1. *For a stationary, continuous-valued stochastic process, the mutual information between past and future, $I_{\text{p-f}}$, is equal to the differential excess entropy, E.*

Proof. The mutual information for a process X, with a past and future of n observations,

$$\begin{aligned} I[\{X_s, -n \leq s < 0\};\{X_s, 0 \leq s < n\}] \\ &= h(X_0, ..., X_{n-1}) - h(X_0, ..., X_{n-1}|X_{-n}, ..., X_{-1}), \\ &= \sum_{i=0}^{n-1} h(X_i|X_{i-1}, ..., X_0) - \sum_{i=0}^{n-1} h(X_i|X_{i-1}, ..., X_{-n}), \\ &= \sum_{i=0}^{n-1} \Big(h(X_i|X_{i-1}, ..., X_0) - h(X_i|X_{i-1}, ..., X_{-n}) \Big), \end{aligned}$$

by the chain rule of differential entropy[41, pg. 253]. Then we consider the mutual information between past and future, by taking the limit of the above expression as $n \to \infty$, which leads to

$$\begin{aligned} I_{\text{p-f}} &= \lim_{n\to\infty} \left[\sum_{i=0}^{n-1} \Big(h(X_i|X_{i-1}, ..., X_0) - h(X_i|X_{i-1}, ..., X_{-n}) \Big) \right], \\ &= \lim_{n\to\infty} \left[\sum_{i=0}^{\infty} \Big(h(X_i|X_{i-1}, ..., X_0) - h(X_i|X_{i-1}, ..., X_{-n}) \Big) \mathbb{1}_{\{i<n\}} \right]. \end{aligned}$$

We define the sequence of measurable functions, $f_n(i)$ as

$$f_n(i) = \Big(h(X_i|X_{i-1}, ..., X_0) - h(X_i|X_{i-1}, ..., X_{-n}) \Big) \mathbb{1}_{\{i<n\}},$$

and we define the function, $g(i)$ as

$$g(i) = h(X_i|X_{i-1}, ..., X_0) - h(\chi).$$

We want to show that $|f_n(i)| \leq g(i)$ for all n and for all $i \in \mathbb{N}$. In this case it is equivalent to showing that $f_n(i) \leq g(i)$, since $f_n(i) \geq 0$ for all $n, i \in \mathbb{N}$,

as the second term of $f_n(i)$ conditions on more random variables, and since conditioning cannot increase entropy this implies that $h(X_i|X_{i-1},...,X_0) \geq h(X_i|X_{i-1},...,X_{-n})$. We consider two cases, $i < n$ and $i \geq n$, separately. In the case, $i \geq n$, we have that $f_n(i) = 0$, and since $g(i) \geq 0$ for all i, this implies that $f_n(i) \leq g(i)$. Considering the second case, $i < n$, we have that

$$f_n(i) = \Big(h(X_i|X_{i-1},...,X_0) - h(X_i|X_{i-1},...,X_{-n}) \Big),$$

and therefore,

$$g(i) - f_n(i) = h(X_i|X_{i-1},...,X_{-n}) - h(\chi).$$

Again, since conditioning does not increase entropy and the characterisation of entropy rate for stationary processes from Theorem B.2.1 this implies that $g(i) - f_n(i) \geq 0$ and therefore $g(i) \geq f_n(i)$ for all n, i such that $i < n$. Then we can apply the monotone convergence theorem [57, pg. 27], since $f_n(i) \to h(X_i|X_{i-1},...,X_0) - h(\chi)$ pointwise, this implies that

$$\begin{aligned} I_{\text{p-f}} &= \lim_{n\to\infty} \sum_{i=0}^{\infty} \Big(h(X_i|X_{i-1},...,X_0) - h(X_i|X_{i-1},...,X_{-n}) \Big) \mathbb{1}_{\{i<n\}}, \\ &= \sum_{i=0}^{\infty} \lim_{n\to\infty} \Big(h(X_i|X_{i-1},...,X_0) - h(X_i|X_{i-1},...,X_{-n}) \Big) \mathbb{1}_{\{i<n\}}, \\ &= \sum_{i=0}^{\infty} \Big(h(X_i|X_{i-1},...,X_0) - h(\chi) \Big). \end{aligned}$$

□

Remark. *This proof is similar to that of Proposition 8 from Crutchfield and Feldman [44]. However, it is more rigorous since the limit is kept out the front of the sum while simultaneously applied to the second term in the sum. This approach using monotone convergence can resolve the issue in their proof.*

In Crutchfield and Feldman [44], they analyse the excess entropy of discrete random variables to understand the convergence rate of the conditional entropy to the entropy rate. We will analyse when the excess entropy converges or diverges for the ARFIMA and FGN classes of processes, over the range of the Hurst parameter. Then we will apply conclusions made to the mutual information between past and future.

We begin by classifying the convergence and divergence of excess entropy for the ARFIMA class of processes

Theorem 3.3.2. *For stationary and invertible ARFIMA(p,d,q) the excess entropy is finite if and only if $d = 0$, i.e. $\mathcal{H} = 1/2$.*

Proof. From Theorem 9.4 of Debowski [50], we have for Gaussian processes that

$$E < \infty \iff \sum_{n=1}^{\infty} n\alpha_n^2 < \infty \text{ and } |\alpha_n| < 1, \forall n \in \mathbb{N},$$

where α_n is the partial autocorrelation function from Definition 2.2.2 We will analyse the two different cases $d \in (-1/2, 1/2) \setminus \{0\}$ and $d = 0$ separately, starting with the former. From Theorem 2.5 of Inoue [94], we have that

$$\alpha_n \sim \frac{|d|}{n},$$

for all $d \in (-1/2, 1/2) \setminus \{0\}$. Since d doesn't depend on n, we get

$$\sum_{n=1}^{\infty} n \left(\frac{|d|}{n}\right)^2 = d^2 \sum_{n=1}^{\infty} \left(\frac{1}{n}\right) = \infty.$$

Which implies that $E = \infty$. Now we consider the case $d = 0$, which are ARMA processes. From Theorem 7.1 of Inuoe [94], we have that

$$|\alpha_n| \leq Mr^n,$$

for a constant $M > 0$ and where $R < r < 1$ and $R = \max\left(\frac{1}{|u_1|}, ..., \frac{1}{|u_q|}\right)$, where $u_1, ..., u_q$ are the complex zeros of the moving average polynomial. Therefore,

$$\begin{aligned} \sum_{n=1}^{\infty} n\alpha_n^2 &\leq \sum_{n=1}^{\infty} n\left(Mr^n\right)^2, \\ &= M^2 \sum_{n=1}^{\infty} nr^{2n} < \infty, , \text{ since } r < 1. \end{aligned}$$

Which implies that $E < \infty$. □

This leads to a simple corollary that the result also applies to the mutual information between past and future.

Corollary 3.3.2.1. *For ARFIMA processes the mutual information between past and future is finite if and only if $\mathcal{H} = 1/2$.*

Proof. This is shown by combining Theorem 3.3.2 and Theorem 3.3.1. □

This is an interesting result which indicates that the boundary between finite and infinite excess entropy is between SRD and LRD/CSRD ARFIMA processes. That is, any persistent correlations, whether positive or negative, induce infinite excess entropy, and not just positive correlattions *i.e.,* LRD, as has been suspected in previous work such as the discussion in section 3.4 of Li [120] and Ding and Xiang [54]. The assumptions in the discussion around the equivalences are due to the second part of Theorem 1 of Li [120] which gives the following statement. *If the spectral density $f(\lambda)$ is continuous and $f(\lambda) > 0$, then $I_{\text{p-f}}$ is finite if and only if the autocovariance function satisfies the condition $\sum_{k=1}^{\infty} k\gamma(k)^2 < \infty$.* Given the result in Theorem 3.3.2 the condition $f(\lambda) > 0$ is critical. Since, in the case of CSRD ARFIMA processes the spectral density has a root at 0. Therefore to classify the behaviour of the excess entropy across the entire range of of $\mathcal{H}$ we need to take a different approach than Li [120] and Ding and Xiang [54], and adding some additional regularity conditions on the infinite autoregressive and moving average representations. This builds upon the work of Inoue [94], which classified the behaviour of the partial autocorrelation function for these processes with the following conditions.

Condition 3.3.1. *$\{X_n\}_{n\in\mathbb{N}}$ is a purely non-deterministic process. That is, for the Hilbert space, $\mathbb{H}$, spanned by $\{X_k\}_{k\in\mathbb{Z}}$ in the probability space $L_2(\Omega, \mathcal{F}, P)$, then*

$$\bigcap_{n=-\infty}^{\infty} \mathbb{H}_{(-\infty,n]} = \{0\}.$$

The next conditions are on the coefficients of the AR(∞), and MA(∞), representations with coefficients $\{a_n\}_{n\in\mathbb{N}}$ and $\{c_n\}_{n\in\mathbb{N}}$ respectively. That is,

$$\epsilon_n = \sum_{k=1}^{\infty} a_k X_{n-k},$$
$$\text{and, } X_n = \sum_{k=1}^{\infty} c_k \epsilon_{n-k}.$$

Condition 3.3.2. *$\{a_n\}_{n\in\mathbb{N}}$ is eventually decreasing to zero.*

Condition 3.3.3. *$\{c_n\}_{n\in\mathbb{N}}$ is eventually decreasing to zero and $c_n \geq 0$ for all $n \geq 0$.*

With the addition of these conditions we are able to classify the behaviour of processes with autocovariance functions of the form $\gamma(n) \sim Cn^{2d-1}$, where $d = \mathcal{H} - 1/2$. Note that we utilise the results from several papers from Inuoe [93, 94] where they consider behaviour of autocovariance functions of the form, $\gamma(n) \sim l(n)n^{2d-1}$, where $l(n)$ is a slowly varying function, that is

$$\lim_{n\to\infty} \frac{l(\lambda n)}{l(n)}, \quad \forall \lambda > 0.$$

Theorem 3.3.3. *For stationary Gaussian processes with an autocovariance function $\gamma(n) \sim Cn^{2d-1}$, for a constant C, obeying conditions 3.3.1, 3.3.2, and 3.3.3, the excess entropy is finite if and only if $d = 0$, i.e., $\mathcal{H} = 1/2$.*

Proof. We will split the proof into the cases where $d \in (0, 1/2)$, $d = 0$, and $d \in (-1/2, 0)$. Note that we have the autocovariance in the form $\gamma(n) \sim l(n)n^{2d-1}$, with $l(n) = C$ trivially a slowly varying function, since $\forall \lambda > 0$ we have that

$$l(\lambda n) = l(n).$$

For $d \in (0, 1/2)$, by Theorem 6.1 of Inoue [94], we have that

$$\alpha_n \sim \frac{d}{n}.$$

Therefore, as in Theorem 3.3.2 we have that $E = \infty$. In the case $d = 0$, we consider the integral $\tilde{l}(n) = \int_B^n \frac{l(n)}{s} ds = \int_B^n \frac{C}{s} ds$. Since the behaviour is independent of B by Inoue [94], we take $B = 1$. The integral diverges as $n \to \infty$ because,

$$\begin{aligned} \int_1^n \frac{C}{s} ds &= C \int_1^n \frac{1}{s} ds, \\ &= C \left[\log s\right]_1^n, \\ &\to \infty \text{ as } n \to \infty. \end{aligned}$$

By Theorem 6.1 of Inoue [94] we have that

$$\begin{aligned} \alpha_n &\sim \frac{l(n)}{2n\tilde{l}(n)}, \\ &= \frac{C}{2nC\log(n)}, \\ &= \frac{1}{2n\log(n)}. \end{aligned}$$

Therefore,

$$\sum_{n=1}^{\infty} n\alpha_n^2 \sim \sum_{n=1}^{\infty} n\left(\frac{1}{2n\log(n)}\right)^2, \\ = \frac{1}{4}\sum_{n=1}^{\infty}\left(\frac{1}{n\log^2(n)}\right).$$

Since we consider the tail behaviour of the sum for divergence, we have that

$$\frac{1}{4}\sum_{n=2}^{\infty} n\left(\frac{1}{n\log^2(n)}\right) < \infty,$$

by Knopp [105, pg. 63]. We conclude that $\sum_{n=1}^{\infty} n\alpha_n^2 < \infty$ and therefore $E < \infty$. Finally we consider $d \in (-1/2, 0)$. By Theorem 6.1 of Inoue [94], we have that

$$\alpha_n \sim \frac{\gamma(n)}{\sum_{k=-n}^{n}\gamma(k)}, \\ \sim \frac{n^{2d-1}}{\sum_{k=-n}^{n} k^{2d-1}}, \text{ since } \gamma(n) \sim Cn^{2d-1}.$$

The denominator is the consecutive sum of powers, which has the following well known asymptotic form

$$\sum_{k=-n}^{n} k^{2d-1} \sim \frac{n^{2d}}{2d}.$$

Note that this asymptotic sum indicates that $\sum_{k=-n}^{n}\gamma(k) \to 0$ as $n \to \infty$, a well-known behaviour for CSRD processes.
Therefore, as in Theorem 3.3.2 we have $E = \infty$. □

We provide an identical corollary to Corollary 3.3.2.1 for processes that meet .

Corollary 3.3.3.1. *For stationary Gaussian processes with an autocovariance function $\gamma(n) \sim Cn^{2d-1}$, for a constant C, obeying conditions 3.3.1, 3.3.2, and 3.3.3, the mutual information between past and future is finite if and only if $\mathcal{H} = 1/2$.*

Proof. This is shown by combining Theorem 3.3.3 and Theorem 3.3.1. □

Theorem 3.3.3 provides a classification of the behaviour of excess entropy for processes where the autocovariance function $\gamma(n) \sim Cn^{2d-1}$, in future we may be able to extend this to the more general case where $\gamma(n) \sim l(n)n^{2d-1}$, where $l(n)$ is a slowly varying function. Using the same argument as above we can show the same result in the case that $d \in (0, 1/2)$, and $d = 0$. Following on from the comments at the end of Section 6 of Inoue [94], it would be interesting to understand how generally the relation,

$$\alpha_n \sim \frac{\gamma(n)}{\sum_{k=-n}^{n} \gamma(k)},$$

holds, and if a potential classification of LRD/SRD/CSRD processes by their excess entropy exists, based on improved knowledge of the asymptotics of the partial autocorrelation function.

3.4 Conclusion

In this chapter, we are concerned with the behaviour of the differential entropy rate to understand and characterise the behaviour of LRD and SRD processes. Analysing two common LRD processes, FGN and ARFIMA(0,d,0), we have shown that the maximum occurs in the absence of correlations, *i.e.*, $\mathcal{H} = 0.5$, and the differential entropy rate tends to the minimum, $-\infty$ as the strength of positive correlations increase, *i.e.*, as we receive more information from correlations, the entropy of the process decreases. However, there is very different behaviour for negatively correlated processes, where ARFIMA(0,d,0) processes do not tend to $-\infty$ as the strength of the negative correlations increases. Further research is required to understand this behaviour for these processes.

In addition, we have made a link, similar to Shannon entropy, between the mutual information between past and future and excess entropy, meaning that the amount of shared information between the complete past of future of a process is the same as the additional information that accrues when converging to the entropy rate, based on past observations. This leads to a characterisation of processes that have power-law decay of their covariance function, by the properties of their mutual information between past and future. This characterisation is that processes which are LRD or CSRD, have infinite mutual information between past and future.

Chapter 4

Shannon Entropy Rate Characterisation of Long Range Dependent Markov Chains

Markov chains are a class of discrete time stochastic processes characterised by the property that the transition probability is only dependent on knowledge of the current state. This is called the Markov property and is the defining property of a large class of processes in discrete and continuous time, called Markov processes.

We will introduce the relevant properties and concepts of Markov chains in the first part of this chapter and in Appendix D. Less often studied are LRD Markov chains. In addition to the Markov property, LRD Markov chains are characterised by the infinite second moment of the return time random variable and the growth of the variance of the counting function of the number of visits to a state. We show here that LRD Markov chains have similar behaviour to LRD Gaussian processes, except on discrete state spaces.

In this chapter, we extend our analysis of entropy rate convergence of stochastic processes by considering Markov chains. We show that there is a connection between the rate of convergence of a Markov chain to its stationary distribution, and the rate of convergence of the conditional entropy to the entropy rate. We show that the convergence rate depends on the existence of moments of the return time distribution. When all moments of the return time random variable are finite we have geometric convergence, shifting to sub-geometric convergence if any infinite moment exists. By adding a common condition for LRD processes, that the complementary cumulative distribution function of the return time random variable has a power law tail, *i.e.*, $\mathbb{P}(T > n) \sim cn^{-\alpha}$. We can then show that the convergence to

the stationary distribution is at a power-law rate, and that the convergence rate is $O(n^{2-2\mathcal{H}})$, where n is the number of data and note that $\mathcal{H}$ and α are related as $\mathcal{H} = 1 - \alpha/2$, similar to other results on LRD processes.

4.1 Markov Chain Background

In this section, we will define Markov chains. This will allow us to analyse their information theoretic properties, such as the entropy rate, and characterisations of LRD on these processes, via their convergence rates. We leave a more through discussion of the relevant properties and concepts to Appendix D. This appendix outlines the background of concepts such as irreducibility, aperiodicity, hitting times and the stationary distribution. These are required to understand LRD Markov chains, their definition and the calculation of the entropy rate of a Markov chain, but the definitions are standard. However, we provide the key definitions here.

Definition 4.1.1. *A discrete-time stochastic process, $\mathcal{X} = \{X_n\}_{n\in\mathbb{Z}^+}$ on a countable state space, Ω, is called a Markov chain if for every n*

$$\mathbb{P}(X_n = x_n | X_{n-1} = x_{n-1}, \ldots, X_1 = x_1) = \mathbb{P}(X_n = x_n | X_{n-1} = x_{n-1}).$$

We will discuss some useful additional properties, the first property is time-homogeneity, which states that the probability of transitioning between states doesn't change over time.

Definition 4.1.2. *We call a Markov chain, $\mathcal{X} = \{X_n\}_{n\in\mathbb{Z}^+}$, time-homogeneous if for every n and all states $i, j \in \Omega$,*

$$\mathbb{P}(X_{n+1} = j | X_n = i) = \mathbb{P}(X_n = j | X_{n-1} = i).$$

In this case, we call the probabilities,

$$p_{ij} = \mathbb{P}(X_n = j | X_{n-1} = i),$$

the transition probabilities. We have that for every state, $i \in \Omega$ that the transition probabilities sum to 1,

$$\sum_{j\in\Omega} p_{ij} = 1,$$

since at each step of the chain the process must be in a state. We generalise the transition probabilities to an arbitrary number of steps, that we call the k-step transition probabilities,

$$p_{ij}^{(k)} = \mathbb{P}(X_{n+k} = j | X_k = i).$$

We provide additional background of Markov chains in Appendix D.

4.2 LRD Markov Chains

LRD processes defined on discrete state spaces in discrete time have been much less studied than on continuous state spaces. The concept of LRD has been defined and characterised on general point processes [46], renewal processes [45] and Markov renewal processes [176].

A key insight for point and renewal processes is that LRD can be defined with respect to the second-order behaviour of the counting function of the number of events in an interval, $N(0,t]$. A definition of LRD for point and renewal processes is given by the variance of this function. A point or renewal process is said to have LRD if the growth of the variance of the function is faster than linear [45, 46], that is

$$\limsup_{t\to\infty} \frac{\mathrm{Var}(N(0,t])}{t} = \infty.$$

This was extended to the case of irreducible Markov chains in discrete time on countable state spaces by Carpio and Daley [27], by using the variance of the counting function of the number of visits of the Markov chain, $\{X_n\}_{n\in\mathbb{Z}^+}$, to state i up until time n, $N_i(0,n]$. That is, the variance of

$$N_i(0,n] = \sum_{k=1}^{n} \mathbb{1}_{\{X_k=i\}}.$$

This is a natural extension, since the rate of increase of the variance of the counting function is a property of the communicating class. This leads to the following definition of LRD on Markov chains.

Definition 4.2.1. *An irreducible and aperiodic Markov chain, $\{X_n\}_{n\in\mathbb{Z}^+}$, is said to be long range dependent if*

$$\limsup_{n\to\infty} \frac{Var(N_i(0,n])}{n} = \infty.$$

Otherwise, we say the Markov chain is short range dependent [27].

Carpio and Daley [27] show that this applies to all states in the communicating class, and therefore is independent of the particular state i used in the definition.

A random variable which characterises the long term behaviour of Markov chains is the return time random variable, T_{ii}, to a particular state i. For example, this is used in classifying Markov chains as positive recurrent, null recurrent and transient. The random variable is defined as

$$T_{ii} = \inf\{n \geq 1 : X_n = i, X_0 = i\}.$$

The return times of LRD Markov chains were shown, in Lemma 1 of Carpio and Daley [27], to have an infinite second moment. Oguz and Anantharan [134] extended this result to show that other functions of Markov chains also have an infinite second moment, given some regularity conditions on the function.

A related concept to infinite moments is that of heavy-tailed or sub-exponential distributions, which are distributions whose tail decay is slower than exponential.

Definition 4.2.2. *A probability distribution of a random variable, X, with cumulative distribution function, F, is called heavy-tailed, or sub-exponential, if and only if* $\forall t > 0$

$$\int_{-\infty}^{\infty} e^{tx} dF(x) = \infty.$$

We will discuss some of the aspects of Markov chain stationary distribution convergence with reference to return time distributions with this property. Note that in this chapter we will be considering discrete return time random variables with heavy tails, and the integral will become an infinite sum.

We aim to answer how LRD affects the rate of convergence of important quantities for Markov chains, which are often used to describe the properties of the Markov chain itself, its stationary distribution and its entropy rate.

In this chapter we show that, like other types of LRD processes, convergence rates of certain quantities of LRD Markov chains are considerably slower than for short range dependent processes. We prove that the convergence of the n-step transition probabilities to the stationary distribution occurs at a rate that is a power law, that is n^c as $n \to \infty$ for some $c \in (0, 1)$, *i.e.,* slower than linear in its rate of convergence.

We extend this idea to the entropy rate, which for Markov chains is a function of the stationary distribution and the n-step transition probabilities and can be thought of as the asymptotic rate of new information for the process. This is an analogue of Chapter 3 showing that the conditional entropy converges to the entropy rate more slowly for LRD discrete time Gaussian processes than their short range dependent counterparts. However, here we consider Markov chains on discrete state spaces which are more relevant to many real contexts, such as the analysis of natural language.

We are able to show that for all positive recurrent ergodic Markov chains the rate of convergence to the entropy rate is the same as the rate of convergence of the n-step transitions to the stationary distribution. In addition, we show that this implies that for LRD Markov chains that the mutual information between past and future is infinite.

4.3 Entropy Rate Convergence relationship with Mixing Time

In this section we discuss the convergence of the conditional entropy of a Markov chain to its entropy rate. The entropy rate is the asymptotic limit of the average information from each additional observation of the Markov chain, and is used as a measure of uncertainty or complexity of a stochastic process. We show that there is an equivalence between the rate of convergence of the n-step transition probabilities to the stationary distribution and the convergence of the conditional entropy to the entropy rate. From Carpio and Daley [27] the convergence rate is a property of the entire communicating class, so any convergence rate in a particular state applies to all states in an ergodic chain. We demonstrate that the behaviour of the excess entropy of LRD Markov chains is consistent with the entropy rate of Gaussian LRD processes. We conclude for ergodic Markov chains that LRD is characterised by slow convergence and an infinite amount of shared information between the past and future of the process.

First, we note that the entropy rate of an ergodic Markov chain is the same as the limit of the conditional entropy, a result commonly seen for stationary Markov chains [41, pg. 75]. However, we extend the result to a Markov chain starting from an arbitrary initial state.

Lemma 4.3.1. *For an ergodic Markov chain χ the entropy rate is equal to the limit*

$$H[\chi] = \lim_{n\to\infty} H[X_n|X_{n-1}, \ldots, X_0].$$

Proof. The proof is omitted as it follows the same argument as Theorem 4.2.1 in Cover and Thomas [41]. □

Next we note that the entropy rate of an ergodic Markov chain is the same as a stationary Markov chain, using Lemma 4.3.1. This is a useful result as it provides the intuition that the entropy rate forgets about its initial state due to the process's ergodicity. We make one additional assumption, required for the calculation of the entropy rate over a countable set, that the conditional entropy given knowledge of which state the process is in, is finite; formally, that is

$$\mathbb{H}\left[X_n|X_{n-1} = i\right] = -\sum_{j\in\Omega} p_{ij} \log p_{ij} < \infty.$$

This is not an onerous assumption, since for interesting analysis we require that the entropy of a random variable is finite. We next provide the entropy rate of ergodic Markov chains.

Theorem 4.3.2. *The entropy rate of an ergodic Markov Chain is*

$$H[\chi] = -\sum_i \sum_j \pi_i p_{ij} \log p_{ij}.$$

Proof. The proof is omitted as it follows the same argument as Theorem 4.2.4 in Cover and Thomas [41]. □

Therefore we have an explicit form for the entropy rate of a Markov chain that is ergodic, rather than just stationary, and therefore the expression is independent of the initial state of the chain.

Next, by analysing the limit of the conditional entropy conditioned on the previous observations of the Markov chain, we show that the convergence to the entropy rate is equivalent to the convergence to the stationary distribution. This provides another perspective on long range dependence, that the convergence to the entropy rate, the average new information from a random variable of a stochastic process, is slower. This equivalence is given by the following theorem.

Theorem 4.3.3. *The convergence of the conditional entropy of an ergodic, positive recurrent Markov Chain to its entropy rate is at the same rate as the convergence to the stationary distribution.*

Proof. We define the initial distribution as $\theta = \{\theta_i\}_{i\in\Omega}$, *i.e.*, $\mathbb{P}(X_0 = i) = \theta_i$. Then the conditional entropy of X_1 given X_0 is

$$\begin{aligned} H[X_1|X_0] &= -\sum_i \sum_j \mathbb{P}(X_0 = i, X_1 = j) \log \mathbb{P}(X_1 = j|X_0 = i), \\ &= -\sum_i \sum_j \theta_i p_{ij} \log p_{ij}. \end{aligned}$$

Considering the conditional entropy of X_n given the history up to step $n-1$, of the previously observed random variables $X_0, \ldots, X_{n-1}$, on the states $i_0, i_1, \ldots, i_n \in \Omega$, where $p_{i_0 i_1} = \mathbb{P}(X_1 = i_1|X_0 = i_0)$,

$$\begin{aligned} &H[X_n|X_{n-1}, \ldots, X_0] \\ &= -\sum_{i_n} \ldots \sum_{i_0} \mathbb{P}(X_0 = i_0, \ldots, X_n = i_n) \log \mathbb{P}(X_n = i_n|X_{n-1} = i_{n-1}, \ldots, X_0 = i_0), \\ &= -\sum_{i_n} \ldots \sum_{i_0} \theta_{i_0} p_{i_0 i_1} \ldots p_{i_{n-1} i_n} \log p_{i_{n-1} i_n}. \end{aligned}$$

Where we have split the joint probability into a path of the conditional probabilities of transitions on the states, $i_0, i_1, \ldots, i_n \in \Omega$, i.e.,

$$\begin{aligned}\mathbb{P}(X_0 = i_0, \ldots, X_n = i_n) &= \mathbb{P}(X_0 = i_0)\mathbb{P}(X_1 = i_1 | X_0 = i_0) \ldots \mathbb{P}(X_n = i_n | X_{n-1} = i_{n-1}), \\ &= \theta_{i_0} p_{i_0 i_1} \ldots p_{i_{n-1} i_n}.\end{aligned}$$

By summing through the intermediate states, utilising the (n-1)-step transition probability of transitioning between states i_0 and i_{n-1}, $p^{(n-1)}_{i_0 i_{n-1}}$, we get

$$\begin{aligned}H[X_n | X_{n-1}, \ldots, X_0] &= -\sum_{i_n}\sum_{i_{n-1}}\sum_{i_0} \theta_{i_0} p^{(n-1)}_{i_0 i_{n-1}} p_{i_{n-1} i_n} \log p_{i_{n-1} i_n}, \\ &= -\sum_{i_n}\sum_{i_{n-1}} p_{i_{n-1} i_n} \log p_{i_{n-1} i_n} \left(\sum_{i_0} \theta_{i_0} p^{(n-1)}_{i_0 i_{n-1}} \right). \quad (4.1)\end{aligned}$$

Note that the term $p_{i_{n-1} i_n} \log p_{i_{n-1} i_n}$ quantifies the information contained in the transitions. As $n \to \infty$, the sum $\sum_i \theta_i p^{(n-1)}_{ij} \to \pi_j$, since a positive recurrent chain has a stationary distribution and by the ergodicity of the chain, it converges to the stationary distribution from any state by Theorem D.0.3. Taking the limit of Equation (4.1) shows that the convergence of the conditional entropy to the entropy rate depends on the rate of convergence of $\sum_i \theta_i p^{(n)}_{ij} \to \pi_j$. □

This theorem shows that the convergence rate of other quantities for Markov chains are intimately connected to the convergence rate of the stationary distribution.

Now we consider some more information theoretic quantities, the excess entropy, which considers the "additional" information that accrues in the convergence to the entropy rate from the conditional entropy and the mutual information between past and future, measuring the amount of information that is shared between the infinite past and infinite future of processes. For processes on countable sets, the excess entropy was shown to be equivalent to the mutual information [44]. We define both here and show that in the case of LRD Markov chains that these two measures are infinite, providing another characterisation of LRD. Note that we denote the excess entropy, E, in line with previous work and use the expectation operator $\mathbb{E}[\cdot]$.

A definition of LRD, suggested in Li [120], is that the mutual information between past and future is infinite. In Chapter 3 we showed that for many classes of LRD processes, such as ARFIMA processes and those meeting some additional conditions, that the excess entropy is infinite [64]. We show that in the case of LRD Markov Chains the excess entropy is infinite, by the limits of the quantities, $Q^n_{ij} = \sum_{r=1}^{n} \left(p^{(r)}_{ij} - \pi_j \right)$, for $i, j \in \Omega$. Carpio and Daley [27]

used Q_{ij}^n to show that the state space must be infinite in the case of LRD. We use it to illustrate the slow convergence, and to reinforce an entropic perspective on LRD.

First we prove a lemma that is used to show the slow convergence behaviour of LRD Markov chains.

Lemma 4.3.4. *For any sequence* $\{x_i\}_{i=1}^{\infty}$ *such that*

$$\lim_{n\to\infty} \sum_{i=1}^{n} x_i \to \infty,$$

we can form a partition $\{R_m\}$*,*

$$R_m = \{i_m, \ldots, i_{m+1} - 1\}$$

with $i_1 = 1$*, such that*

$$\sum_{i\in\{R_m\}} x_i \geq 0$$

for all $m \in \mathbb{N}$*.*

Proof. For a limit of a diverging partial sum

$$\sum_{i=1}^{\infty} x_i \to \infty,$$

there exists by definition an $N \in \mathbb{N}$ such that

$$\sum_{i=1}^{N} x_i \geq a,$$

for any $a \in \mathbb{R}$. Therefore we choose $a = 0$, $i_1 = 1$ and $i_2 - 1 = N$. This defines the first set in the partition

$$R_1 = \{i_1, \ldots, i_2 - 1\},$$

such that

$$\sum_{i\in R_1} x_i \geq 0.$$

We then create a new sequence $x_i^{(1)}$ which begins with first element x_{N+1}, that is the new sequence starts at the index immediately after N. The

new sequence diverges since removing a finite portion of the beginning of a divergent sequence produces a sequence that still diverges. That is,

$$\lim_{n\to\infty} \sum_{i=1}^{n} x_i^{(1)} \to \infty.$$

We can repeat this argument for R_2, with $a = 0$, $i_2 = N+1$, and $i_3 - 1 = M$, for an $M \in \mathbb{N}$. As before we define a new sequence $x_i^{(2)}$ with its first element x_{M+1}, the element with index $M+1$, and repeat the process. Therefore, we can continue and creates a partition, $\{R_m\} = R_1, R_2, R_3, \ldots$ with the required properties. That is,

$$R_m = \{i_m, \ldots, i_{m+1} - 1\}$$

such that

$$\sum_{i \in R_m} x_i \geq 0,$$

for all $m \in \mathbb{N}$. □

Theorem 4.3.5. *A countable state LRD Markov chain has infinite excess entropy.*

Proof. By Carpio and Daley [27], we have for LRD Markov chains

$$\lim_{n\to\infty} \sum_{r=1}^{n} \left(\sum_k \theta_k p_{ki}^{(r-1)} - \pi_i \right) = \infty.$$

From Lemma 4.3.4 we can form sets

$$R_m = \{i_m, \ldots, i_{m+1} - 1\}$$

such that

$$S_m^{(i)} = \sum_{r \in R_m} \left(\sum_k \theta_k p_{ki}^{(r-1)} - \pi_i \right)$$

and $S_m^{(i)} \geq 0$ for all $m \in \mathbb{N}$, and for any given state i. Since the transition probabilites, p_{ij} are such that $0 \leq p_{ij} \leq 1$, we have

$$-p_{ij} \log p_{ij} \; S_m^{(i)} \geq 0$$

for all i, j and m. By Tonelli's theorem, we have that

$$\sum_{i,j}\sum_{m} -p_{ij}\log p_{ij} S_m^{(i)} = \sum_{m}\sum_{i,j} -p_{ij}\log p_{ij} S_m^{(i)} \tag{4.2}$$

The LHS of Equation 4.2 is

$$\sum_{i,j}\sum_{m} -p_{ij}\log p_{ij}\ S_m^{(i)} = \infty,$$

by Carpio and Daley [27]. The RHS of equation 4.2 is

$$\begin{aligned}
\sum_{m}\sum_{i,j} -p_{ij}\log p_{ij} S_m^{(i)} &= \sum_{m}\sum_{i,j} -p_{ij}\log p_{ij} \sum_{r\in R_m}\left(\sum_k \theta_k p_{ki}^{(r-1)} - \pi_i\right) \\
&= \sum_{m}\sum_{i,j}\sum_{r\in R_m} -p_{ij}\log p_{ij}\left(\sum_k \theta_k p_{ki}^{(r-1)} - \pi_i\right) \\
&= \sum_{m}\left[\sum_{i,j}\sum_{r\in R_m}\left(-p_{ij}\log p_{ij}\sum_k \theta_k p_{ki}^{(r-1)} + p_{ij}\log p_{ij}\pi_i\right)\right] \\
&= \sum_{m}\left[\sum_{i,j}\sum_{r\in R_m}\left(-p_{ij}\log p_{ij}\sum_k \theta_k p_{ki}^{(r-1)}\right) + \sum_{i,j}\sum_{r\in R_m} p_{ij}\log p_{ij}\pi_i\right].
\end{aligned}$$

We can swap the order of the summations since the terms all have the same sign, which gives

$$\begin{aligned}
\sum_{m}\sum_{i,j} -p_{ij}\log p_{ij} S_m^{(i)} &= \sum_{m}\left[\sum_{r\in R_m}\sum_{i,j}\left(-p_{ij}\log p_{ij}\sum_k \theta_k p_{ki}^{(r-1)}\right) + \sum_{r\in R_m}\sum_{i,j} p_{ij}\log p_{ij}\pi_i\right], \\
&= \sum_{m}\sum_{r\in R_m}\left[\sum_{i,j}\left(-p_{ij}\log p_{ij}\sum_k \theta_k p_{ki}^{(r-1)}\right) + \sum_{i,j} p_{ij}\log p_{ij}\pi_i\right], \\
&= \sum_{n=1}^{\infty} H[X_n|X_{n-1},\ldots,X_1] - H[\chi].
\end{aligned}$$

Therefore, the excess entropy of LRD Markov chains is infinite. □

This result leads to a corollary that classifies LRD Markov chains.

Corollary 4.3.5.1. *The mutual information between past and future of a Markov chain is infinite if the Markov chain is LRD*

Proof. Theorem 4.3.5 and Proposition 8 of Crutchfield and Feldman [44] imply the result. □

Therefore, we have shown that for Markov chains that LRD implies that the excess entropy is infinite. This supports the notion that this definition of LRD exhibits the "right" behaviour for Markov chains. As is common with other definitions of LRD, it is characterised by slow convergence to quantities, such as sample mean, and we have shown in this section that this behaviour extends to a common way of measuring uncertainty of stochastic processes, the entropy rate. Most processes that have been developed that exhibit this behaviour are defined on a continuous state space, so these results show that even for discrete valued spaces this behaviour exists. This behaviour is the result of the infinite second moment of the return time random variable, hence this is the simplest discrete valued model of which this behaviour occurs. This result might also lead to useful means of discriminating between SRD and LRD sequences of discrete values, such as sequences of words.

Other discrete space models have been show to exhibit LRD, such as Markov renewal processes, which are a Markov chain with the time spent in a state occurring randomly according to a distribution. However in this case, the LRD behaviour has been defined by an infinite second moment of the time spent in a state, the dwell time, which then leads to the return time also having an infinite moment. Some possible extensions are to analyse the convergence rate to the stationary distribution and entropy rate of Markov renewal processes and semi-Markov chains. For these processes LRD behaviour can come from both sources, driven by the return time behaviour from the Markov chain and those driven by the dwell time distribution.

4.4 Convergence to the Stationary Distribution of Long Range Dependent Markov Chains

We next examine the actual rate of convergence as a function of $\mathcal{H}$. In this section we introduce some of the main concepts and relevant results regarding the convergence of the limit of n-step transition probabilities to the stationary distribution for Markov chains. This subject has been well studied, in particular conditions where the Markov chain converges at a geometric rate, *i.e.,* decays as ρ^n for some ρ such that $0 < \rho < 1$, are well known. It was also noted in Carpio and Daley [27] that the convergence of the n-step transitions probabilities to the stationary distribution is "slow" for LRD processes. We aim to calculate this convergence for LRD Markov chains by showing the rate of convergence is related to the existence of a corresponding moment of the return time, with the slowest convergence occurring for LRD Markov chain where the second moment of the return time doesn't exist. This

reinforces previous characterisations for LRD on other stochastic processes which show slow convergence is a typical behaviour.

In addition to the concepts above we need a notion of distance between probability mass functions. Specifically, the distance between the n-step transition probabilities and the stationary distribution. Here we use the total variation norm.

Definition 4.4.1. *The total variation distance between two probability distributions, μ and ν on a support, Ω, with an associated sigma algebra, $\mathcal{F}$ is defined as*

$$d(\mu, \nu) = ||\mu - \nu||_{TV} = \sup_{A \in \mathcal{F}} |\mu(A) - \nu(A)|.$$

This definition gives the total variation as the maximum difference between the two probability distributions across all possible events. An extension of this distance is called the f-norm, which is used in the statements of more general convergence theorems where convergence is of the quantity, $f(X_n)$, to its mean value, given an arbitrary function $f : \Omega \to [1, \infty)$ of the states.

Definition 4.4.2. *The f-norm of two probability distributions, μ and ν on a support, Ω is defined as*

$$||\mu - \nu||_f = \sup_{g : |g| \leq f} |\mu(g) - \nu(g)|,$$

where $\mu(g) = \sum_{i \in \Omega} \mu(i) g(i)$, for an arbitrary function g.

Note that the definition here applies to any function g that is dominated by f. The total variation and f-norms are equivalent using the function $f = 1$ [169].

A classic theorem, Theorem 4.4.1 below, classifies the convergence rate of all finite state Markov chains. Characterising the convergence of finite state Markov chains is simpler than the countable state case, as every recurrent Markov chain is positive recurrent and all moments of the return time are finite [88, Theorem 7.3.1]. Given the finite state space, this implies that the convergence rate for finite state Markov chains is geometric.

Theorem 4.4.1 (Theorem 4.9 [118])**.** *For an ergodic Markov chain on a finite state space Ω with stationary distribution π and probability transition matrix P, there exists $\alpha \in (0, 1)$ and $C > 0$ such that $\forall i, j \in \Omega$*

$$\max_{i \in \Omega} ||p_{ij}^{(n)} - \pi_j||_{TV} \leq C \alpha^n.$$

However, since LRD Markov chains have an infinite second moment of the return time random variable, and as Carpio and Daley [27] note, that LRD Markov chains must therefore have an infinite state space. Geometric convergence for Markov chains on countably infinite state spaces requires more conditions, since moments of the return time can be infinite. An important concept, introduced by Kendall [102], is the following.

Definition 4.4.3 (Geometric Ergodicity)**.** *A Markov chain, $\{X_n\}_{n\in\mathbb{Z}^+}$, is geometrically ergodic if there exist numbers c_i, π_i and $0 \leq \rho_i < 1$ for every state $i \in \Omega$ such that*

$$||p_{ii}^{(n)} - \pi_i||_{TV} \leq c_i \rho_i^n.$$

This concept has been prominent in the theory of Markov chains, and has been applied in many contexts, in particular in queueing theory and Monte Carlo Markov chain techniques, where it is important to understand the length of time it takes for a process to converge to its stationary distribution. Geometric ergodicity implies a fast convergence rate, as it can be bounded by an exponentially decaying function.

An extension of Theorem 4.4.1 has been developed for Markov chains on general, not necessarily countable, state spaces. It requires additional definitions to state its conditions. First we will define the sampled chain of a Markov chain.

Definition 4.4.4. *Let $a = \{a_n\}_{n\in\mathbb{Z}^+}$ be a distribution, then we define the sampled chain of a Markov chain $\{X_n\}_{n\in\mathbb{Z}^+}$ for a state i and a subset of the state space, A, to be*

$$K_a(i, A) = \sum_{n=0}^{\infty} p_{iA}^{(n)} a_n,$$

where $p_{iA}^{(n)}$ is the n-step transition probability of moving from state i to a subset A.

Now we define the concept of a petite set.

Definition 4.4.5. *A set C is called petite if the sampled chain satisfies the following bound*

$$K_a(i, A) \geq \nu(A),$$

for all $i \in C$ and for all subsets A, and for a non-trivial measure ν, that is $\nu(A) \neq 0$.

When Ω is countable, every state $i \in \Omega$ forms a singleton petite set and we use these results from general state spaces to refer to petite sets of a single countable state.

The following theorem summarises some important implications and characterisations of geometric ergodicity. The notation $p^{\infty}(C)$ is the limiting probability of being in a subset C.

Theorem 4.4.2 (Geometric Ergodic Theorem [131, Theorem 15.0.1])**.** *For an ergodic Markov chain, $\{X_n\}_{n\in\mathbb{Z}^+}$, on a countable state space, the following conditions are equivalent:*

1. *The chain $\{X_n\}_{n\in\mathbb{Z}^+}$ is positive recurrent with stationary distribution, π, and there exists a petite set $C, 0 < \rho_C < 1$ and $0 < M_C < \infty$ and $p^{\infty}(C) > 0$, such that for all $i \in C$*

$$|p_{i,C}^{(n)} - p^{\infty}(C)| \leq M_C \rho_C^n.$$

2. *There exists some petite set C and $\gamma > 1$ for all $i \in C$ such that*

$$\sup_{i\in C} \mathbb{E}_i[\gamma^{T_{ii}}] < \infty,$$

where $\mathbb{E}_i[\cdot] = \mathbb{E}[\cdot|X_0 = i]$.

Note that we have removed some equivalent conditions that are irrelevant to our discussion.

Part 2 of Theorem 4.4.2 is a condition on the radius of convergence of the probability generating function of the return time random variable, and when considering sets consisting of a single point, it reduces to a condition on the return time distribution. We define the probability generating function of the return time distribution as

$$F_{ii}(z) = \sum_{n=1}^{\infty} \mathbb{P}(T_{ii} = n) z^n = \mathbb{E}[z^{T_{ii}}].$$

for $z \in \mathbb{C}$. By Theorem 4.4.2, if the radius of convergence of $F_{ii}(z)$ is greater than 1, then the chain is geometrically ergodic. For a positive recurrent Markov chain, the radius of convergence is at least 1 since the return time is finite almost surely, and hence $\sum_{n=1}^{\infty} \mathbb{P}(T_{ii} = n) = 1$.

Lemma 4.4.3. *The return time distribution of a Markov chain is heavy-tailed if and only if the convergence to the stationary distribution is slower than geometric.*

Proof. The definition of a heavy-tailed distribution, is equivalent the moment generating function, $E[e^{tT_{ii}}]$ being infinite for all $t > 0$ because a heavy tail from Definition 4.2.2 is equivalent to having infinite moments [67, pg. 11]. This implies that for any heavy-tailed return time distribution, that $F_{ii}(e^t) = \infty, \forall\ t > 0 \iff \mathbb{E}[e^{tT_{ii}}] = \infty,\ \forall\ t > 0$. For the probability generating function, $F_{ii}(z),\ t > 0 \iff z > e^0 = 1$. Which implies that the radius of convergence is exactly 1. By the discussion of the implications of Part 2 of Theorem 4.4.2 above, this implies that the convergence is slower than geometric. □

Hence, LRD Markov chains must converge more slowly than geometric convergence. We now use the knowledge of the moments of the return time to provide a convergence rate for LRD processes and introduce similar results to geometric ergodicity and in this discussion we will use convergence rates of the form $r(n) = (n+1)^{\beta}$.

The direct analogue of the classification of geometric ergodicity for general rate functions is given by the following theorem.

Theorem 4.4.4 (Theorem 2.1 [169])**.** *For an ergodic Markov chain, a function $f : \Omega \to [1, \infty)$ and a rate function, $r(n) : \mathbb{Z}^+ \to \mathbb{R}^+$. The following statements are equivalent:*

1. *There exists a petite set, C, such that*

$$\sup_{i \in C} \mathbb{E}_i \left[\sum_{k=0}^{T_C - 1} r(k) f(X_k) \right] < \infty,$$

   where T_C is the return time to the set C and $\mathbb{E}_i[\cdot] = \mathbb{E}[\cdot | X_0 = i]$.

2. *The sequence, $r(n)||p^n_{i,\cdot} - \pi(\cdot)||_f \to 0$ as $n \to \infty$ for all C such that*

$$\sup_{i \in C} \mathbb{E}_i \left[\sum_{k=0}^{T_B - 1} r(k) f(X_k) \right] < \infty,$$

   for all subsets $B \in \mathcal{F}$ where T_B is the return time to subset B.

Any of these conditions implies

$$r(n)||p^{(n)}_{i,\cdot} - \pi(\cdot)||_f \to 0, \quad \forall i \in \Omega.$$

We have shown in Lemma 4.4.3, that any infinite moment of the return time distribution implies that the Markov chain cannot converge geometrically.

We define a condition that we will use throughout this chapter, concerning the tail decay of the return time random variable.

Condition 4.4.1. *The complementary cumulative distribution function of the return time random variable T_{ii} has a power-law tail. That is, $\mathbb{P}(T_{ii} > n) \sim cn^{-\alpha}$, where $\alpha > 1$ and $c > 0$.*

Next, we show that we can link the maximum convergence rate of the Markov chain to the supremum of finite moments of the return time, under Condition 4.4.1.

To use the first part of Theorem 4.4.4, we require a lemma to that when f is the constant function equal to 1, *i.e.*, $f = 1$ that $\mathbb{E}_i\left[\sum_{k=0}^{T_{ii}-1} r(k)\right] < \infty$ is equivalent to an easier to analyse expression for power law decay. The behaviour of the return time random variable is a property of the communication class, and therefore for an irreducible Markov chain the power-law behaviour of the return time to a state i applies to all states.

Lemma 4.4.5. *For an ergodic Markov chain where $\mathbb{P}(T_{ii} > n) \sim cn^{-\alpha}$,*

$$\mathbb{E}_i\left[\sum_{k=0}^{T_{ii}-1} r(k)\right] < \infty,$$

if and only if

$$\sum_{k=1}^{\infty} r(k)k^{-\alpha} < \infty.$$

Proof. We have

$$\begin{aligned}\mathbb{E}_i\left[\sum_{k=0}^{T_{ii}-1} r(k)\right] &= \sum_{n=1}^{\infty}\left(\sum_{k=0}^{n-1} r(k)\right)\mathbb{P}(T_{ii} = n)\\ &= \sum_{n=1}^{\infty} r(n)\mathbb{P}(T_{ii} > n),\end{aligned}$$

where the second equality follows by Tonelli's theorem for a positive random variable. □

With this, we can state the next result, Lemma 4.4.6, that the rate of convergence, via the exponent of a power-law, is dependent on the existence of a corresponding moment of the return time random variable.

Lemma 4.4.6. *The rate of convergence of the n-step transition probabilities to the stationary distribution of an ergodic Markov chain with Condition 4.4.1, is $O(n^{1-\alpha})$. Specificaly, for any $0 < \beta < \alpha - 1$*

$$(n+1)^{\beta}||p_{i,\cdot}^{(n)} - \pi(\cdot)||_{TV} \to 0.$$

Proof. From Theorem 4.4.4, we can show that the return time random variable in Part 1 converges using the function $f(X_n) = 1, \forall n$, since we are considering the convergence rate from a singleton, that is the state i, which by definition is a petite set. Now considering $f = 1$ and the petite set $C = \{i\}$, the condition in Part 1 becomes,

$$\mathbb{E}_i \left[\sum_{k=0}^{T_{ii}-1} r(k) \right] < \infty.$$

By, Lemma 4.4.5 this condition is equivalent to

$$\sum_{k=1}^{\infty} r(k) k^{-\alpha} < \infty,$$

for an ergodic Markov chain with Condition 4.4.1.

We can apply Theorem 4.4.4, to show the exponents under which convergence occurs. We can see that we require a function $r(n) = (n+1)^{\beta}$ such that, $\beta - \alpha < -1$, since any sum $\sum_{n=0}^{\infty}(n+1)^{\gamma} = \sum_{n=1}^{\infty} n^{\gamma}$, diverges for $\gamma \geq -1$. This implies that we require $\beta < \alpha - 1$ and we require $\beta > 0$ for convergence to occur. So any rate between these two will converge. □

Condition 4.4.1 requires a return time with power law, *e.g.*, $\mathbb{P}(T_{ii} > n) \sim cn^{-\alpha}$. We use

$$\alpha = \sup \left\{ \delta : \mathbb{E}[T_{ii}^{\delta}] < \infty \right\},$$

which we call the moment index. The range of the moment index is $\alpha \geq 0$, however we are only considering $\alpha > 1$ in this discussion since we consider positive recurrent Markov chains. From the previous discussion and Lemma 4.4.3, we can conclude that if all moments of the return time random variable exist, then the convergence to the stationary distribution is geometric. From Lemma 4.4.6, if there exist any infinite moments of the return time random variable and a power-law tail in the return time random variable, then the convergence is a power-law with exponent of the moment index minus 1, *i.e.*, $\alpha - 1$.

This gives an interesting link between the convergence rate of an LRD Markov chain and the Hurst parameter. This is summarised in the following theorem.

Theorem 4.4.7. *The rate of convergence of the n-step transition probabilities to the stationary distribution of a LRD Markov chain power-law decaying complementary cumulative distribution function for the return time,* $\mathbb{P}(T_{ii} > n) \sim n^{-\alpha}$, $1 < \alpha < 2$, *is* $O(n^{2-2\mathcal{H}})$. *That is, for any* $0 < \beta < \mathcal{H}$

$$(n+1)^{2-2\beta} ||p_{ii}^{(n)} - \pi_i||_{TV} \to 0, \forall i.$$

Proof. From Carpio and Daley [27] and Theorem 1 of Daley [45] we have that the Hurst parameter is linked to the moment index by the following relationship,

$$\mathcal{H} = \frac{1}{2}(3 - \alpha). \tag{4.3}$$

Since, the exponent α in the complementary cumulative distribution function represents the moment index in the distribution [67, pg. 32], and by rearranging (4.3) that $\alpha - 1 = 2 - 2\mathcal{H}$. Then the result follows by applying Lemma 4.4.6. □

This result echoes previous work in the area of LRD, *e.g.,* [13], which shows that the convergence rate to important quantities is slower for LRD processes and is related to the Hurst parameter. Interestingly, this result illustrates that the convergence rate slows as the Hurst parameter tends to one, *i.e.,* as the degree of LRD increases. As $\mathcal{H} \to 1$ the exponent tends to 0, and the moment index is close to 1 and at that stage the expectation of the return time is finite. If the moment index is ≤ 1, then the expectation becomes infinite, and the Markov chain is null-recurrent, *i.e.*, $\mathbb{E}[T_{ii}] = \infty$ and $\alpha = 1$.

Finally, we will discuss the behaviour of the excess entropy under the additional assumption that the return time random variable has a power-law tail, *i.e.,* $\mathbb{P}(T_{ii} > n) \sim cn^{-\alpha}$, where $\alpha > 1$ and $c > 0$. Using the same argument as Lemma 4.4.6, the rate of convergence to the stationary distribution is $O(n^{1-\alpha})$. Which by Theorem 4.3.3 implies that the convergence rate of the conditional entropy to the entropy rate is also $O(n^{1-\alpha})$. This gives

$$\begin{aligned} E(n) &= \sum_{r=1}^{n} H[X_r|X_{r-1}, \ldots, X_1] - H[\mathcal{X}], \\ &\sim \sum_{r=1}^{\infty} n^{1-\alpha}. \end{aligned}$$

Which implies that in the case of power-law tail of the return time random variable that $E < \infty \iff \alpha > 2$. That is, the excess entropy is finite if and only if the second moment of the return time random variable is finite. This leads to the following result.

Corollary 4.4.7.1. *For Markov chains with the complementary cumulative distribution function of the return time has a power law tail,* i.e., $\mathbb{P}(T_{ii} > n) \sim cn^{-\alpha}$, *where* $\alpha > 1$ *and* $c > 0$, *the Markov chain is LRD if and only if* E *is infinite.*

Chapter 5

A Survey of Entropy Rate Estimation

The estimation of the entropy of a random variable has long been an area of interest in Information Theory. From the original definition by Shannon [154], the interest in development of information theory and entropy as a concept was motivated by aiming to understand the uncertainty of sources of information and in the development of communications theory. In real systems, understanding this uncertainty allows more robust models and a better understanding of complex phenomena.

Estimation of the entropy of random variables has been reviewed on several occasions, with reviews that have covered the estimation of Shannon, differential, and other types of entropy measures. A recent survey by Verdu [175], reviews techniques for empirical estimation of many information measures, such as entropy, relative entropy and mutual information, for both discrete and continuous data. Amigó *et. al.* [5] surveyed generalised entropies, for further quantification of complexity and uncertainty of random variables. Rodriguez *et al.* [40] survey and review the performance of 18 entropy estimators for short samples of data, assessing them on their bias and mean squared error. A comparison of different generalised entropy measures, and their performance, was recently performed by Al-Babtain *et. al.* [1].

In this chapter, we review techniques for estimating the entropy *rate*, a measure of uncertainty for stochastic processes. This is a measure of the average uncertainty of a stochastic process, when measured per sample. Shannon's initial work considered the problem of quantifying the uncertainty of Markov sources of information [154]. We will be considering estimation techniques of the entropy rate for both discrete and continuous data, therefore covering both Shannon and differential entropy rate estimation. There are a variety of estimation properties and ways of assessing the quality of estima-

	Modelling Estimate	
Entropy Rate Estimate	Parametric	Nonparametric
Parametric	[11, 24, 25, 26, 29, 35, 36, 37, 71, 83, 100, 125, 132, 135, 136, 141, 144, 160, 166, 184]	[9, 10, 28, 33, 60, 70, 81, 104, 117, 122, 139, 140, 159, 163, 164, 165, 168]
Nonparametric	N/A	[7, 47, 79, 98, 99, 106, 142, 143, 147, 162, 172]

Table 5.1: Comparison of entropy rate estimation techniques into categories based on parametric/nonparametric techniques. The modelling estimate refers to the quantity that is estimated in the technique and the entropy rate estimate refers to the full entropy rate expression used. For example, if estimating entropy rate of a Markov chain using plug-in estimation. Then the modelling estimates may be nonparametric for the transition probabilities, p_{ij} and the stationary distribution, π_j. However, the entropy rate estimator is a parametric estimator for the Markov model. Hence, there are no nonparametric/parametric estimators because nonparametric entropy estimators do not use a model.

tors that will be used in this section and throughout the thesis; a discussion is given in Appendix E.

There are two main estimation paradigms that are used in statistical estimation, parametric and nonparametric estimation. Parametric techniques assume a model for the stochastic process that generates the data, and fit parameters to the model [42]. In many cases, these parameters are estimated and then used directly in an entropy rate expression, which we call plug-in estimation. Nonparametric approaches, on the other hand, make very few assumptions on the process that generates the data. However they can contain assumptions about properties, such as stationarity [42]. Fewer assumptions for an estimator can lead to more robustness. This review will cover techniques using both of these approaches, outlining what assumptions are used in the generation of the estimates. The material has been published in the paper "A Review of Shannon and Differential Entropy Rate Estimation" [65].

The parametric estimation techniques reviewed here model the data as Gaussian processes, Markov processes, hidden Markov models and renewal processes. For Gaussian processes, due to the equivalence of entropy rate estimation and spectral density estimation, which we discuss below, we in-

	Time	
State Space	Discrete	Continuous
Discrete	[7, 11, 29, 35, 36, 37, 71, 74, 75, 79, 83, 98, 99, 100, 106, 125, 132, 135, 136, 141, 143, 144, 160, 166, 172, 184]	[71]
Continuous	[7, 9, 10, 24, 25, 28, 33, 47, 60, 70, 81, 104, 117, 122, 139, 140, 142, 147, 159, 163, 164, 165, 168]	N/A

Table 5.2: Comparison of entropy rate estimation techniques. They are partitioned into 4 categories based whether they are discrete or continuous time, and whether the work on discrete or continuous valued data.

troduce some literature on spectral density estimation, such as maximum entropy and maximum likelihood techniques.

Nonparametric estimators are often based on limit theorems of an expression of the entropy rate, with estimation being made on a finite set of data. We review and present assumptions and properties of nonparametric entropy rate estimation techniques for Shannon entropy rate, which are based on limit theorems of string matches. For differential entropy rate estimation, we present 3 techniques that were developed as measures of complexity of time series, rather than strictly as entropy rate estimators. In some special cases, such as first order Markov chains, these estimators have been shown to converge to the entropy rate and therefore, in practice, have been used as entropy rate estimators. Then we present another approach using conditional entropy estimates, based on observations of a finite past, that provides an exact estimate, given some assumptions. There are far fewer techniques that have been developed for continuous-valued random variables, which is not surprising given the history of development of information theory for transmission of data.

5.1 Parametric approaches

In this section we will discuss parametric approaches to estimate the entropy rate of a process from observed data. Parametric estimators assume a model for the data, estimate some aspects of the model from the data and then directly calculate the entropy rate from those estimates. The three model types

used are Gaussian processes, Markov processes, and renewal/point processes.

5.1.1 Gaussian Processes

First we will cover a class of processes that are defined by the assumption that the finite dimensional distributions are normally distributed. Since the spectral density is the Fourier transform of the autocovariance, all the information for the process is encoded in the spectral density.

The entropy rate of a Gaussian process is given by,

$$h(\chi) = \frac{1}{2}\log(2\pi e) + \frac{1}{4\pi}\int_{-\pi}^{\pi} \log(f(\lambda))d\lambda, \tag{5.1}$$

where $f(\lambda)$ is the spectral density of the process [41, pg. 417].

This reduces the estimation task down to estimating the spectral density of the process. That is, using an approach to create an estimate of the spectral density function, $\hat{f}(\lambda)$, and plugging it in to the expression above. There are several methods to estimate the spectral density of a Gaussian process, and hence produce a estimator of the entropy rate. Note that we can use this framework even in the cases of discrete-valued, discrete-time processes, using sampling techniques which can be used to calculate the integral in (5.1). A variety of parametric and nonparametric techniques have been developed to estimate the spectral density of a Gaussian process. We will refer to these as either parametric/nonparametric or parametric/parametric for the classification by their entropy estimate and modelling estimate type.

5.1.1.1 Maximum Entropy Spectral Estimation

A common technique used for the inference of spectral density is maximum entropy spectral estimation. This is a fitting paradigm that selects the estimate that maximises the entropy, that is, has the highest uncertainty, given the current knowledge.

These techniques were introduced by Burg [24, 25], when aiming to model seismic signals by fitting stochastic models. He showed that given a finite set of covariance constraints for a process, $\mathbb{E}[X_i X_{i+k}] = \alpha_k,\ k = 0, 1, \ldots, p$, then the process that is the best fit for the constraints, given a maximum entropy approach, is the class of autoregressive processes, AR(p),

$$X_n = -\sum_{k=1}^{p} a_k X_{n-k} + \epsilon_n,$$

where $\epsilon_n \sim \mathcal{N}(0, \sigma^2)$ is normally distributed and a_k and σ^2 are selected to fit the constraints [32].

This type of analysis can be generalised to auto-regressive moving-average, ARMA(p,q), models of the form

$$X_n = -\sum_{k=1}^{p} a_k X_{n-k} + \sum_{k=1}^{q} b_k \epsilon_{n-k},$$

where the additional parameters, b_k, are selected to fit the behaviour of the noise process. Maximum entropy spectral analysis in this case also has to consider the function of the noise, called the impulse response function. It was shown by Franke [68, 69] that ARMA is the maximum entropy process given a finite set of constraints on the covariances and constraints on the impulse response function, $\mathbb{E}[X_i \epsilon_{i-k}] = \sigma_\epsilon^2 h_k, k = 1, \ldots, q$, where σ_ϵ^2 is the variance of the noise variables and h_k are the parameters of the impulse responses.

The entropy rate of the AR(p) and ARMA(p,q) classes of processes does not need to perform the integration over the spectral density function, given in 5.1, because the rate is known to be [64]

$$h(\chi) = \frac{1}{2} \log(2\pi e \sigma_\epsilon^2).$$

That is, the new information at each step of the process arises purely from the innovations, and if we can estimate the variance of the innovations then we can infer the entropy rate directly. Note that σ_ϵ^2 is the variance of the innovation process, not the variance of the AR/ARMA process itself. This has been extended to the ARFIMA(p,d,q) class of processes, where a process passed through a linear filter $(1-L)^d, -\frac{1}{2} < d < \frac{1}{2}$, of the lag parameter L, *i.e.*, $LX_n = X_{n-1}$ is an ARMA(p,q) process, with the same entropy rate [64]. However, for a fixed process variance, the entropy rate in this case is dependent upon the fractional parameter, d.

5.1.1.2 Maximum Likelihood Spectral Estimation

In contrast to maximum entropy techniques, there are a class of techniques using a likelihood-based approach. These select model parameters based on likelihood function, which is the probability of parameters that would have generated the observations. In contrast to maximum entropy techniques, maximum likelihood requires a model of the data, from which the likelihood function is calculated.

These were first developed by Capon [26], to estimate the power spectrum from an array of sensors. Each sensor's signal is modelled as, $x_i = s + n_i$, where x_i is the observed value at a sensor i, s is the signal and n_i is the noise

at sensor i. The maximum likelihood assumption is used in the density of the noise, a multivariate normal distribution, and then a maximum likelihood estimate is made for the underlying signal.

Connections between the maximum entropy and maximum likelihood paradigms have been found in some aspects of spectral estimation. Landau [113] makes a connection between the maximum likelihood estimate of a spectral measure based on a one parameter distribution and the maximum entropy spectral measure, where the maximum entropy measure is the uniform average over all of the maximum likelihood spectral measures. In the one-parameter case, maximum entropy is the uniform average over the parameters of the maximum likelihood estimators.

These approaches can then be used for an entropy rate estimate, calculating (5.1) above by plugging in the inferred spectral density function.

5.1.1.3 Nonparametric Spectral Density Estimation

The spectral density of a Gaussian process can be estimated directly without additional modelling assumptions, and then used in (5.1) to estimate the entropy rate.

A common technique to estimate the spectral density is called the periodogram, which uses the fact that the spectral density is the Fourier transform of the autocovariance function. Therefore, we can calculate the plug in estimate of the spectral density estimate as

$$\hat{f}(\lambda) = \sum_{j=-\infty}^{\infty} \hat{\gamma(j)} e^{ij\lambda} d\lambda,$$

where the autocovariance function can be estimated from observed data as

$$\hat{\gamma}(k) = \frac{1}{n-k} \sum_{i=1}^{n-k} X_i X_{i+k}.$$

However, this can cause issues as it may not converge for large sample sizes. This motivated the research of the maximum entropy processes, given different autocorrelation constraints [25].

Some important work in the development of the periodogram on time series data is from Bartlett [9, 10] and Parzen [139, 140] showing the consistency of the periodogram. Smoothing techniques have been developed and expanded in work by Tukey [168] and Grenander [81].

Other techniques for nonparametric spectral density have been developed. Some examples include Stoica and Sundin [159] by considering the estimation

as an approximation to maximum likelihood estimation. Other nonparametric techniques are robust to data from long memory processes, which have a pole at the origin of the spectral density, by Kim [104]. Finally numerous Bayesian techniques have been developed for smoothing [117], parametric inference of the periodogram [28, 70], robust to long memory [122], using MCMC to sample a posterior distribution [33, 60] and using Gaussian process priors [163, 164, 165].

5.1.2 Markov Processes

Markov processes have been used to model information sources since Shannon's introduction of information theory [154]. In this section, we discuss entropy rate estimation assuming the Markov property, that is for a process $\{X_i\}_{i\in\mathbb{N}}$,

$$\mathbb{P}(X_n = x_n | X_{n-1} = x_{n-1}, \ldots, X_1 = x_1) = \mathbb{P}(X_n = x_n | X_{n-1} = x_{n-1}).$$

There are two main types of Markov processes considered, firstly a simple Markov chain, and secondly hidden Markov models (HMM). We mention Markov jump processes at the end, which have had substantially less attention.

5.1.2.1 Markov Chains

The entropy rate of a stationary Markov chain with state space, Ω, is given by

$$H(\chi) = \sum_{i\in\Omega}\sum_{j\in\Omega} \pi_i p_{ij} \log p_{ij}, \tag{5.2}$$

where the $p_{ij} = \mathbb{P}(X_n = j | X_{n-1} = i)$ form the probability transition matrix and π_i is the stationary distribution for the Markov chain [41, Theorem 4.2.4]. For this approach, an implicit assumption of an ergodic Markov chain is required, for the existence of the stationary distribution.

A few different approaches have been developed to estimate transition probabilities and the stationary distribution, which utilise parametric or nonparametric estimators.

The approach that has received most attention is to estimate the stationary distribution, and the probability transition matrix directly, which was inspired by the description of plug-in estimators for single samples by Basharin [11]. Maximum likelihood estimation techniques have been developed by Ciuperca and Girardin [35], on a finite state space, Girardin and Sesboue [74, 75] on two-state chains, and Ciuperca and Girardin [36] on

countable state spaces. These utilise maximum likelihood estimators for π_i and p_{ij}, given observations of the chain $X = (X_0, \ldots, X_n)$,

$$\hat{p}_{ij} = \frac{N_{ij}[0,n)}{N_i[0,n)}, \quad \text{and,} \quad \hat{\pi}_i = \frac{N_i[0,n)}{n},$$

where

$$N_{ij}[0,n) = \sum_{m=1}^{n} \mathbb{1}_{\{X_m=j, X_{m-1}=i\}}, \quad \text{and,} \quad N_i[0,n) = \sum_{j\in\Omega} N_{ij}[0,n),$$

are the counting functions of transitions from state i to j and visits to state i respectively.

Whether estimating from one long sample or many groups of samples, the estimator from plugging these values into the entropy rate expression (5.2) are strongly consistent and asymptotically normal [35, 36, 75].

For the countable case, for any finite sample there will be transitions that have not been observed which are then set to 0, *i.e.*, $p_{ij} = 0$ if $N_{ij}[0,n) = 0$, however in the limit as $n \to 0$ the entropy rate still converges. These results have been extended to more general measures using extensions of the entropy rate, such as Renyi Entropy [37].

Kamath and Verdu [100] have analysed the convergence rates for finite samples and single paths of estimators of this type. They showed that convergence of the entropy rate estimators can be bounded using the convergence rate of the Markov chain and the number of data observed.

A similar technique on finite state Markov chains was introduced by Han *et al.* [83], by enforcing a reversibility condition on the Markov chain transitions, in particular $\pi_i p_{ij} = \pi_j p_{ji}$. Using the stationarity of the transition function of the Markov chain they define an estimator by utilising Shannon entropy estimators, of the conditional entropy $H(X_2|X_1 = i)$, and then the overall estimator is

$$\hat{H} = \sum_{i\in E} \hat{\pi}_i \hat{H}(X_2|X_1 = i),$$

where $\hat{\pi}_i$ is the stationary distribution estimate.

An estimator was proposed by Chang [29] for finite Markov chains with knowledge of the probability transition matrices, and calculates the convergence rate to the entropy rate estimate,

$$\hat{H}_N = \frac{\sum_{n=0}^{N-1} H(X_n)}{N},$$

given an initial state, $X_0 = x$ and where $H(X_n)$ is the Shannon entropy given knowledge of the current state, $X_n \in \Omega$. This is the same as using the maximum likelihood estimator of π_i, considering the probabilities as parameters, and then having a known conditional entropy estimate, as in the previous approach by Han *et al.* [83]. Chang was able to show that there is an exponential rate of convergence of this technique to the real value [29]. A similar result is obtained by Yari and Nikooravesh [184], showing an exponential convergence rate for this type of estimator under an assumption of ergodicity.

A final approach by Strelioff *et al.* [160] utilises Bayesian techniques to calculate the entropy rate of a Markov chain, using the connection to statistical mechanics. The model parameters, the probability transitions of the kth order Markov chain, are inferred as a posterior using a prior distribution, incorporating observed evidence. This is formulated as

$$\mathbb{P}(\theta_k|M_k)\mathbb{P}(D|\theta_k, M_k) = \mathbb{P}(D, \theta_k|M_k),$$

where D is the data, M_k is a kth order Markov chain and θ_k are the parameters, transition probabilities, of the Markov chain. The same framework can be applied to other information theoretic measures, the Kullback-Leibler divergence.

5.1.2.2 Hidden Markov Models

A generalisation of Markov chains is given by hidden Markov Models, where we observe a sequence, $\{Y_i\}_{i\in\mathbb{Z}^+}$ where there is a hidden underlying Markov chain, $\{X_i\}_{i\in\mathbb{Z}^+}$, and the probabilities of the observations of the hidden Markov model only depend on the current state of the Markov chain,

$$\mathbb{P}(Y_n = y_n|Y_{n-1}, \ldots, Y_1, X_n, \ldots, X_1) = \mathbb{P}(Y_n = y_n|X_n).$$

Hence, this also exhibits the Markov property with dependence on the latent Markov chain.

In general there is no known expression to directly calculate the entropy rate of a hidden Markov model [61, 62, 95, 145], so we can't just describe the techniques with respect to a plug-in expression for this class of models. However, some upper and lower bounds have been given by Cover and Thomas [41, pg. 69], and a proof of convergence of the bounds to the true value. It was shown that the entropy rate function is analytic in its parameters in [82], and it has been shown that the entropy rate function of a hidden Markov model varies analytically in its parameters, with some assumptions on the positivity of the transition matrix of the embedded Markov chain.

In the more specific case of binary valued models, where both the Markov chain $\{X_i\}_{i\in\mathbb{Z}^+}$ and observed random variables $\{Y_i\}_{i\in\mathbb{Z}^+}$ are binary valued, there have been expressions derived based on a noise model using a series expansion and analysing the asymptotics [183, 187, 188], and some analysis which links the entropy rate to the Lyapunov exponents, arising in dynamical systems [95]. Nair et al. [132] generated some upper and lower bounds, depending on the stationary distribution of the Markov chain and the entropy of a Bernoulli random variable. Lower bounds were further refined by Ordentlich [136], by creating an inequality that utilises a related geometrically distributed random variable. The exact expression remains elusive and is an active topic of research, however as pointed out by Jacquet *et al.* [95], the link with Lyapunov exponents highlights the difficulty of this problem in general.

Although there are no explicit estimators for HMMs, Ordentlich and Weissman [135] created an estimator for the binary sequence $\{Y_i\}_{i\in\mathbb{Z}^+}$,

$$H(Y) = E\left[H_2\left(\frac{e^{Y_i}}{1+e^{Y_i}} \star p \star \delta\right)\right],$$

where

$$H_2(p) = -p\log_2 p - (1-p)\log_2(1-p),$$

is the binary entropy function, $\star$ is the binary convolution operator, p and δ are the probability of the embedded Markov chain changing state and the probability of observing a different state from the Markov chain. Given these simplifications, we can get an expression in terms of the expectation of the random variable and the stationary distribution. Luo and Guo [125] utilised a fixed point expression that can be developed on the cumulative distribution function. Then a conditional entropy expression is exploited to calculate an entropy rate estimate,

$$H(X_1|X_0, Y_{-\infty}^1) = E\left[H_2\left(\left(1+e^{-\alpha X_0 - r(Y_1) - L_2}\right)^{-1}\right)\right],$$

and

$$\alpha = \log\left((1-\epsilon)/\epsilon\right),$$
$$r(y) = \log\frac{\mathbb{P}_{Y|X}(y+1)}{\mathbb{P}_{Y|X}(y-1)},$$

where H_2 is the binary entropy function, $\mathbb{P}_{Y|X}()$ is the conditional probability of random variable Y given X and L_2 is the log-likelihood ratio. Then they

computed this numerically to form estimates, using a technique that exploits the fixed-point structure in a set of functional equations.

Gao *et. al.* [71] use a nonparametric approach using limit theorems discussed in Section 5.2.1, which is applied to other processes such as Markov chains. However, with some assumptions, results can be achieved using limit theorems and fitting parameters to data. Travers [166] uses a path-mergability condition, if there exist paths that emit a symbol from the process $\{Y_i\}_{i\in\mathbb{Z}^+}$,

$$\delta_i(w) = \left\{ j \in E : \mathbb{P}_i \left(X_0^{|w|-1} = w, Y_{|w|} = j \right) \right\},$$

such that for two distinct states i and j, there is a state k that can be reached from both states while creating the same path, *i.e.*,

$$k \in \delta_i(w) \cap \delta_j(w).$$

Then entropy rate estimates are made nonparametrically of the conditional entropy,

$$H_T(\chi) = H(X_T | X_{T-1}, \ldots, X_1),$$

which under the the stationarity assumption this converges to the entropy rate. Given these assumptions, the estimates converge to the true value in the total variation norm at an exponential rate.

Peres and Quas [141] then tackle the problem of finite state hidden Markov models with rare transitions. The analysis is performed by setting some rare transitions to 0. In this case, they have defined the entropy rate as the average over the possible paths w,

$$H(Y) = \lim_{N\to\infty} \frac{1}{N} \sum_{w\in\Omega^N} \mathbb{P}(Y_1^N = w) \log \mathbb{P}(Y_1^N = w).$$

Under these assumptions, some lower and upper bounds of the expression above were found. These bounds are composed of the sums of the entropy rate of the Markov chain alone, and the entropy of the conditional distribution of the observed variables given the latent Markov chain.

5.1.2.3 Other Markov Processes

In addition Markov and hidden Markov chains, some less studied Markov process have had parametric entropy rate estimators developed.

Dumitrescu [56] analysed Markov pure-jump processes, which are processes that have an embedded discrete-time Markov chain with jumps occurring at random times, T_t, for the t-th jump, where the rates are given by a

generator matrix, $Q = (q_{ij})_{i,j\in\Omega}$. In this case, Dumitrescu [56] proved that the entropy rate is

$$H(\chi) = \sum_{i\in\Omega} \pi_i \sum_{j\neq i} q_{ij} \log q_{ij} + \sum_{i\in\Omega} \pi_i \sum_{j\neq i} q_{ij}, \tag{5.3}$$

for π, the stationary distribution of the Markov chain.

Regnault [144] showed that, similar to the results of Ciuperca and Girardin [35, 36], that the stationary distribution could be estimated consistently and is asymptotically normal, for both: one long sample paths and an aggregation of multiple sample paths. Consistency and asymptotic normality of the generator matrix, $\hat{Q}$, also proved, which are estimated using

$$\hat{q}_{ij} = \begin{cases} \frac{N_{ij}[0,n)}{R_i[0,n)}, & \text{if } R_i[0,n) \neq 0, \\ 0, & \text{otherwise,} \end{cases}$$

where $R_i[0,n)$ is the total time spent in state i. Regnault then proved that plugging these estimates into the parametric form of the entropy rate in (5.3) results in consistent and asymptotically normal estimates of the entropy rate, for the case of estimation from one long single path and estimation of multiple paths.

5.1.3 Renewal/Point Processes

Another important class of stochastic processes are renewal processes. These processes are a sequence of independent realisations of an inter-event distribution. We define the renewal process $S = \{S_i\}_{i\in\mathbb{N}}$, where S_i is the time of the ith event, and the inter-event times $X = \{X_i\}_{i\in\mathbb{N}}$, and note that $S_i = \sum_{j=0}^{i} X_j$. In a renewal process the X_i are all independent. A key description of a renewal process is the counting function of events, which is defined similarly to the Markov chain case above, $N[0,n) = \sum_{j=0}^{\infty} \mathbb{1}_{\{S_j \leq n\}}$, where each jump increments $N[0,n)$ by 1. The entropy rate in the case of discrete-time inter-event distribution, p_i, is

$$\begin{aligned} H(S) &= \lambda H(X), \\ &= -\lambda \sum_{j=1}^{\infty} p_j \log p_j, \end{aligned}$$

where $\lambda = 1/\mathbb{E}[X_1]$.

Gao *et al.* [71] defined a technique for estimating the entropy rate of discrete-time renewal processes, for a discrete distribution of inter-event

times, $p_j, j = 1, 2, \ldots$, to model binary valued time series. The estimator is simply,

$$H(S) = -\hat{\lambda} \sum_{j=1}^{\infty} \hat{p_j} \log \hat{p_j}.$$

This was shown to be a consistent estimator of entropy rate, however in practice it was shown that long strings can be undersampled unless the process was observed for an extremely long time. This is another example of a nonparametric model inside of a parametric estimator.

Alternatively, you could estimate p_j parametrically, *e.g.*, assume X is a geometric random variable and then $p_j = p(1-p)^j$ and then estimate the probability, the parameter p and plug this into the entropy rate estimator.

5.2 Nonparametric Approaches

In this section we will be discussing nonparametric estimators of the Shannon and differential entropy rate. In contrast to the previous section, the estimators presented here make very few assumptions about the form of the data generating process. However, there are still assumptions that are required to enable the analysis, in particular the stationarity or ergodicity of the process, to allow for limit theorems which are used to develop estimators with the desired properties. Nonparametric methods are robust to the type of distribution and parameter choices of models [73, pg. 3]. There has been more research interest for Shannon entropy rate estimation, rather than differential entropy rate. However, there has been considerable research into the estimation of differential entropy, see Beirlant *et. al.* [12]. The interest into differential entropy estimation techniques continues, particularly with the increase in computational power to enable efficient calculation of kernel-density based techniques [20].

5.2.1 Discrete-Valued, Discrete-Time Entropy Rate Estimation

In this section we will briefly describe some entropy rate estimators for discrete-valued, discrete-time processes. We will consider techniques that utilise completely nonparametric inference of quantities that can be used for entropy rate inference. Nonparametric estimators have a rich history in information theory as ways of characterising the complexity and uncertainty of sources of generating data, particularly when considering communication theory and dynamical systems.

The first estimator we discuss is based on the Lempel-Ziv compression algorithm [186]. The estimation technique is based on a limit theorem on the frequency of string matches of a given length, for each $n \in \mathbb{Z}^+$. Given the length of the prefix sequences of a process starting at digit i, $x_i, x_{i+1}, \ldots$, we define,

$$L_i^n(x) = \min\{L : x_i^{i+L-1} \neq x_j^{j+L-1}, 1 \leq j \leq n, j \neq i\},$$

where $x_i^{i+n} = x_i x_{i+1} \ldots x_{i+n}$. This is the length of the shortest prefix of $x_i, x_{i+1}, \ldots$ which is not a prefix of any other $x_j, x_{j+1}, ...$ for $j \leq n$. A limit theorem was developed by Wyner and Ziv [181], based on the string matching, which states,

$$\lim_{n\to\infty} \frac{L_i^n(x)}{\log n} \to \frac{1}{H(\chi)}, \text{ in probability.}$$

This was extended to almost sure convergence, by Ornstein and Weiss [137]. Utilising the idea of this theorem, estimation techniques were developed which utilise multiple substrings and average the L_i^n's instead of estimating from one long string, to make accurate and robust estimates with faster convergence to the true value. The following statement, by Grassberger [79], was suggested heuristically,

$$\lim_{n\to\infty} \frac{\sum_{i=1}^n L_i^n(x)}{n \log n} = \frac{1}{H(\chi)}, \text{ a.s.}$$

This expression was shown, by Shields [156], to not hold except in the cases of simple dependency structures, such as i.i.d. processes and Markov chains. However, a weaker version does hold for general ergodic processes, which states that for a given $\epsilon > 0$, all but a fraction of at most ϵ of the $\sum_{i=1}^n L_i^n(x)/n \log n$, are within the same ϵ of $1/H(\chi)$ [156].

This is converted to an estimation technique by taking a suitably large truncation, and calculating the above expression for $1/H(\chi)$. However, to make consistent estimates for more complex dependency structures, where the limit expression above does not hold, additional conditions are required. Kontoyiannis and Suhov [107], and Quas [143] extended this concept to a wider range of processes, firstly to stationary ergodic processes that obey a Doeblin condition, *i.e.*, there exists an integer $r \geq 1$ and a real number $\beta \in (0, 1)$ such that for all $x_0 \in A, \mathbb{P}(X_0 = x_0 | X_{-\infty}^{-r}) \leq \beta$, with probability one, and secondly to processes with infinite alphabets and to random fields satisfying the Doeblin condition.

Kontoyiannis and Suhov [107] followed the results of Shields [156] and Ornstein and Weiss [137], to show that the above estimator is consistent in

much further generality with the addition of the Doeblin condition. They also state that without the condition, $1/h(\chi)$ is the asymptotic lower bound of the expression.

Another class of estimators, which was initially suggested by Dobrushin [55], uses a distance metric on the "closeness" of different strings. We let ρ be a metric on the sample space Ω, and define sequences of length T, as $x_i^{i+T} = (x_i, x_{i+1}, ..., x_{i+T})$, with each of the n sequences being independent. A nearest neighbour estimator is defined as,

$$\hat{h_n} = -\frac{1}{n \log n} \sum_{j=1}^{n} \log \left(\min_{i:i \neq j} \rho(x_i^{i+T}, x_j^{j+T}) \right).$$

Grassberger suggested this as an estimator with the metric $\rho(x, y) = \max\{2^{-k} : x_k \neq y_k\}$ [79]. This is an equivalent formulation using the L_i^n quantity, and therefore the same results from Shields apply. Similar techniques for nearest neighbour estimation were developed by Kaltchenko *et al.* [99], and the convergence rate for the nearest neighbour estimator was shown by Kaltchenko and Timofeeva [98]. Another related estimator was developed by Vatutin and Mikhailov [172], where they calculated the bias and consistency for nearest neighbour estimation.

A generalisation of the nearest neighbour entropy estimator was introduced as a measure called Statentropy by Timofeev [162]. This estimator is defined as,

$$\hat{h_n} = -\frac{1}{n \log n} \sum_{j=1}^{n} \log \left(\overset{(k)}{\min_{i:i \neq j}} \rho(x_i^{i+T}, x_j^{j+T}) \right),$$

where $\min^{(k)}$ is the kth order statistic, *i.e.*, the kth smallest value of the pairwise comparisons. Hence, this is a generalisation of the nearest neighbour estimator, by considering the kth smallest value, rather than the minimum. This estimator has been shown to be consistent, with convergence rates to the entropy rate developed by Kaltchenko *et al.* [98].

5.2.2 Continuous-Valued, Discrete-Time Entropy Rate Estimation

We consider some non-parametric estimators of entropy rate for continuous-valued data in two different classes, entropy measures that are adapted from Shannon entropy, that we can use for comparison of complexity of a system,

and absolute measures, which are intended to accurately estimate the value of differential entropy rate for a system.

The measures adapted from Shannon entropy include two closely related approaches, approximate [142] and sample entropy [147], which utilise pairwise comparisons between substrings of realisations of the process to calculate a distance metric. Another popular approach is permutation entropy which utilises the frequency of different permutations of order statistics of the process [8], and then calculates the estimate using an analogue of Shannon entropy on the observed relative frequencies [7].

These techniques were developed to quantify the complexity of continuous-valued time series, and therefore the intention is to compare time series as opposed to provide an absolute estimate. These types of measures, from dynamic systems literature, have been successful in the analysis of signals to detect change [2, 31, 111]. From the probabilistic perspective we have an interest in the accurate, nonparametric estimation of differential entropy rate from data, without any assumptions on the distribution of the underlying source, and to compare complexity using this quantity.

The final technique we consider, specific entropy [47], is an absolute measure of the entropy rate. Due to computational advances, the technique uses nonparametric kernel density estimation of the conditional probability density function, based on a finite past, and uses this as the basis of a plug-in estimator.

We present each of these techniques in more detail below.

5.2.2.1 Approximate Entropy

Approximate entropy was introduced by Pincus [142], with the intention of classifying complex systems. However it has been used to make entropy rate estimates, since it was shown in the original paper to converge to the true value in the cases of i.i.d. processes and first-order finite Markov chains. Given a sequence of data, $x_1, x_2, ..., x_N$, we have parameters m and r, which represent the length of the substrings we use for comparison and the maximum distance, according to a distance metric, between substrings to be considered a match. Then we create a sequence of substrings, $u_1 = [x_1, ..., x_m], u_2 = [x_2, ..., x_{m+1}], ..., u_{N-m+1} = [x_{N-m+1}, ..., x_N]$ and we define a quantity,

$$C_i^m(r) = \frac{1}{N-m+1} \sum_{j=1}^{N-m+1} \mathbb{1}_{\{d[u_i,u_j] \leq r\}},$$

where $d[x(i), x(j)]$ is a distance metric. Commonly used metrics for this measure are the l_∞ and l_2 distances.

The following quantity, used in the calculation of the approximate entropy, is defined in Eckmann and Ruelle [59] and used in Pincus [142],

$$\Phi^m(r) = \frac{1}{N-m+1} \sum_{i=1}^{N-m+1} \log C_i^m(r).$$

We now define the approximate entropy, $ApEn(m,r)$, which is,

$$ApEn(m,r) = \lim_{N\to\infty} \left[\Phi^m(r) - \Phi^{m+1}(r)\right].$$

For finite sequences of length N this is positively biased because of the logarithm function $\mathbb{E}[\log(X)] \leq \log(\mathbb{E}[X])$ by Jensen's inequality [52] and the counting of some substrings twice. The bias in this estimator decreases as the number of samples, N, gets larger [52].

Pincus showed in his initial paper, that approximate entropy would converge to the entropy rate for i.i.d and finite Markov chains [142]. However, this doesn't hold in more general cases. It has been noted that the approximate entropy corresponds to the differential Renyi entropy rate of order 1 [110].

Approximate entropy is also quite sensitive to the two parameters, m, and r, and hence care must be taken when selecting these parameters [52, 185]. It is recommended that m has a relatively low value, *e.g.*, 2 or 3, which will ensure that the conditional probabilities can be estimated reasonably well [52]. The recommended values for r, are in the range of $0.1\sigma - 0.25\sigma$, where σ is the standard deviation of the observed data [52]. Another approach has been suggested by Udhayakumar *et. al.* [170] to replace r by a histogram estimator based on the number of bins, and generate an entropy profile based on multiple different rs, to reduce the sensitivity to this parameter.

5.2.2.2 Sample Entropy

A closely related technique for estimating the entropy rate is sample entropy [147], which was developed to address the issues of bias and lack of relative consistency in approximate entropy. The sample entropy, SampEn, is a simpler algorithm than ApEn, with a lower time complexity to make an estimate and eliminating self-matches in the data. Similar to the approximate entropy, it has been noted that the sample entropy corresponds to the differential Renyi entropy rate of order 2 [110].

We define sample entropy, by using very similar objects to approximate entropy. Given a time series, $x_1, ..., x_N$, of length N, we calculate substrings

$u_i^m = [x_i, ..., x_{i+m-1}]$ of length m, and choose the parameter, r, for the maximum threshold between strings. We now define two related quantities,

$$A = \sum_{i=0}^{N-m} \mathbb{1}_{\{d[u_i^{m+1}, u_j^{m+1}]<r\}},$$
$$B = \sum_{i=0}^{N-m+1} \mathbb{1}_{\{d[u_i^{m}, u_j^{m}]<r\}},$$

where $d[u_i^m, u_j^m]$ is a distance metric, with the usual distance metrics l_∞ and l_2. Finally we define the sample entropy as,

$$SampEn = -\log \frac{A}{B}.$$

As A will be always less than or equal to B, this value will always be non-negative.

Sample entropy removes the bias that is introduced via the double counting of substrings in approximate entropy, however sample entropy does not reduce the source of bias that is introduced by the correlation of the substrings used in the calculation [52, 147].

5.2.2.3 Permutation Entropy

In addition to the two related entropy rate estimation techniques, we introduce permutation entropy developed by Bandt and Pompe [7]. Unlike the previous two, this has not been shown to converge to the true entropy rate, for particular stochastic processes. However, it was developed for the same purpose, to quantify the complexity of processes generating time series data. Further development of the theory was undertaken by Bandt and Pompe, justifying the development of permutation entropy as a complexity measure [8].

Given a set of discrete-time data, $x_1, \ldots, x_N$, we consider permutations, $\pi \in \Pi$ of length n which represent the numerical order of the substring data. For example, with $n = 3$ three consecutive data points $(2, 7, 5)$ and $(3, 9, 8)$ are examples of the permutation 021, and $(5, 1, 3)$ and $(7, 4, 5)$ are examples of the permutation 201, the the numbers in the permutation represent the ordering of the substring. For every permutation π, the relative frequencies are calculated as

$$p(\pi) = \frac{|\{t | t \leq T - n, x_{t+1}, ..., x_{t+n} \text{ has type } \pi|\}}{T - n + 1}.$$

Hence, we are working with approximations to the real probabilities, however we could recover these by taking the limit as $T \to \infty$ by the Law of Large Numbers [57, pg. 73] using a characteristic function on observing the permutation, with a condition on the stationarity of the stochastic process.

The permutation entropy of a time series, of order $n \geq 2$, is then defined as,

$$H(n) = -\sum_{\pi \in \Pi} p(\pi) \log p(\pi).$$

Permutation entropy has a parameter which controls the length of the order permutations considered, the order n. The number of permutations for an order scales as $n!$, which creates a time complexity issue as the required computations grows very quickly in the size of the order. Hence, the minimum possible data required to observe all of the possible permutations of order n, is $n!$ data. However, it is claimed that the permutation entropy is robust to the order of the permutations used [7]. In practice smaller n's are used, such as $n = 3, 4, 5$ due to the growth of the number of permutations which requires more data to observe all of the permutations [7]. There is another parameter, embedding delay, which is the period of the elements that are considered. That is, for a sample, $x_1, x_2, \ldots$ for an embedding delay of τ the elements used in the permutation entropy calculation are $x_1, x_{1+\tau}, x_{1+2\tau}$. In this work we will be using $\tau = 1$.

5.2.2.4 Specific Entropy Rate

The specific entropy rate was defined by Darmon [47], to provide a differential entropy rate estimation technique that has a stronger statistical footing than the previously defined estimation techniques. The intent of the development of this quantity was to create a measure of the complexity of a continuous-valued, discrete-time time series, as a function of its state. Then a differential entropy rate estimate is made by taking a time average of the specific entropy rate estimates, and therefore can be applied in particular to ergodic processes. The approach is to consider the short-term predictability of a sequence, by utilising a finite history of values to create a kernel density estimate of the conditional probability density function. Then use the kernel density estimate to plug-in to the differential entropy rate formula, when using the conditional density function. For the calculation of this quantity, a parameter of the length of the history, p, is used for a kernel density estimate of the conditional probability density function.

The definition of the specific entropy rate, makes a finite truncation of the conditional entropy version of the entropy rate, from Theorem B.2.1. One

condition is required in the formulation of the theoretical basis, which is that the process being measured is conditionally stationary. That is, given the conditional distribution function of X_{t+1}, conditional on $(X_t, \ldots, X_{t-p+1}) = \mathbf{X}$, does not depend on the value of t for a fixed length of history being considered, p. In the paper by Darmon [47], they show that the conditional entropy up to order p, depends on the state specific entropy rate of a particular history $(x_p, \ldots, x_1) = \mathbf{x_1^p}$ and the density of the possible pasts $(X_p, \ldots, X_1) = \mathbf{X_1^p}$. This is shown by an argument which establishes that,

$$h\left(X_t|\mathbf{X_{t-p}^{t-1}}\right) = -E\left[E\left[\log f\left(X_t|\mathbf{X_{t-p}^{t-1}}\right)\right]\right].$$

Where the first expectation is with respect to $f(\mathbf{x_1^p})$, and the second expectation is with respect to $f(X_{p+1}|\mathbf{x_1^p})$. Given this relationship and the law of total expectation, the specific entropy rate, of order p, $h_t^{(p)}$, is defined as

$$\begin{aligned} h_t^{(p)} &= h\left(X_t|\mathbf{X_{t-p}^{t-1}} = \mathbf{x_{t-p}^{t-1}}\right), \\ &= -E\left[\log f\left(X_t|\mathbf{X_{t-p}^{t-1}}\right)\right], \\ &= -\int_{-\infty}^{\infty} f\left(x_{p+1}|\mathbf{x_1^p}\right) \log f\left(x_{p+1}|\mathbf{x_1^p}\right) dx_{p+1}. \end{aligned}$$

Hence, the specific entropy rate estimator, $\hat{h}_t^{(p)}$, defined by plugging in the estimate of the density obtained by kernel density estimation, $\hat{f}\left(x_{p+1}|\mathbf{x_1^p}\right)$, is

$$\hat{h}_t^{(p)} = -E\left[\log \hat{f}\left(x_{p+1}|\mathbf{x_1^p}\right)\right].$$

Using the specific entropy rate an estimate of the differential entropy rate of order p, $\hat{h^{(p)}}$, is defined as

$$\begin{aligned} \hat{h}^{(p)} &= \frac{1}{T-p}\sum_{t=p}^{T} \hat{h}_t^{(p)}, \\ &= \frac{1}{T-p}\sum_{t=p}^{T} -E\left[\log \hat{f}\left(x_{p+1}|\mathbf{x_1^p}\right)\right], \end{aligned}$$

which is the time average of all the specific entropy rates across the observed states.

Darmon [47] implemented a version of the specific entropy rate by using kernel density estimation to estimate the conditional entropy, based on the past. The specific entropy rate implementation relies on some parameters to construct the kernel density estimation, which is the length of the past, p and

the $p+1$ bandwidths, $k_1, \ldots, k_{p+1}$ that are used in the kernel density estimation [47]. The parameter choice can have large impacts on the quality of the estimation, in particular depending on how long the past that is considered. The suggested technique for selecting p is a cross-validation technique which removes an individual observation and l observations either side. Then the following expression is minimised for its parameters $p, k_1, \ldots, k_{p+1}$,

$$CV(p, k_1, \ldots, k_{p+1}) = -\frac{1}{T-p} \sum_{t=p+1}^{T} \log \hat{f}_{-t:l} \left(X_t | X_{t-p}^{t-1} \right).$$

where $\hat{f}_{-t:l}$ is the conditional density with the points removed [47]. A suggested approach is to take $l = 0$ and only remove the individual observation [48]. In practice, it is advised to fix p and then calculate the bandwidths due to the computational complexity of the cross-validation [47].

Chapter 6

Robust Estimation for LRD Processes

Estimation of entropy rate is a classical problem in information theory. The entropy rate is the asymptotic limit of the per sample average entropy of a discrete-time stationary stochastic process as defined in Definition 2.1.7. It is used as a measure of the complexity of the process and thus to perform comparisons, and detect anomalies.

Estimation of entropy rate is comparatively easy when the underlying stochastic process is SRD. The reality is that many real data sequences are LRD. Nonparametric approaches to estimation have the very significant advantage that they do not depend on a fitting a model of the data and hence have a degree of robustness missing from parametric estimators, in particular when estimating processes exhibiting LRD.

A great deal of theory has been developed on nonparametric entropy rate estimation from data from finite alphabets [107, 156], since these models form the basis of discrete codes for communications. These estimation techniques have been extended to countably infinite alphabets [106, 143]. However, these approaches cannot be directly applied to continuous-valued stochastic processes, which are a main topic of interest in this chapter.

There are several approaches that are designed for estimating the (differential) entropy rate of continuous-valued stochastic processes: approximate [142], sample [147] and permutation entropy [7]. These estimators have been used in estimating the entropy rate for processes, particularly for those that are memoryless or have short memory. For example, approximate entropy has been shown to converge to the entropy rate for independent and identically distributed processes and first order Markov chains in the discrete-valued case [142]. However, at best, these estimators are sensitive to their parameter choices; at worst we shall see that they have severe defects when

Estimation Technique	Values		Estimation Quality			Computation	
	Discrete	Continuous	Consistent	Asymp. unbiased	Correlation Length	Complexity	Time (s)
Grassberger [79]	✔	✖	✖	✖	$\approx \log(N)$		
Kontoyiannis and Suhov [107]	✔	✖	✔	✔	$\approx \log(N)$		
Statentropy [162]	✔	✖	✔	✔	$\approx \log(N)$		
Vatutin and Mikhailov [172]	✔	✖	✖	✖	N		
Approximate Entropy [142]	✔	✔	✖	✖	m	$O(N^2)$	10.01
Sample Entropy [147]	✔	✔	✖	✖	m	$O(N^2)$	263.0
Permutation Entropy [7]	✔	✔	✖	✖	n	$O(n!N)$	0.82
Specific Entropy Rate [47]	✖	✔	✔	✔	p	$O(N^2p)$	504,219.9
NPD Entropy	✖	✔	✔	✔	$\approx \log(N)$	$O(N\log(N))$	39.96

Table 6.1: Comparison of differential entropy rate estimators. The discrete-value (only) estimators have more desirable properties: being consistent and asymptotically unbiased. Approximate, sample and permutation entropy can be applied to either discrete or continuous valued sequences but these are biased and inconsistent. Correlation length refers to the longest lag at which correlations are included into the entropy into the estimate where the length of data is N, the length of substrings matched in approximate and sample entropy is m, the order of permutations used in permutation entropy is n, and p is a specified parameter of specific entropy rate. Note also that although specific entropy rate behaves relatively well, the computation time of the author's R implementation are prohibitive. We have only included the complexity and computation time for the continuous-valued estimators that we test in this chapter. The computation time is the total time to make estimates on 1000 time series of length 1000, with more details of the experiment in Section 6.4.2.

applied to LRD processes.

In this chapter, we compare the existing approaches and develop a new nonparametric differential entropy rate estimator – NPD Entropy – for continuous-valued stochastic processes. It combines the best of the existing discrete alphabet nonparametric estimators with the standard signal processing techniques to obtain an estimate that is more reliable than the alternatives, in particular when applied to LRD processes. We have implemented this estimator in Python as a package, NPD-Entropy, that is available on GitHub[1].

Table 6.1 outlines our results. Notably, several discrete alphabet approaches come with guarantees of consistency but cannot be applied directly to continuous-valued processes as they are based on matching strings. On the other hand, the main approaches that have been applied to continuous values – approximate, sample and permutation entropy – are not consistent, except in special parametric cases, and what's more the use of (short) finite windows limits their ability to cope with processes with extended correlations.

[1] https://github.com/afeutrill/npd_entropy

We show, for instance, that these estimation techniques do not make accurate estimates for the entropy rate for processes whose dependency structure has slowly decaying correlations.

We examine these entropy rate estimates performance on data generated by long range dependent (LRD) processes. We apply the estimation approaches to two common LRD processes: Fractional Gaussian Noise and ARFIMA(0,d,0), for which we have expressions for the differential entropy rate that can be directly evaluated, see equations 3.4 and 3.6 respectively. Thus we can show exactly how bad some estimators are when applied to an even slightly challenging data set.

Another alternative – specific entropy rate – was developed as a technique to calculate the predictive uncertainty for a specific state of a continuous-valued process [47]. This approach utilises more rigorous statistical foundations to estimate the entropy rate of a state given the observation of a finite past. The technique is able to make accurate entropy rate estimates by calculating the average over the states. This is able to capture the complex dependency structure with past observations. However this comes at a large computational cost, and hence cannot be used for large sequences, or for streaming data to make online estimates (see Table 6.1 for computation time comparisons).

Our estimation technique – NPD Entropy – utilises the extensive research into nonparametric estimation of the entropy rate for discrete alphabets to provide estimation techniques that are robust to strongly correlated processes. We utilise a connection between the Shannon entropy and differential entropy [41, pg. 248], and then extend it to the case of Shannon and differential entropy rates. The technique quantises the continuous-valued data into discrete bins, then makes an estimation of the quantised process using discrete alphabet techniques. Then the differential entropy rate estimate is calculated by adjusting by the quantitative difference between the differential entropy rate and the Shannon entropy rate of the quantised process.

We show that NPD Entropy inherits many useful estimation properties from discrete alphabet estimators, including consistency and bias. Hence, by choosing a finite alphabet rate estimator that has a set of the desired properties we can ensure that NPD Entropy also has these properties for inference on continuous-valued data. We show that NPD Entropy estimates perform well in the estimation of differential entropy rate of stochastic processes which have more complex dependency structure, in particular LRD processes

We also compare the runtime performance of techniques and find that quantised estimation can make much faster estimates than any approach of comparable accuracy.

6.1 Performance of existing estimation techniques

In this section we will analyse samples of continuous valued, discrete time LRD data using the existing continuous-value entropy rate estimators. Most analysis and tests of Approximate and Sample entropy have been based on i.i.d. processes or finite state Markov chains *i.e.,* processes with either very short or no correlations. However, many real processes, in particular the types of processes for which complexity measures are useful, exhibit long-range correlations. To create sample data for testing we created 50 samples of 2000 data points, and averaged the estimates to get the values presented. Figure 6.1 shows just how bad common estimators are for FGN. Both the shape and scale of the entropy estimate curves are quite wrong (with the exception of the shape of the sample entropy). Note that the adapted Shannon entropy estimators are all positive, hence will not be able to estimate the negative values of the differential entropy rate.

One might be concerned that the results stem from particular parameter choices, or other details of the estimates, so we investigate further below.

Figures 6.2 and 6.3 show the entropy rate estimates for the Sample and Approximate Entropy for both Fractional Gaussian Noise and ARFIMA(0,d,0) for two different parameters $m = 2, 3$ and $r = 0.2$ as recommended in Delgado-Bonal and Marshak [52].

Sample Entropy approximates the the shape of the real entropy rate functions but provides large overestimates. This reinforces the use of Sample Entropy as a measure of complexity of a time series, as long as it is not claimed to be an estimate of differential entropy rate. Unfortunately the name implicitly makes this claim.

Approximate Entropy, however, fails to even approximate trend or range of values. Interestingly, this technique is also quite sensitive to changes in the value of m.

Figure 6.4 shows the Permutation Entropy of order $n = 3$. These results are indicate that permutation entropy is not a good choice for strongly correlated process as all estimates exist on a very small scale, and there seems to be little difference in measured complexity across the range of $\mathcal{H}$. The maximum Permutation Entropy for any process, with $n = 3$, is $H(\pi) = \log_2(3!) \approxeq 2.585$, and all of the estimates are within 0.2 of the maximum value. Hence, this measure does not seem to able to pick up the entropy rate trend in LRD data. Although the trend looks promising on a small scale, once we compare the behaviour to the actual entropy rate, it is a poor estimator.

Naively, these estimates can be improved by extending the lengths of the windows being used, however, for LRD processes large windows would be required. Unfortunately the computational complexity of the approaches (see

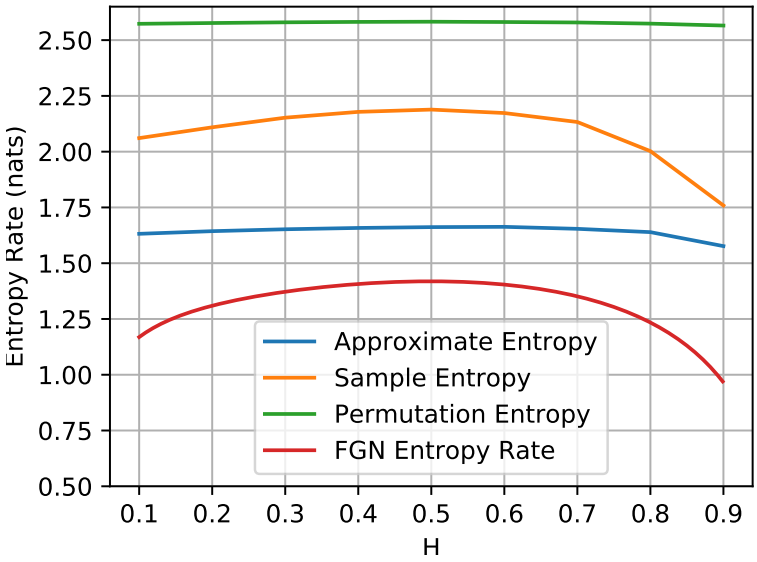

Figure 6.1: Approximate, sample and permutation entropy estimates for FGN. The red curve shows the true entropy rate for FGN processes. Note the wide discrepancies in the estimates.

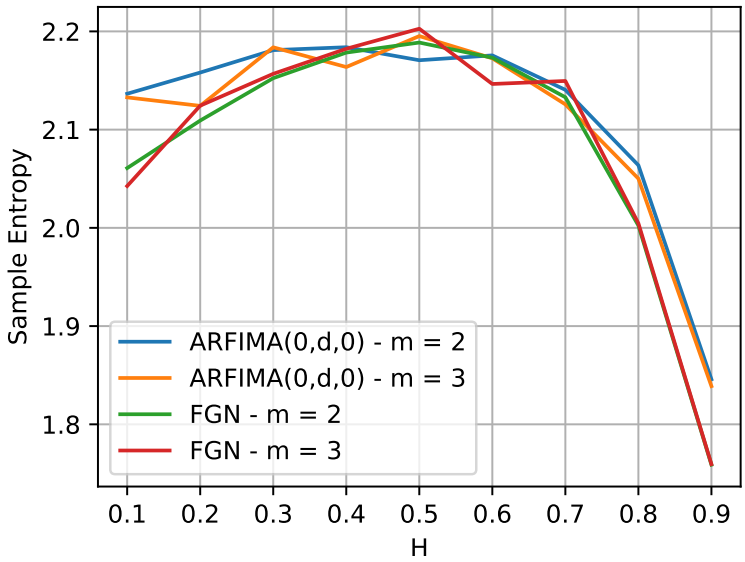

Figure 6.2: Entropy-rate estimates of FGN and ARFIMA(0,d,0) with process variance $\sigma^2 = 1$ using Sample Entropy ($r = 0.2$). The estimates show the rough trend of the LRD processes, however it overestimates the differential entropy by a large amount. See Figure 6.1, note that the parameter choice affects the results but not in a useful manner.

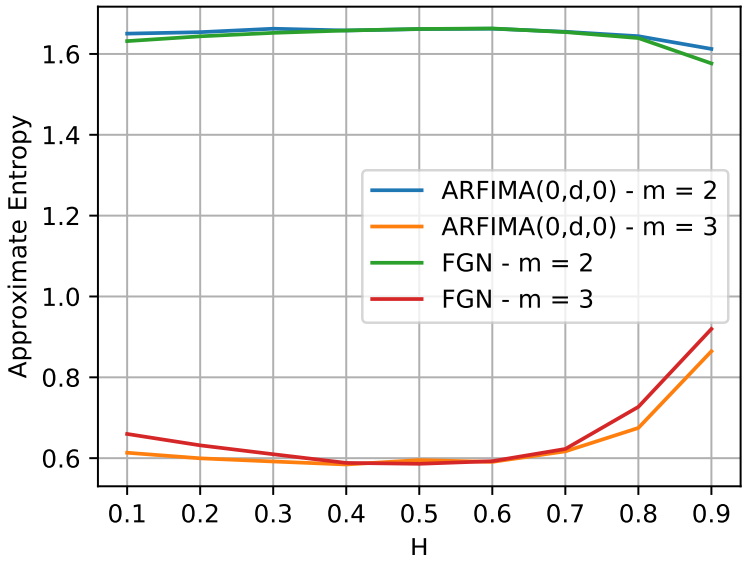

Figure 6.3: Entropy-rate estimates of FGN and ARFIMA(0,d,0) with process variance $\sigma^2 = 1$ using Approximate Entropy ($r = 0.2$). It fails to even approximate trend or range of values. Interestingly, this technique is also quite sensitive to changes in the value of m.

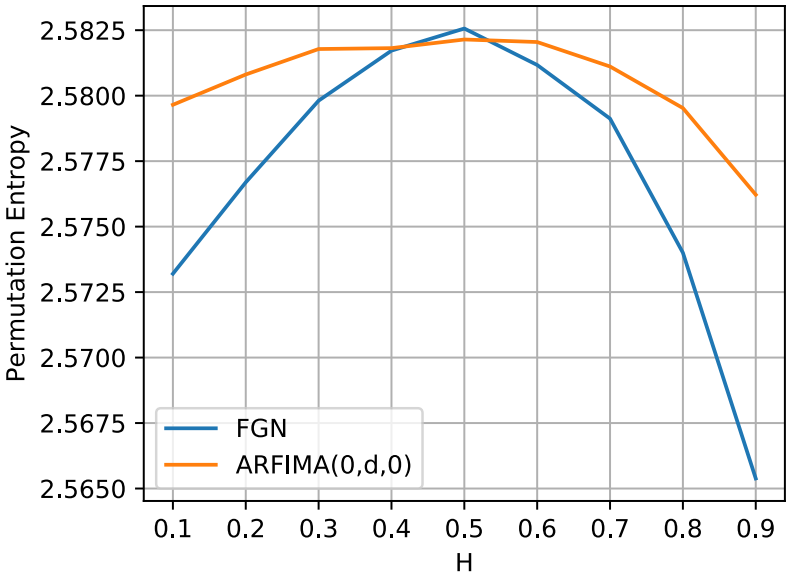

Figure 6.4: Entropy-rate estimates of FGN and ARFIMA(0,d,0) with process variance $\sigma^2 = 1$ using Permutation Entropy ($n = 3$). This appears to capture the trend of the entropy rate function, however the scale of the changes is wrong. Noting that the maximum Permutation Entropy for any process is $\log_2(3!) \approx 2.585$ the range displayed is very small.

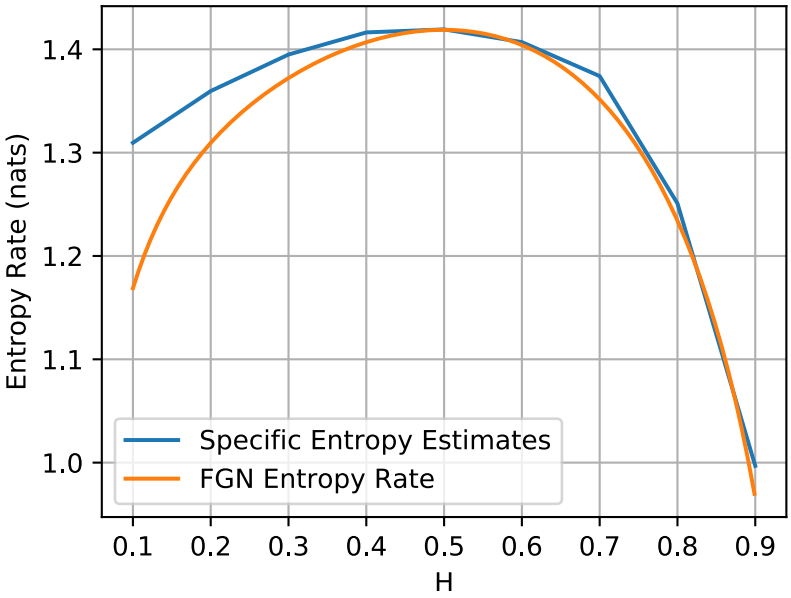

Figure 6.5: Specific Entropy rate ($p = 10$) estimates of the entropy rate of FGN. These estimates are very good over the range of $\mathcal{H}$, however the estimates start to diverge for smaller values of $\mathcal{H}$.

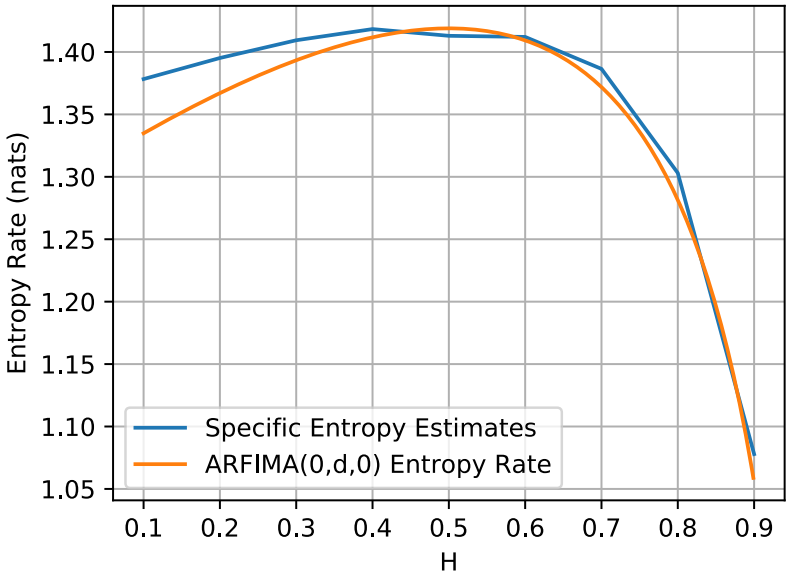

Figure 6.6: Specific Entropy rate ($p = 10$) estimates of the entropy rate of FGN. These estimates are very good over the range of $\mathcal{H}$, however the estimates diverge as $\mathcal{H}$ tends to 1.

Table 6.1) grows with increases window sizes and we see stability problems at least with Approximate Entropy, so there does not appear to be a suitable trade-off between computational cost and accuracy.

Estimates derived from Specific Entropy rate are shown in Figures 6.5 and 6.6, with $p = 10$. Specific entropy rate provides good agreement with the entropy rate for LRD FGN and ARFIMA(0,d,0), for $\mathcal{H} > \frac{1}{2}$, with greater divergence occurring in the CSRD parameter range for small $\mathcal{H}$ values. However, note that the computational cost for these estimates is very high, being over 500,000 seconds to make 1000 estimates using 1000 data points.

6.2 The Link between Shannon and Differential Entropy Rates

Motivated by the problems in existing continuous-value entropy rate estimators, as illustrated in the previous section, we make a connection between the differential and Shannon entropy rate, which we use as the basis of an estimation technique. We propose an approach where we quantise the continuous valued process into a discrete valued process, then apply estimators from the discrete valued domain, and then translate back to differential entropy. Note that we will be using natural logarithms for differential entropy, and therefore we will be using the units of nats throughout the chapter.

Given the definitions of Shannon and differential entropy, Definitions 2.1.1 and 2.1.6 respectively, we utilise a connection that exists between these quantities for a quantised version of continuous data. From Cover and Thomas [41, pg. 248], a link is defined between differential entropy and Shannon entropy, for a quantised window size Δ and the associated Shannon entropy $H(X^{\Delta})$, where $X^{\Delta} = x_i$, if $i\Delta \leq X < (i+1)\Delta$. This is made explicit in Theorem 2.1.6 of the background with $H\left(X^{\Delta}\right) + \log\left(\Delta\right) \to h(X)$ as $\Delta \to 0$.

Hence, if we quantise and apply Shannon entropy estimators we may be able to make a useful estimation of differential entropy, particularly with finer quantisations and more data. We extend this relationship to the case of Shannon entropy rate and differential entropy rate, by considering the joint entropy of a quantised process and then taking the limit as $n \to \infty$. We clarify the relationship between the entropy rates in the following theorem, which is an extension of Theorem 8.3.1 of Cover and Thomas [41].

Theorem 6.2.1. *If the joint density function, $f(x_1, ..., x_n)$, of a stationary stochastic process, $\chi = \{X_m\}_{m\in\mathbb{N}}$ is continuous and Riemann integrable $\forall n \in \mathbb{N}$ and $f(x_1, ..., x_n) \log f(x_1, ..., x_n)$ is Riemann integrable $\forall n \in \mathbb{N}$, then*

$$H(\chi^{\Delta}) + \log(\Delta) \to h(\chi), \text{ as } \Delta \to 0.$$

Remark. *An implication, as noted in Cover and Thomas [41] is that the entropy rate of an n-bit quantisation of a continuous-valued, discrete-time stochastic process, χ is approximately $h(\chi) + n$, when using* $\log_2$ *in the expression above.*

Proof. The beginning of the proof follows an identical argument to Theorem 8.3.1 of Cover and Thomas [41], but for multiple random variables. Note that Cover and Thomas use the i as an index for the bin of the quantised value, and for this we use the index j We use i for the index of the random variable of the stochastic process. For all finite $n \in \mathbb{N}$, the joint density of the finite collection of random variables is given by $f(x_1, ..., x_n)$. We partition the range of each random variable X_i into bins of length Δ with the random variables indexed by $i \in \{1, \ldots, n\}$. Similar to equation 8.23 of Cover and Thomas [41], by the mean value theorem for multiple variables [182], there exists $(x_1^{(j)}, ..., x_n^{(j)}) \in \mathbb{R}^n$, such that

$$f(x_1^{(j)}, ..., x_n^{(j)})\Delta^n = \int_{j_1\Delta}^{(j_1+1)\Delta} ... \int_{j_n\Delta}^{(j_n+1)\Delta} f(x_1, ..., x_n)dx_1 \ldots dx_n.$$

Where the $x_i^{(j)}$ are such that,

$$j_i\Delta \leq x_i^{(j)} < (j_i + 1)\Delta.$$

Where we denote by $x_i^{(j)} \in \mathbb{R}$ the quantised value of the bin indexed by j_i for the ith random variable X_i. Then we consider the quantised random variables, X_i^Δ, like equation 8.24 of Cover and Thomas [41], by partitioning each random variable, for $i \in \{1, \ldots, n\}$, into bins indexed by $j_i \in \mathbb{Z}$,

$$X_i^\Delta = x_i^{(j)}, \text{ if } j_i\Delta \leq X_i < (j_i + 1)\Delta.$$

Then, like equation 8.25 of Cover and Thomas [41], the probability that $\left(X_1^\Delta, ..., X_n^\Delta\right) = \left(x_1^{(j)}, ..., x_n^{(j)}\right)$ is given by

$$\begin{aligned}\mathbb{P}(\left(X_1^\Delta, ..., X_n^\Delta\right) = \left(x_1^{(j)}, ..., x_n^{(j)}\right)) &= \int_{j_1\Delta}^{(j_1+1)\Delta} ... \int_{j_n\Delta}^{(j_n+1)\Delta} f(x_1, ..., x_n)dx_1 \ldots dx_n, \\ &= f(x_{j_1}, ..., x_{j_n})\Delta^n.\end{aligned}$$

Where we use the notation

$$\mathbb{P}\left(\left(X_1^\Delta, ..., X_n^\Delta\right) = \left(x_1^{(j)}, ..., x_n^{(j)}\right)\right) = \mathbb{P}(x_1^{(j)}, ..., x_n^{(j)}).$$

We define $\mathbb{P}(X_n^\Delta = x_n'^{(j)} | X_{n-1}^\Delta = x_{n-1}'^{(j)}, ..., X_1^\Delta = x_1'^{(j)})$ with the simpler notation

$$\begin{aligned}\mathbb{P}(x_n'^{(j)} | x_{n-1}'^{(j)}, ..., x_1'^{(j)}) &= \int_{j_n\Delta}^{(j_n+1)\Delta} f(x_n | x_{n-1}, ..., x_1) dx_n, \\ &= f(x_n'^{(j)} | x_{n-1}'^{(j)}, ..., x_1'^{(j)})\Delta.\end{aligned}$$

Where the second equality is due to the single variable mean value theorem. Note that we differentiate that the value for the conditional density, $x_n'^{(j)}$, for the single value mean value theorem and the multivariate density, $x_n^{(j)}$, used in the multivariate mean value theorem may be different.

Hence, following the argument of Cover and Thomas [41], we consider the conditional Shannon entropy of the quantised random variable,

$$\begin{aligned}H(X_n^\Delta | X_{n-1}^\Delta, ..., X_1^\Delta) &= -\sum_{j_1=-\infty}^{\infty} \dots \sum_{j_n=-\infty}^{\infty} \mathbb{P}(x_1^{(j)}, ..., x_n^{(j)}) \log(\mathbb{P}(x_n'^{(j)} | x_{n-1}'^{(j)}, ..., x_1'^{(j)})), \\ &= -\sum_{j_1=-\infty}^{\infty} \dots \sum_{j_n=-\infty}^{\infty} f(x_1^{(j)}, ..., x_n^{(j)})\Delta^n \log(f(x_n'^{(j)} | x_{n-1}'^{(j)}, ..., x_1'^{(j)})\Delta), \\ &= -\sum_{j_1=-\infty}^{\infty} \dots \sum_{j_n=-\infty}^{\infty} f(x_1^{(j)}, ..., x_n^{(j)})\Delta^n \log(f(x_n'^{(j)} | x_{n-1}'^{(j)}, ..., x_1'^{(j)})) \\ &\qquad - \sum_{j_1=-\infty}^{\infty} \dots \sum_{j_n=-\infty}^{\infty} f(x_1^{(j)}, ..., x_n^{(j)})\Delta^n \log(\Delta).\end{aligned}$$

The joint density, $f(x_1, ..., x_n)$, and conditional density $f(x_n | x_{n-1}, ..., x_1)$ are Riemann integrable $\forall n \in \mathbb{N}$, which implies that as $\Delta \to 0$,

$$\begin{aligned}&-\sum_{j_1=-\infty}^{\infty} \dots \sum_{j_n=-\infty}^{\infty} f(x_1^{(j)}, ..., x_n^{(j)})\Delta^n \log(f(x_n'^{(j)} | x_{n-1}'^{(j)}, ..., x_1'^{(j)})) \\ &\qquad \to -\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(x_1, ..., x_n) \log(f(x_n | x_{n-1}, ..., x_1)) dx_1 \dots dx_n, \\ &\qquad = h(X_n | X_{n-1}, \dots, X_1).\end{aligned}$$

We have

$$\sum_{j_1=-\infty}^{\infty} \dots \sum_{j_n=-\infty}^{\infty} f(x_1^{(j)}, ..., x_n^{(j)})\Delta^n = 1,$$

since it is the sum over all possibilities of a probability mass function. Therefore, we get as $\Delta \to 0$,

$$H(X_n^\Delta | X_{n-1}^\Delta, \dots, X_1^\Delta) + \log(\Delta) \to h(X_n | X_{n-1}, \dots, X_1).$$

From this point on we diverge from the single variable proof of Theorem 8.3.1 of Cover and Thomas [41], since we require analysis of the two limits as $\Delta \to 0$ and $n \to \infty$ simultaneously. Now we define a function

$$g(k,n,m) = H(X_i^{2^{-n}} | X_{i-1}^{2^{-m}}, ..., X_{i-k}^{2^{-m}}) + \log(2^{-n}).$$

We note that

$$g(k+1,n,m) \leq g(k,n,m),$$

since conditioning cannot increase entropy. We have that

$$g(k,n+1,m) \leq g(k,n,m),$$

and

$$g(k,n,m+1) \leq g(k,n,m),$$

since the entropy is not increased by considering a smaller bin size, and by conditioning on smaller bins.

Since for functions, $f(n,m)$, with

$$f(n+1,m), f(n,m+1) \leq f(n,m)$$

we have from Schilling [153, pg. 29]

$$\inf_{n\in\mathbb{N}} \inf_{m\in\mathbb{N}} f(n,m) = \inf_{n\in\mathbb{N}} f(n,n) = \inf_{m\in\mathbb{N}} \inf_{n\in\mathbb{N}} f(n,m). \tag{6.1}$$

Therefore by the definition of the differential entropy rate

$$\begin{aligned}
h(\chi) &= \lim_{k\to\infty} h(X_k | X_{k-1}, ..., X_1), \\
&= \lim_{k\to\infty} \lim_{\Delta\to 0} H(X_k^{\Delta} | X_{k-1}^{\Delta} \ldots, X_1^{\Delta}) + \log(\Delta), \text{ by Theorem 2.1.6} \\
&= \inf_{k\in\mathbb{N}} \inf_{n\in\mathbb{N}} \inf_{m\in\mathbb{N}} g(k,n,m), \text{ since the infima of g is equivalent to the limit,} \\
&= \inf_{k\in\mathbb{N}} \inf_{n\in\mathbb{N}} g(k,n,n), \text{ by the expression (6.1) with respect to } k \text{ and } m, \\
&= \inf_{n\in\mathbb{N}} \inf_{k\in\mathbb{N}} g(k,n,n), \text{ by the expression (6.1) swapping the order of infima,} \\
&= \inf_{n\in\mathbb{N}} \lim_{k\to\infty} \left[H(X_i^{2^{-n}} | X_{i-1}^{2^{-n}}, ..., X_{i-k}^{2^{-n}}) + \log(2^{-n}) \right], \\
&= \lim_{\Delta\to 0} \left[H(\chi^{\Delta}) + \log(\Delta) \right].
\end{aligned}$$

□

Remark. *In the process of proving this theorem we showed that in the limit as* $\Delta \to 0$ *there is a link between the Shannon conditional entropy of the quantised process and the conditional differential entropy. This is,*

$$h(X_n^{\Delta}|X_{n-1}^{\Delta},\ldots,X_1^{\Delta}) + \log(\Delta) \to h(X_n|X_{n-1},\ldots,X_1),$$

for any collection of random variables of length n*.*

As an example, if we consider the quantisation of length $\Delta = 1$, then the quantised process should be close to the differential entropy as,

$$\begin{aligned} H(\chi_1^{\Delta}) + \log(1) &= H(\chi_1), \\ &\approx h(\chi), \end{aligned}$$

and we can make an approximation of the differential entropy rate, which we will use as the basis of an estimation technique in the following section. However, there will be an error which is due to the difference between the real differential entropy rate, and the approximation of the integral at the quantisation size, Δ.

Another aspect to note is that as the quantisation window gets finer, *i.e.*, as Δ gets smaller, the term in the estimator, $\log(\Delta) \to -\infty$. There is a potential concern that this disparity between the correction term and the actual estimate could lead to numerical errors. Thus, smaller quantisations aren't necessarily better. There may be a range of values of Δ, where the estimation is closest to the real value, and we can trade off these potential errors. In the next section, we will define an estimator using the connection in Theorem 6.2.1 and investigate empirically the ideal range of choices of Δ for practical estimation.

6.3 NPD-Entropy Estimator

Theorem 6.2.1 gives us a way to convert an estimator of entropy rate for discrete-valued, discrete-time stochastic processes into an estimator for a continuous-valued process. We call this strategy NPD-Entropy estimation and explore its properties in this section. In particular we develop links from the properties of the Shannon entropy rate estimator that is selected to the corresponding NPD-estimator. In the following arguments we assume that the differential entropy rate and Shannon entropy rate of the quantised sequence exist.

We begin with the definition of NPD-Entropy, short for Non-Parametric Differential Entropy estimator.

Definition 6.3.1 (NPD-Entropy)**.** *The NPD-Entropy estimator of the differential entropy rate, $h(\mathcal{X})$, of a continuous-valued, discrete time stationary stochastic process, $\mathcal{X} = \{X_i\}_{i\in\mathbb{N}}$, using a Shannon entropy rate estimator, $H(\mathcal{X}^\Delta)$, of the corresponding quantised discrete-time, discrete-valued stochastic process $\mathcal{X}^\Delta = \{X_i^\Delta\}_{i\in\mathbb{N}}$ with window size, Δ, is defined as*

$$\hat{h}_{NPD}(\mathcal{X}) = \hat{H}(\mathcal{X}^\Delta) + \log(\Delta).$$

We analyse some properties of NPD-Entropy and start by considering the consistency of the estimation technique.

Theorem 6.3.1. *NPD-Entropy, $\hat{h}_{NPD}$, is a consistent estimator as $\Delta \to 0$, if and only if the associated Shannon entropy rate estimator, $\hat{H}$, is consistent,* i.e., *as $\Delta \to 0$ the following two are equivalent:*

$$\lim_{n\to\infty} \hat{H}(\mathcal{X}_n^\Delta) \to H(\mathcal{X}^\Delta),$$
$$\lim_{n\to\infty} \hat{h}_{NPD}(\mathcal{X}_n) \to h(\mathcal{X}),$$

where n denotes the length of the data sequence to which the estimator is applied.

Proof. Appling the definition of consistency gives

$$\lim_{n\to\infty} \hat{h}_{NPD}(\mathcal{X}_n) \quad = \quad \lim_{n\to\infty} \hat{H}(\mathcal{X}_n^\Delta) + \log(\Delta).$$

Applying Theorem 6.2.1 and the consistency of the Shannon entropy rate estimator gives the result. □

Remark. *This is a very general result, that only relies on the existence of a consistent estimator for discrete valued stochastic processes. The argument is agnostic to the mode of convergence used, and the strength of the convergence result depends on which mode of convergence is used for the consistency of the Shannon entropy rate estimator. That is, if we have convergence in probability of the Shannon entropy rate estimator, then we have convergence in probability of the resulting differential entropy rate estimator.*

Note that this result gives consistency of the estimator in the limit. In practice, we will fix the window size to use for online estimation. However, this will result in an estimator that is overbiased.

We continue this discussion by considering some other properties that are useful for estimation, in particular the bias, variance and mean squared error. We summarise these facts in the following theorem, and show how we can inherit properties from discrete estimation.

Theorem 6.3.2. *The NPD-Entropy, $\hat{h}_{NPD}$, constructed from an estimator of the Shannon entropy rate, $\hat{H}$ has the following properties:*

1. *Bias$_H[\hat{H}_n(\chi^\Delta)] \to$ Bias$_h[\hat{h}_{NPD}(\chi)]$,*
2. *Var$(\hat{H}_n(\chi^\Delta)) \to$ Var$(\hat{h}_{NPD}(\chi))$,*
3. *MSE$(\hat{H}_n(\chi^\Delta)) \to$ MSE$(\hat{h}_{NPD}(\chi))$,*

as $\Delta \to 0$.

Proof. From the definition of bias [146, pg. 126] we have

$$\begin{aligned}
\text{Bias}_H[\hat{H}_n(\chi^\Delta)] &= E[\hat{H}_n(\chi^\Delta)] - H(\chi^\Delta), \\
&\to E\left[\hat{h}_{NPD}(\chi) - \log(\Delta)\right] - (h(\chi) - \log(\Delta)), \text{ as } n \to \infty, \\
&= E[\hat{h}_{NPD}(\chi)] - h(\chi), \\
&= \text{Bias}_h[\hat{h}_{NPD}(\chi)].
\end{aligned}$$

Where we can exchange the limit as $n \to \infty$ with the expectation in the second equality, since the expectation of $\hat{H}$ is finite.

The result for the mean squared error of the estimator follows by an identical argument, substituting, $\hat{h}_{NPD} - \log(\Delta)$ and $h - \log(\Delta)$ for $\hat{H}$ and H respectively as $n \to \infty$, therefore we omit the argument.
We consider another characterisation of the mean squared error, known as the bias-variance decomposition [85, pg. 24],

$$\text{MSE}[\hat{H}] = \text{Bias}_H[\hat{H}]^2 + \text{Var}(\hat{H}),$$

and placing in terms of the variance and substituting the bias and mean square error equivalences as $n \to \infty$, we get

$$\begin{aligned}
\text{Var}(\hat{H}) &= \text{Bias}_H[\hat{H}]^2 - \text{MSE}[\hat{H}], \\
&\to \text{Bias}_h[\hat{h}_{NPD}]^2 - \text{MSE}[\hat{h}_{NPD}], \text{ as } n \to \infty, \\
&= \text{Var}(\hat{h}_{NPD}).
\end{aligned}$$

□

Note that the NPD Entropy estimator itself will be biased for any finite quantisation window, since the results hold in the limit as $n \to \infty$.

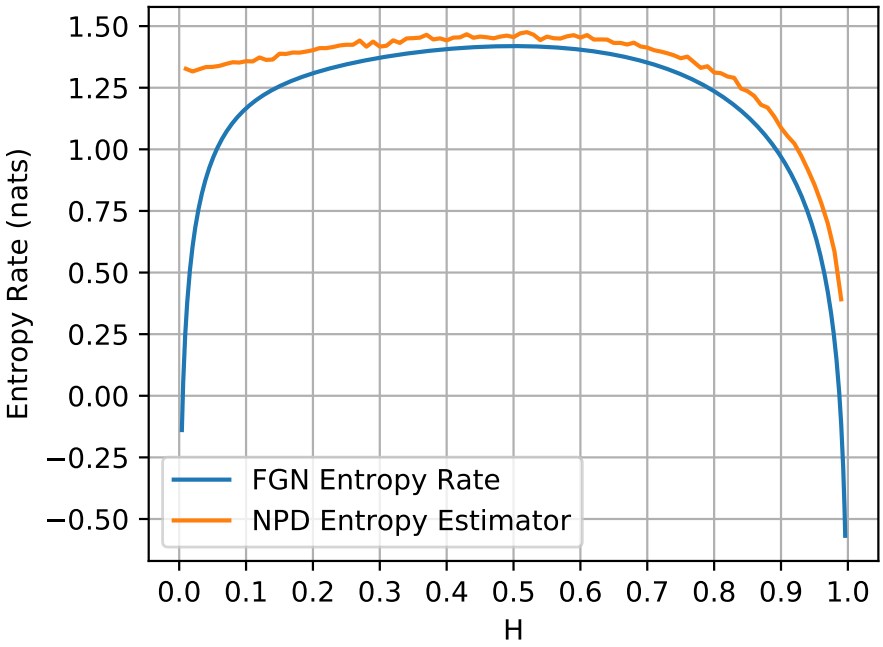

Figure 6.7: NPD Entropy estimates with $\Delta = 1$ compared to actual entropy rate of FGN. The estimates have a small bias, with the bias increasing as $\mathcal{H} \to 0$ and the entropy rate function asymptotically tends to $-\infty$.

6.4 Evaluation of Performance

To test this concept, we have implemented an estimator using Lempel-Ziv string matching [186] matching, as described by Grassberger [79] to create a quantised entropy rate estimator. The conditions in Kontoyiannis and Suhov [106] ensure that we produce a consistent Shannon entropy rate estimator. We will use these to then make estimates of the differential entropy rate.

The estimator was implemented in Python using the NumPy library [84]. The estimator, NPD-Entropy is available in the GitHub repository, referred to in the beginning of the chapter.

We were able to parallelise the implementation of NPD Entropy to increase the performance of the estimation technique. This is because we can calculate the length of each prefix sequence, $L^n(x)$, independently, without knowledge of any other prefix sequence. This has been implemented using the Python library Numba [112], which is able to parallelise the loop to calculate the length of the prefixes, $L(x)$. This approach is able to speed up the

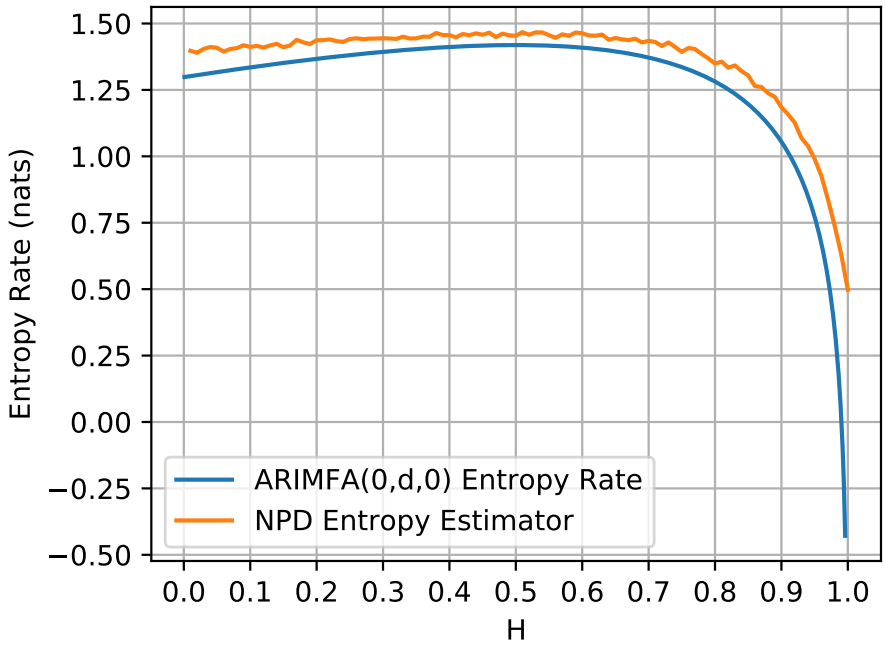

Figure 6.8: NPD Entropy estimates with $\Delta = 1$ compared to actual entropy rate of ARFIMA. Note the close correspondence, with a small bias.

implementation, and hence allows the algorithm to be able to be run online, for quick estimation of entropy rates of streaming data.

The estimators were tested on data generated by FGN and ARFIMA(0,d,0) processes. We generated 50 test samples of 2000 data points for each process, where the length was chosen to capture the longer term trends of sequences of LRD processes. The estimates for each process were then averaged, to show the mean estimate at each value of $\mathcal{H}$. The FGN realisations were generated using the Davies-Harte method [49], and the python package *fbm* [66]. The ARFIMA(0,d,0) realisations were generated using the Durbin-Levinson algorithm [130], from the R package *arfima* [173].

From observing both Figure 6.7 and Figure 6.8, utilising a quantised interval size of 1 (*i.e.*, $\Delta = 1$), we see overall good agreement, except in the region $\mathcal{H} \to 0$ for FGN. Note that in the case of ARFIMA that $\mathcal{H} = d + \frac{1}{2}$. In all cases the NPD-Entropy estimator was better at picking up the underlying trends of the entropy rate functions for strongly correlated processes than the measures adapted from Shannon entropy in Section 6.1, but specific entropy rate produced closer estimates of the differential entropy rate at a large cost in computation time.

There are a few places where errors can be introduced in practical estimation using this estimation technique. The limit theorems and the results on the inherited properties, from Section 6.3, of the estimator hold as $\Delta \to 0$, which for practical purposes we are unable to achieve. However, this leaves us with trying to understand the possible errors that have been introduced. Potential sources of estimation error are from the the quantised Shannon entropy rate estimate, and the approximation of the integral for the differential entropy rate estimate, *i.e.*, as $\sum f(x)\Delta \to \int f(x)$, due to the result holding in the limit. As the quantisation window decreases, for $\Delta < 1$ and as $\Delta \to 0$, small differences from the true value and estimated values can be enlarged by the adjustment to the Shannon entropy rate estimate of $\log(\Delta)$. To test the deviation from true value, as a function of the quantisation size, we produced estimates for a range of different quantisation window sizes, $\frac{1}{3}, \frac{1}{2}, 1, 2, 3$, and across the range of the Hurst parameter, $[0, 1]$. Given that we have standardised all of the process variances to be $\sigma^2 = 1$, the window size is in direct proportion with the variance. This means that small fluctuations of the process will produce the next realisation within the same bin, and therefore the measured uncertainty will come from larger shifts. In general we would tailor the window size to the variance, since we want to balance the number of bins and the expected size of movements between random variables of the process.

The results for FGN are shown in Figure 6.9. The most accurate of these across the whole range is at $\Delta = 1$, which is surprising as we should be more accurate as the quantisation size decreases. However, we are balancing a few different sources of error and in this case the best result is to quantise at a window size of 1. As the Hurst parameter tends to 1, the finer quantisations become more accurate estimators This is because as the correlations become stronger the variation in the process will be more subtle, and the process appears more smooth, hence the smaller quantisation windows are required to estimate the uncertainty.

The results are similar for ARFIMA(0,d,0) processes, with the closest estimate, across the entire parameter range, to the true value of the entropy rate being when the quantisation window was of size, $\Delta = 1$. For the rest of the quantisation sizes the estimates got worse as they got further away from 1, i.e. the next closest were $\Delta = \frac{1}{2}$ and 2. The finer quantisations, $\Delta = \frac{1}{2}, \frac{1}{3}$ became more accurate as $\mathcal{H}$ increased, since the changes in the time series occur on a much smaller scale the positive correlations increase. For ARFIMA(0,d,0) processes, the finer quantisations are better able to capture the divergence towards $-\infty$ as $\mathcal{H} \to 1$. Similar to FGN, the finer quantisations are required to pick up the deviations as $\mathcal{H} \to 1$. We recommend that $\Delta = 1$ is used in general, and decreasing the window size as the Hurst parameter tends to 1, to capture the smaller variations in the process as the

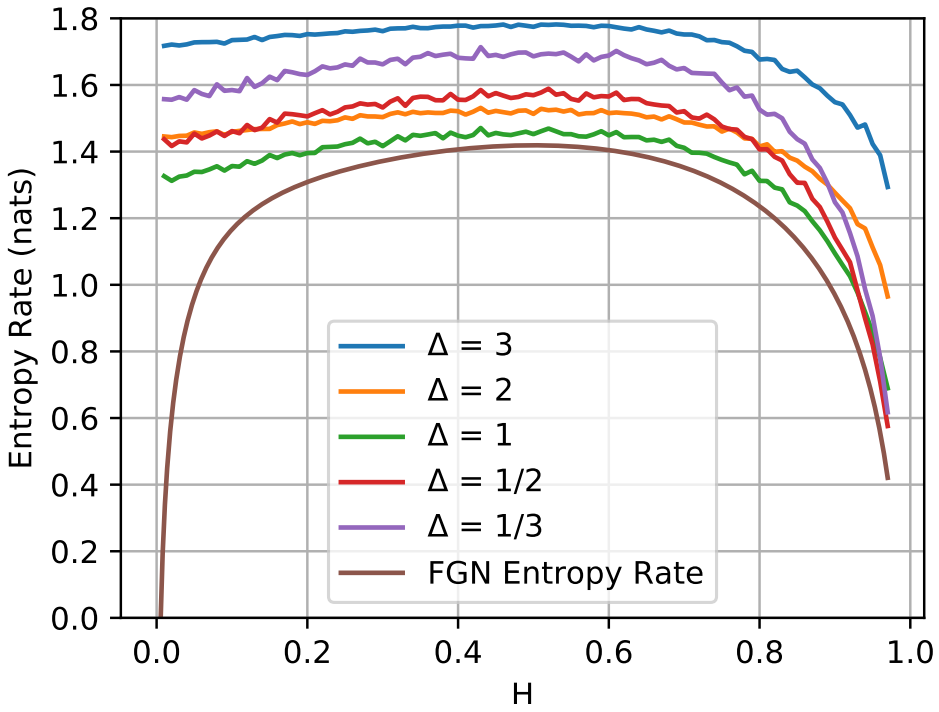

Figure 6.9: Comparison of the NPD-Entropy estimates and true values of differential entropy rate of FGN with $\Delta = \frac{1}{3}, \frac{1}{2}, 1, 2, 3$. The window size of one has the best estimates, however as the Hurst parameter tends to 1 the finer quantisations, *i.e.*, $\frac{1}{3}$ and $\frac{1}{2}$, approximate the asymptotic decrease better.

autocorrelation of the time series increases to 1.

6.4.1 Robustness of estimator to non-stationarity

In statistics and probability theory, a common assumption that is made to enable analysis of stochastic processes is stationarity, meaning that the probability distribution is time invariant over the stochastic processes, see Definition 2.1.8.

In reality, many real world processes are not exactly stationary, hence we need to understand how robust the quantised estimation technique is to non-stationarity. In the previous section we have shown that the estimator has good performance in the cases of LRD processes. However, we would like to test the performance against processes that have a convergent entropy

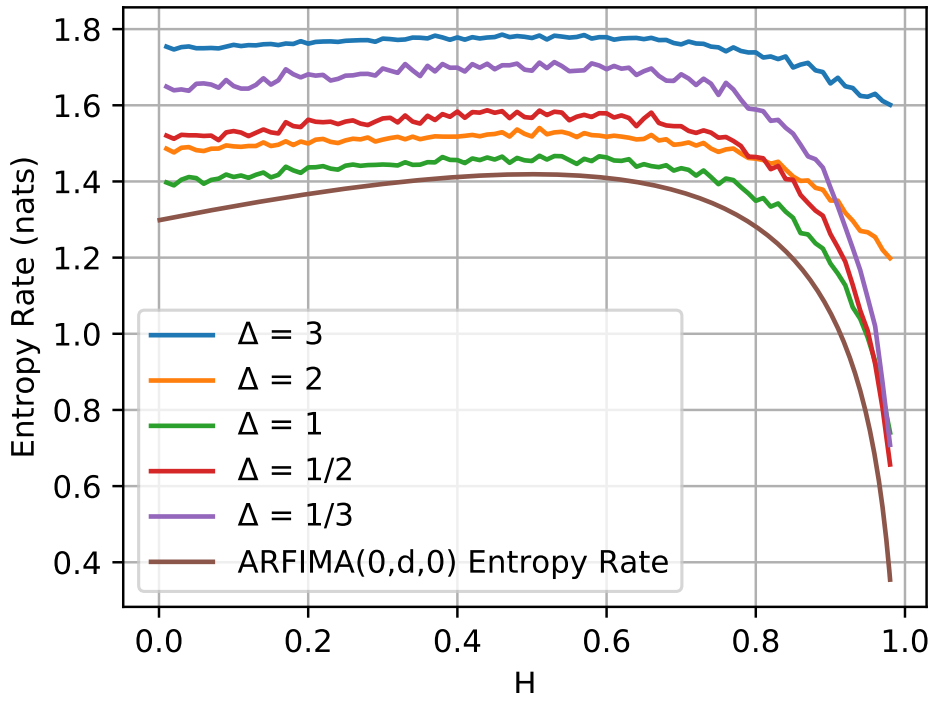

Figure 6.10: Comparison of the NPD-Entropy estimates and true values of differential entropy rate of ARFIMA(0,d,0) with $\Delta = \frac{1}{3}, \frac{1}{2}, 1, 2, 3$. Similar to FGN, the window size of 1 provides the best estimates over the range of Hurst parameter values, however we also see that the finer quantisations become more accurate as $\mathcal{H}$ increases towards 1.

rate, but have moment statistics that vary with time. In this section, we will analyse some well defined non-stationary processes, with an entropy rate that converges, and use the estimators to make estimates of their entropy rates, and compare to the true value. We test two different approaches, the first being a process with a non-stationary mean that alternates deterministically with a common variance, and a random walk with increasing variance.

The initial test for a departure from stationarity is by considering a mean-shifting process. We define this as $X_n = \mu_n + \epsilon_n$, where $\epsilon_n \sim \mathcal{N}(0, 1)$ and independent and we use the example where μ_n alternates periodically between 0 and 1 every 100 samples. The mean shifts are predictable so the entropy rate is

$$h\left(\mu_n + \epsilon_n\right) = h\left(\epsilon_n\right)$$

$$= \frac{1}{2}\log(2\pi e) \approx 1.419,$$

since there is no uncertainty in the mean. The changing mean makes accurate estimation of differential entropy rate more difficult because all of the estimators we are aware of assume stationarity.

We test this process with 50 realisations of 2000 data points. The mean and variance of the estimates are shown in Table 6.2. NPD-Entropy, with a choice of parameter $\Delta = 1$, estimated the entropy rate to be 1.554, which overestimates the true value by a small amount, similar to specific entropy rate. Sample and permutation entropy both overestimate the true value, with permutation entropy having an extremely low variance, close to its maximum value, $\log(3!) \approx 2.585$. Approximate entropy was a large underestimate, similar to the behaviour with $m = 3$ for both the ARFIMA(0,d,0) and FGN estimates.

The second test for robustness to a departure from non-stationarity is to consider a simple model that has a stationary mean, but varying second order statistics, a Gaussian Random Walk, $\{Z_n\}_{n\in\mathbb{N}}$. This model is a discrete time stochastic process, where each step $X_i \sim \mathcal{N}(0,1)$ and independent and we consider the sum of the steps $Z_n = \sum_{i=1}^{n} X_i$. The entropy rate can be derived by considering the limiting conditional entropy to give $h(Z_n) = h(X_n) \approx 1.419$ once again. However, the process has non-stationary second moment $\mathrm{Var}(Z_n) = n$.

As in the previous test, 50 realisations of the model were made and then estimates were made and the mean and variance calculated. Most of the estimators exhibited behaviour similar to the mean-shift process, but the NPD-estimator performed worse.

The usage of specific entropy rate depends on estimator of the conditional entropy. We used Darmon's [47] implementation in R, using kernel density estimation to estimate the conditional entropy. These initial results suggest that specific entropy rate is relatively robust, but its alternatives are not. However NPD-Entropy is somewhat robust to mean shifts, if not to changes in variance. The variance of the Gaussian walk grows proportionally to the number of random samples of the process, as opposed to the mean shifting process which has constant variance of the normally distributed noise. This may explain the difficulty of estimating the true value, as all of the bins are likely to have small numbers of observations, unless large amounts of data points are observed. Therefore, we have to be careful in how processes depart from stationarity, to understand how closely we can estimate the true entropy rate.

Estimation Technique	Mean Shift Mean	Mean Shift Variance	Gaussian Walk Mean	Gaussian Walk Variance
Approximate Entropy	0.482	0.0008	0.099	0.0007
Sample Entropy	2.263	0.015	2.336	0.205
Permutation Entropy	2.582	0.000005	2.497	0.0002
Specific Entropy Rate	1.478	0.0005	1.523	0.0007
NPD-Entropy	1.554	0.002	2.385	0.083

Table 6.2: Differential entropy rate estimates applied to two non-stationary processes: a Gaussian mean shift process and a Gaussian walk, both with entropy rates 1.419. Each process was simulated 50 times. Sample entropy and permutation entropy both provide large overestimates of the true entropy, and approximate entropy provides large underestimates. NPD-entropy behaves only slightly worse than specific entropy rate on the mean-shift process, but somewhat worse on the Gaussian walk. Estimator variances are included to inform about the relative size of the errors.

6.4.2 Complexity Analysis of Estimation Techniques

The worst-case asymptotic time complexities of the estimators tested here are shown in Table 6.1. Approximate entropy has two loops, first to calculate every $C_i^m(r)$, the number of strings that exceed a threshold, by fixing one of the substrings, then a second loop over all of the $C_i^m(r)$'s, *i.e.*, for a pairwise comparison of all substrings. Hence, the time complexity is approximately $O(N^2)$, where N is the length of data. Sample entropy requires two separate loops, which run over the substrings of order m and $m+1$, all calculating the number of pairs that exceed a threshold, *i.e.,* two loops of order N^2, which also results in a pairwise comparison and hence the time complexity is $O(N^2)$. Permutation entropy takes the relative frequency of each possible permutation, of order n, of which there exist $n!$ permutations. This is performed across the entire data set, and hence the worst-case time complexity for permutation entropy is $O(n!N)$, and hence is extremely sensitive to the choice of order n. With appropriate choices of order this can be used as a quick measure of complexity of time series, however it grows extremely quickly with the order of permutation and hence there is a trade off here between how much data is available with what order can be selected for this technique.

Specific entropy rate's complexity depends on the choice of estimator of the conditional entropy. We used Darmon's [47] implementation in R which used the package `np` for nonparametric kernel density estimation [47], to

estimate the conditional entropy. The technique calculates the bandwidth parameters and then calculates the product and univariate kernel estimates for windows of length p. Hence, to make an estimate we do N kernel density evaluations, and therefore Np for the calculation of the estimate of the joint density per window, and hence with a loop over all of the windows we get an approximate complexity of $O(N^2p)$. Although this does not appear substantially worse that the other estimators, the actual amount of computation is dramatically larger.

NPD Entropy leverages Shannon entropy rate estimators. In our case we use the Grassberger [79] estimate, which utilises string-matching based on Lempel-Ziv algorithm [186], which has a complexity of $O(N \log(N)$.

Complexity estimates are useful to understand scaling properties, but the actual computation times for finite data can be dominated by non-asymptotic terms, and so we test performance directly using 1000 estimates of $N = 1000$ data points. Tests were performed on a 2.4GHz Intel Core i5 processor with 8GB of RAM, running MacOS 10.14.6. The results are shown in Table 6.1. Approximate entropy was much slower than sample entropy, with permutation entropy being the fastest by far. However NPD-Entropy is faster than all but permutation entropy and runs extremely quickly compared to specific entropy rate. The specific entropy rate was considerably slower than all the measures with the 1000 estimates being in the order of 6 days.

Given the time complexity, we suggest using NPD-Entropy as an efficient way of classifying complexity of a continuous-valued system. NPD-Entropy provides a more accurate and robust measurement with comparable or better computation times than most alternatives. Specific entropy rate is superior for accurate measures, but computationally prohibitive except where accuracy is the only criteria.

6.5 Conclusion

We have defined a new technique for the nonparametric estimation of the differential entropy rate. We made an explicit link between the differential entropy rate and the Shannon entropy rate of the associated quantised process. This forms the basis of the estimation technique, NPD Entropy, by quantising the continuous data and utilising the existing theory of Shannon entropy rate estimation. We have shown that this estimation technique inherits statistical properties from the Shannon entropy rate estimator and performs better than other differential entropy rate estimators in the presence of strongly correlated data.

In addition, we have investigated the robustness to non-stationarity of the

estimation technique, and provided over-estimates of the entropy rate. More research is required to analyse the robustness to non-stationarity and understand if the influence is similar to the Shannon entropy rate case. Finally, we have demonstrated the utility of NPD Entropy and shown that it can be run quickly and efficiently to make decent estimates if used in an online mode.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

The development and analysis of LRD model has been of large interest over the previous few decades. This surge in research activity was inspired by discoveries by Harold Hurst, showing that real-world phenomena, in particular the flood levels of the Nile river can have strong correlations with the past. Models of LRD phenomena that do not account for this behaviour can be poor representations of the real system.

In this thesis, we introduce a new perspective on LRD, by characterising the properties of an information theoretic measure, the entropy rate, by the slow convergence and by the infinite sum of a quantity related to the convergence. Then we consider the robust estimation of information theoretic quantities for data generated by LRD processes, given the impact of the entropy rate characterisation. Intuitively, there should be more shared information between the past and future values of an LRD process, due to the strong correlations with the past. This point of view has been considered by a few authors, with a suggestion by Li [120] to define the boundary between short and long memory by the finiteness of the mutual information between past and future. We show that for Markov chains with power-law tail of the return time random variable that Li's idea was correct, however in the case of Gaussian processes that CSRD processes also have infinite mutual information between past and future. Confirming the intuition that there is an infinite amount of information shared between the past and future, and therefore has slow convergence rates to the entropy rate from the conditional entropy. Given the difficulties of quantifying the entropy rate of LRD processes, we investigated the robust estimation of the entropy rate, and find that this task is difficult to perform quickly and accurately.

In Chapter 3 we analysed the entropy rate function, as a function of the Hurst parameter, for LRD Gaussian processes. Utilising two common models, FGN and ARFIMA, we illustrated the behaviour of the entropy rate function, and show that these two processes behave differently in their information theoretic properties, despite being closely related in the time and frequency domain. We proved that two information theoretic measures, the excess entropy and mutual information between past and future are equal for Gaussian processes, which we then used to characterise LRD and CSRD Gaussian processes as those where the measures are infinite. Thus we showed that LRD and CSRD Gaussian processes are characterised by an infinite amount of shared information between and past and future. Which implies that these processes have a "slower" convergence rate as these quantities are related to the convergence of conditional entropy to the entropy rate.

Continuing to characterise the behaviour of LRD, we turned our attention to Markov chains defined on discrete state spaces in Chapter 4. We showed that the convergence of the conditional entropy to the entropy rate is at the same rate as the convergence of the n-step transition probabilities to the stationary distribution, thus allowing us to phrase this as a Markov chain mixing time problem. As a consequence, we proved that the mutual information between past and future is once again infinite for LRD Markov chains, providing the same characterisation of LRD as the previous chapter on Gaussian processes. Finally, we utilised the literature on Markov chain mixing to show that LRD Markov chains do in fact have slower convergence, and this depends on the Hurst parameter, and therefore the strength of positive correlations.

This work indicates that may be able to relate strong positive and negative correlations on general stochastic process, by information theoretic properties. This raises the natural questions such as "Are there more classes of stochastic processes for which LRD and CSRD is equivalent to the mutual information between past and future being infinite?". In addition, we have discovered that the convergence of the conditional entropy to the entropy rate, via observing the process, has slower convergence for LRD processes. Thus identifying another type of quantity whose convergence rate is impacted by LRD. This result may hold in full generality for LRD processes, but more research is required for the potential characterisation.

We shifted our attention to the question of estimation in Chapters 5 and 6. In Chapter 5 we provided a review of the state-of-the-art of the estimation of the entropy rate for both Shannon and differential entropy rate. Providing in-depth analysis of both parametric and non-parametric techniques, and highlighting the relatively fewer number of techniques for the estimation of differential entropy rate.

Continuing on the discussion of entropy rate estimation in Chapter 6, we illustrated the difficulties and issues with the estimation of entropy rate for LRD processes. Using the current state of the art estimation techniques we find that the current techniques either have issues with accuracy or computational complexity. We developed a technique, NPD Entropy, to estimate the differential entropy rate that uses Shannon entropy estimation on quantised data, motivated by a theorem we developed which connects the differential entropy rate and the Shannon entropy rate of a quantised process. This technique provides a good balance between the accuracy and computation time, and the differential entropy rate estimator is able to inherit important properties of the Shannon entropy rate estimator.

The entropy rate is used as a measure of complexity or uncertainty of a stochastic process. Therefore, its robust estimation is useful in many applied settings as a way of classifying the complexity of a data source. By utilising our technique, differential entropy rate estimation is translated to a Shannon entropy rate estimation problem, for which many estimators have been developed from limit theorems, with elegant estimation properties.

In this thesis, we have demonstrated that there are useful characterisations and insights that can be gained from approaching LRD from the perspective of information theoretic measures. These characterisations can be used in applied contexts to provide understanding of LRD in real-phenomena, and then have been used to form estimation techniques that robust to the influence of LRD.

7.2 Future Work

There are a few extensions and generalisations to the work in this thesis, to both the characterisation and estimation of the entropy rate for LRD processes.

From the initial discoveries in the characterisation of the behaviour of the entropy rate for LRD Gaussian processes and Markov chains, this begs the question of whether the behaviour can be generalised to more classes of stochastic processes. For continuous-valued processes, the proof of the characterisation of the convergence rate of the conditional entropy to the entropy rate relies on a characterisation of the entropy rate for Gaussian processes, therefore new techniques would be required to answer this question. More general techniques may be able to make advances on whether the mutual information between past and future is infinite for LRD and CSRD processes, from the information theoretic perspective.

Another related question, alluded to in Chapter 3, is finding which class

of processes have maximum entropy under the constraint of power-law decaying covariances. From previous results for AR and ARMA processes, and Figure 3.6 it appears likely that this would be the ARFIMA class. This has applications to the theory of estimation, and model selection.

As has been noted earlier in the thesis, LRD on discrete state spaces is much less studied. Some obvious generalisations of the analysis presented for Markov chains, could apply to semi-Markov chains, and potentially to more general point processes defined on discrete spaces. There are some additional difficulties that arise in the analysis of these more general processes, instead of transitioning state at each time step, the time between transitions follows a probability distribution. However, this behaviour means that the LRD behaviour can be driven by the infinite second moment of the distribution of time between transitions, in addition to the infinite second moment of the return time to the same state as in the Markov chain case. This means that LRD can occur on a finite state space, in contrast to Markov chains that require infinite states for the infinite second moment of the return time. The interaction of these two causes of LRD behaviour on discrete state spaces may provide some interesting research avenues.

There are improvements that can be made to the differential entropy rate estimator, NPD Entropy. Analysis to understand if other estimation properties of interest can be inherited from a chosen Shannon entropy rate estimation technique. Accurate quantification of the error between the estimated and true value would be very useful, in being able to generate confidence intervals and to quantify the uncertainty in the estimate. Following this, further understanding of the relationship between the errors and improved methodology for generating bin sizes, to achieve the theoretical asymptotic properties. Outside of NPD Entropy, continuing the development of robust differential entropy rate estimation is valuable, as we have shown that no techniques provide robustness to non-stationarity, given the underlying assumptions of stationarity in many techniques. Therefore, developing estimators robust to real-world data would be valuable, which could be used to understand the complexity and uncertainty in many real-world data generating sources, and provide insight on real-world phenomena.

Appendix A

NPD Entropy Estimator Package

We developed an implementation of the estimator defined in Chapter 6, NPD Entropy, in Python. This can be found on GitHub at the following url, https://github.com/afeutrill/npd_entropy. In this appendix we will briefly describe the estimation code and its function.

A.1 Installation

The code can be installed in two ways, via cloning the GitHub repository or more simply through pip. From the command line, input this command.

```
pip install npd_entropy
```

A.2 Functionality

As described above, the code implements the NPD Entropy estimator as defined in Chapter 6. The estimation technique was originally defined in Grassberger [79], with the description of the technique given in Chapter 5.

Given the technique that we have developed it requires a quantisation of the input sequence of continuous-valued data. We provide three different quantisation approaches, these are:

- partitioning of the data into equally spaced bins, equally spaced counting up from the origin,
- partitioning of the data into equally spaced bins, equally spaced centred at the origin,
- partitioning the data into bins that have the same amount of probability mass, assuming a normally distributed random variable.

The code for the respective quantisations is given in the block below.

```
import numpy as np
from numba import jit, prange
from scipy.special import erfinv

def quantise_series(series, delta):
"""
Returns a quantised version of a continuous-valued, discrete-time
    process. This takes the floor to the nearest increment of 1/res
Parameters:
series - numpy array: A time series of continuous valued variables
delta - float: the bin width of the quantisation
Returns:
Numpy array of a discrete quantised version of the series
"""
series = np.array(series)
return np.floor(series*1/delta) * delta

def quantise_series_around_zero(series, delta):
"""
Returns a quantised version of a continuous-valued, discrete-time
    process. Quantises centred around zero.
Parameters:
series - numpy array: A time series of continuous valued variables
delta - float: the bin width of the quantisation
Returns:
Numpy array of a discrete quantised version of the series
"""
series = np.array(series)
return np.rint(series*1/delta) * delta

def quantise_series_normal_quantiles(series, num_quantiles):
"""
Returns a quantised version of a continuous-valued, discrete-time
    process. This returns discrete quantiles of the same probability
    mass, according to a normal distribution
Parameters:
series - numpy array: A time series of continuous valued variables
num_quantiles - int: The number of possible quantiles to use as bins
    for the quantisation
```

```
    Returns:
    Numpy array of a discrete quantised version of the series
    """
    series = np.array(series)
    bins = np.array([np.sqrt(2) * erfinv(2 * (i/num_quantiles) - 1) for i
        in range(num_quantiles+1)])
    return np.digitize(series, bins)
```

All of these techniques return a discrete version of the original data.

To calculate the entropy rate estimate we pass the quantised series into the estimator. The implementation of the estimator is given by the following code, which performs string matching across many offset sub-sequences of the quantised data. Then the Shannon entropy rate estimator is returned for the quantised data.

```
    @jit(nopython = True)
    def grassberger_estimate(series, n):
    """
    Implementation of the estimator described in P. Grassberger "
        Estimating the information content of symbol sequences and
        efficient codes", IEEE Transactions on Information Theory 1989.
    Parameters:
    series - numpy array: A discrete-valued time series.
    n - int: Number of past values to start for string matching.
    Returns:
    Shannon entropy estimate of the discrete-valued time series
    """
    L_num_array = np.zeros((n,))
    for i in prange(n):
    L_num = 0
    for num in np.arange(n):
    for j in np.arange(1, n-num+1):
    # Comparing the series to back in time only
    if np.all(series[n+i:n+i+num] == series[n+i-j-num:n+i-j]):
    break
    # If the end of the loop with no matches, then set L_num to the
        previous length
    if j == n - num:
    L_num = num - 1
    # Tests whether non-match has occurred, then exits
    if L_num > 0:
    break
```

```
    L_num_array[i] = L_num
    reciprocal = 1/(n * np.log(n)) * L_num_array.sum()
    H_est = 1/reciprocal
    return H_est
```

Given the Shannon entropy rate estimate we then calculate the differential entropy rate estimate by adding the $\log \Delta$ term. The NPD Entropy estimate is given by the following function.

```
def npd_entropy(H_est, delta):
"""
Generic implementation of NPD Entropy, which calculates a differential
    entropy rate estimate, h_est, given a Shannon entropy rate
    estimate, H_est.
Parameters:
H_est - float: Shannon entropy rate estimate, from any technique
delta - float: the bin width of the quantisation
Returns:
Differential entropy rate estimate of a quantised estimate
"""
return H_est + np.log(delta)
```

A.3 Usage

An example of the usage of the functions to quantise the series is given by the following command, after importing the npd-entropy package.

```
import npd_entropy
quantised_series = npd_entropy.quantise_series(series, 1)
shannon_entropy_rate_estimate = npd_entropy.grassberger_estimate(
    quantised_series, 100)
npd_entropy.npd_entropy(shannon_entropy_rate_estimate, 1)
```

The first command here, returns the quantised version of the series. This result is then given to the Shannon entropy rate estimator. Note that for string matching that we need to consider some history to form the string matches, that is greater than $\log n$ long. The final part of the estimation is adding the term to correct between Shannon entropy rate of the quantised data and the differential entropy rate of the original data.

Appendix B

Entropy

This appendix provides additional definitions and results related to the concept of entropy, and in particular, differential entropy, where these definitions and results are direct analogues of the Shannon entropy case. We start with some a discussion of expanded results in differential entropy.

B.1 Differential Entropy

After the definition of differential entropy in Definition 2.1.6 we define the joint differential entropy for a collection of continuous random variables as a natural extension to the multivariate case,

Definition B.1.1. *The joint differential entropy of a collection of continuous random variables,* $X_1, ..., X_n$*, with support on,* $\Omega_1 \times ... \times \Omega_n$*, with a joint density function,* $f(x_1, ..., x_n) = f(\mathbf{x})$*, is*

$$h(X_1, ..., X_n) = \int_{\Omega_1} ... \int_{\Omega_n} f(\mathbf{x}) \log f(\mathbf{x}) d\mathbf{x}.$$

Like differential entropy, the joint differential entropy can also be negative, by generalising the above example to n dimensions, *e.g.*, take the uniform distribution on an n-dimensional subset of $\mathbb{R}^n$ and take the limit as $a \to 0$. An example is the joint density of the multivariate Gaussian distribution, with density function

$$f(\mathbf{x}) = \frac{1}{\left(\left(\sqrt{2\pi}\right)^n |\Sigma|^{\frac{1}{2}}\right)} e^{-\frac{1}{2}\mathbf{x}^T \Sigma^{-1} \mathbf{x}},$$

where Σ is the covariance matrix and with zero mean for every dimension. Which has the joint differential entropy

$$\begin{aligned} h(X_1, ..., X_n) &= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(\mathbf{x}) \log f(\mathbf{x}) d\mathbf{x}, \\ &= \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(\mathbf{x}) \left(-\frac{1}{2} \log \left((2\pi)^n |\Sigma| \right) - \frac{\mathbf{x}^T \Sigma^{-1} \mathbf{x}}{2} \right) d\mathbf{x}, \\ &= -\frac{1}{2} \log \left((2\pi)^n |\Sigma| \right) - \frac{E[\mathbf{X}^T \Sigma^{-1} \mathbf{X}]}{2}. \end{aligned}$$

Then we can manipulate the last term, $\mathbf{X}^T \Sigma^{-1} \mathbf{X} = tr\left(\mathbf{X}^T \Sigma^{-1} \mathbf{X}\right)$, since it is a 1×1 matrix, where $tr(A)$ is the trace function of a matrix A. This implies

$$\begin{aligned} E[\mathbf{X}^T \Sigma^{-1} \mathbf{X}] &= E\left[tr\left(\mathbf{X}^T \Sigma^{-1} \mathbf{X}\right) \right], \\ &= tr\left(E[\mathbf{X}^T \mathbf{X}] \Sigma^{-1} \right), \\ &= tr(I) = n, \end{aligned}$$

since $E[X^T X] = \Sigma$ for a mean zero process and where the second equality is given because of the linearity of expectation. Therefore the entropy rate of the multivariate Gaussian distribution is

$$h(X_1, ..., X_n) = \frac{1}{2} \log \left((2\pi e)^n |\Sigma| \right).$$

We define the conditional entropy in a similar manner as the joint entropy, as the expectation of $-\log f(x|y)$.

Definition B.1.2. *The conditional differential entropy, $h(X|Y)$ of two random variables X and Y, with a joint density $f(x, y)$, is defined as*

$$h(X|Y) = -\int_{\Omega_1} \int_{\Omega_2} f(x, y) \log f(x|y) dx dy.$$

A theorem of interest that links these two concepts is the chain rule for differential entropy, which is a direct analogue of the Shannon entropy case. This theorem is very useful as it gives approaches to analyse the joint entropy given information of the conditionals, which are estimated from observed data.

Theorem B.1.1 (Chain rule of Differential entropy [41, pg. 253])**.**

$$h(X_1, \dots, X_n) = \sum_{i=1}^{n} h(X_i | X_{i-1}, \dots, X_1).$$

Relative measures are also defined as for Shannon entropy, these serve the same function for differential entropy, quantifying the difference between probability density functions. We define the relative entropy, or Kullback-Leibler divergence, for differential entropy.

Definition B.1.3. *The relative entropy, $D(f(x)||g(x))$, of two probability density functions $f(x)$ and $g(x)$ is defined as*

$$D(f(x)||g(x)) = \int_{\Omega} f(x) \log \left(\frac{f(x)}{g(x)} \right) dx.$$

Note that this has the same properties as the discrete version, and are consistent with the discussion after Definition 2.1.4. Even though differential entropy may be negative, the relative entropy is non-negative, with equality if and only if $f = g$ almost everywhere [41, Theorem 8.6.1, pg. 252]. The density is finite when the support of f is completely contained in the support of g. We now define the mutual information for two random variables. The fact that the relative entropy is non-negative has resulted in a lot of authors to use this as a measure of complexity or uncertainty, instead of differential entropy. However, it is a relative measure, and needs to be quantified with respect to another probability distribution. Therefore, it is often used to quantify the difference between the distribution with the "highest uncertainty" on the support set, *e.g.*, relative to the uniform distribution over a finite interval.

Definition B.1.4. *The mutual information, $I(X;Y)$, between two random variables X and Y, with a joint density $f(x,y)$ is defined as*

$$I(X;Y) = \int_{\Omega_1} \int_{\Omega_2} f(x,y) \log \left(\frac{f(x,y)}{f(x)f(y)} \right) dxdy.$$

Similar to the relative entropy this has the same properties as the discrete case. The quantity is non-negative with equality if and only if $f(x,y) = f(x)f(y)$.

B.2 Differential Entropy Rate

Finally we will introduce the entropy rate for stochastic processes on continuous state spaces, the differential entropy rate. We will quickly present the main definitions for completeness, however, as many of the definitions and theorems are the natural extensions of the concepts defined previously for Shannon entropy we will not present the analogues of simple extensions and results for differential entropy.

Definition B.2.1. *The differential entropy rate for a continuous-valued, discrete-time stochastic process, $\chi = \{X_i\}_{i\in\mathbb{N}}$, is defined as,*

$$h(\chi) = \lim_{n\to\infty} \frac{1}{n} h(X_1, ..., X_n).$$

An example of a process which is non-stationary but has a differential entropy rate is the Gaussian walk, $S = \{S_n\}_{n\in\mathbb{N}}$. This is defined as the process of sums of i.i.d. normally distributed random variables, *i.e.,* $S_n = \sum_{i=1}^{n} X_i$, where $X_i \sim \mathcal{N}(0, \sigma^2)$. The process has mean 0 for all n, however it is non-stationary as the variance depends on the number of steps, n, as

$$\mathrm{Var}(S_n) = \mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = n\sigma^2.$$

However, the entropy rate converges and is equal to

$$\begin{aligned}\lim_{n\to\infty} \frac{h(S_1, ..., S_n)}{n} &= \lim_{n\to\infty} \frac{\sum_{i=1}^{n} h(S_i|S_{i-1}, \ldots, S_1)}{n}, \\ &= \lim_{n\to\infty} \frac{nh(X_i)}{n}, \\ &= h(X_i),\end{aligned}$$

as each random variable X_i is independent.

This example illustrates that the entropy rate of process that is independent of its history is the same as the entropy of a single random variable of the process. This very intuitive as it shows that the history is not considered for sequences of independent random variables. In chapters 3 and 4 we also consider the convergence rate of the conditional entropy, from observing the history, to the entropy rate. By showing that i.i.d. processes achieve their entropy rate from the first random variable, this implies that it has already converged, and therefore that in the i.i.d. case the convergence is instantaneous.

An equivalent characterisation of the differential entropy rate for stationary stochastic processes, using conditional entropy, is given by the following theorem. This is a direct analogue of the Shannon entropy case, and is extended to differential entropy utilising an identical argument [41, pg. 416].

Theorem B.2.1. *For a stationary stochastic process, $\{X_i\}_{i\in\mathbb{N}}$, the differential entropy rate is equal to,*

$$h(\chi) = \lim_{n\to\infty} h(X_n|X_{n-1}, ..., X_1),$$

where the limit exists.

This states that for a stationary stochastic process, the differential entropy rate is the limit of the new information that we get from each new random variable, after observing the infinite past.

Appendix C

Long Range Dependence

In this appendix we provide some additional results that describe and give more context to the phenomenon of LRD.

A related concept to LRD is self-similarity, which involves the scale invariance of a stochastic process. A constant theme throughout LRD is scaling phenomena and power laws, such as seen with the R/S statistic and the sample mean which we discuss below. The self similarity property has been connected to LRD since the development of the theory by Mandelbrot [126], Mandelbrot and Van Ness [127].

Definition C.0.1. *A continuous-time stochastic process, $(X(t))_{t\in T}$ is called self similar if there exists a $\mathcal{H} > 0$, such that*

$$(X(ct)) \stackrel{D}{=} \left(c^{\mathcal{H}} X(t)\right).$$

For the relevance to LRD, the examples we will introduce have self-similarity with the Hurst parameter.

FBM exhibits self-similarity, since if we consider the process, $B^{\mathcal{H}}(ct)$, for $c > 0$ we see

$$\begin{aligned} E[B_{\mathcal{H}}(ct)B_{\mathcal{H}}(cs)] &= \frac{1}{2}\left((ct)^{2\mathcal{H}} + (cs)^{2\mathcal{H}} - |ct - cs|^{2\mathcal{H}}\right), \\ &= c^{2\mathcal{H}}\left(\frac{1}{2}\left(t^{2\mathcal{H}} + s^{2\mathcal{H}} - |t - s|^{2\mathcal{H}}\right)\right). \end{aligned}$$

This is equal to the covariance function of the process $c^{\mathcal{H}}B^{\mathcal{H}}(t)$, and FBM is self-similar with parameter $\mathcal{H}$.

There are strong links between the concepts of power-law tails of probability distributions and the phenomenon of LRD. In addition to the asymptotic decay of the covariance function, other quantities of interest in LRD processes

also feature asymptotic power laws, *e.g.*, partial sum variance [14, Theorem 2.2] and variance of sample means [13]. Mandelbrot in his early investigations into LRD, named two related phenomena the Noah effect, after extremely heavy tailed events, *e.g*, extreme rainfall, and the Joseph effect, after long periods of correlated events, *e.g.*, periods of correlated high or low flood levels in the Nile River [128]. These phenomena are linked fundamentally through their scaling, where the Noah effect is scaling in the spatial domain, and the Joseph effect is scaling in the temporal domain [80]. Given these links the work of Mandelbrot focused on a developing models to explain both of these phenomena, and resulted in classes of models that had these effects which had both Gaussian and non-Gaussian marginal distributions, of which the non-Gaussian marginals had the property that the probability distributions scaled as a power law, that is for a density function, $f(x)$,

$$\begin{aligned} & f(x) \sim x^{-\alpha}, \\ \implies & f(cx) \sim c^{-\alpha} x^{-\alpha}. \end{aligned}$$

In particular, the power-laws to model the extreme events have parameterisations where the variance is infinite, *i.e.,* $E\left[(X-\mu)^2\right] = \infty$. This approach was inspired by the results from errors in telecommunications networks by Berger and Mandelbrot [16], and Mandelbrot [126]. Although the links between the two phenomena are very strong, and power-laws will be a consistent theme throughout the thesis, it was shown that infinite variance power-law tails alone was not enough to induce LRD behaviour, as some examples of processes with infinite variance having the growth of R/S statistic at the rate of $n^{\frac{1}{2}}$ [151, pg. 180].

A related asymptotic property that involves power law tails is called regular variation, which comes up consistently in this area of probability.

Definition C.0.2. *A function $L : (0,\infty) \to (0,\infty)$ is called a regularly varying function, if for every $a > 0$*

$$\lim_{x\to\infty} \frac{L(ax)}{L(x)} = g(a) < \infty.$$

Another related concept is called slowly varying, that is defined with the limit above being $g(a) = 1$ for all $a > 0$. Often this comes up in the characterisation of probability distributions or conditions on tails behaving like a power-law of a negative exponent, *e.g.*, $p(x) \sim L(x)x^{-\alpha}$, for a slowly varying function $L(x)$ [38].

Classes of point processes and renewal processes, known as fractal based processes have been shown to exhibit the phenomena of LRD, various examples are defined by Lowen and Teich [124]. These are defined like regular

point processes, such as the Poisson process, with the interarrival distributions given by an infinite variance power-law probability distribution. We present a theorem of Taqqu et al. [161], which connect the two concepts of the Noah effect and the Hurst effect. The heavy tailed source, *i.e.,* Noah effect, is an Alternating (or On-Off) fractal renewal processes, which are a subclass of fractal renewal process on a binary state space, where the time until jumps between the two states is given by a power-law probability distribution. This was inspired by the analysis of ethernet packet data.

Definition C.0.3. *Let $\{W(t)\}_{t\geq 0}$ be a stationary binary valued time series, then*

$$W(t) = 1, \text{ the process is "On" at time } t, \text{ with mean duration } \mu_1$$
$$W(t) = 0, \text{ the process is "Off" at time } t, \text{ with mean duration } \mu_2$$

If the "On" times are i.i.d. and the "Off" times are i.i.d. (not necessarily with the same distribution). Then $\{W(t)\}_{t\geq 0}$ is an alternating renewal process.

We want to take a superposition of these types of processes, where we define the superposition of processes below.

Definition C.0.4. *The superposition of M i.i.d. On/Off Processes $\{W^{(m)}(t)\}_{t\geq 0}$ at time t is defined as:*

$$W_M^*(t) = \sum_{m=1}^{M} W^{(m)}(t)$$

The aggregated count in the interval $[0, Tt]$ is:

$$W_M^*(Tt) = \int_0^{Tt} (\sum_{m=1}^{M} W^{(m)}(u))du.$$

Then we can state the theorem, and point out that the aggregation of a bunch of heavy-tailed infinite variance point processes can create LRD in the limiting process.

Theorem C.0.1 (Theorem 2 [161])**.** *For large M and T, the aggregated count process $\{W_M^*(Tt)\}_{t\geq 0}$, for fractal Alternating On/Off processes, behaves statistically like*

$$TM\frac{\mu_1}{\mu_1+\mu_2}t + T^{\mathcal{H}}\sqrt{L(T)M}\sigma B_{\mathcal{H}}(t)$$

or more precisely,

$$\lim_{T\to\infty}\lim_{M\to\infty}\frac{(W_M^*(Tt)-TM\frac{\mu_1}{\mu_1+\mu_2}t)}{T^{\mathcal{H}}\sqrt{L(T)M}}=\sigma B_{\mathcal{H}}(t),$$

where $L(T)$ is a slowly varying function.

An alternate second order definition for LRD involving the variance of the partial sums, and the growth rate greater than linear, called LRD in terms of Allen variance (LRD-AV) in some literature [86], however we will see that this definition is equivalent to the autocorrelation and spectral density definitions.

Definition C.0.5. *A stochastic process, $\{X_i\}_{i\in\mathbb{Z}^+}$ is LRD-AV if the sequence of partial sums $S_n=\sum_{i=0}^{n}X_i$, has the property*

$$\lim_{n\to\infty}\frac{Var(S_n)}{n}=\infty.$$

This definition presents another second-order approach to defining LRD. This is related to the other definitions because the variance of the partial sum can be transformed to a form depending on the covariance function,

$$\begin{aligned}Var(S_n)&=\sum_{i=1}^{n}\sum_{j=1}^{n}Cov(X_i,X_j),\\&=\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma(|i-j|),\\&=n\gamma(0)+2\sum_{i=1}^{n-1}(n-i)\,\gamma(i).\end{aligned}$$

This leads to the following result, after dividing through by n and taking the limit, showing that the definitions are equivalent and is an adaption of Proposition 6.1.1 [151] of Samorodnitsky, where they use the definition of absolute summability for LRD, *i.e.,* $\sum_{k=1}^{\infty}|\gamma(k)|$.

Theorem C.0.2. *The sum $\sum_{k=-\infty}^{\infty}\gamma(k)<\infty$, if and only if*

$$\lim_{n\to\infty}\frac{Var(S_n)}{n}=\gamma(0)+2\sum_{i=1}^{\infty}\gamma(i)<\infty.$$

In the case of the definition of LRD using $\sum_{k=-\infty}^{\infty} |\gamma(k)| = \infty$, the two definitions of LRD-AV and LRD are not equivalent. A stronger condition is required for equivalence, that the autocovariance decays as a power law *i.e.,* equivalent to Definition 2.2.4. An in-depth discussion of different definitions of LRD and their interactions is given in Section 5.2 of [72].

An example of the phase transition behaviour is shown by a link between the scaling of partial sum processes, $S_n(t) = \{S_{[nt]}\}_{t\geq 0}$ of an increment process and the strength of correlations.

Theorem C.0.3 (Proposition 9.2.6 [151]). *Let $X_n = Y(n) - Y(n-1)$ be the increment process of a self-similar process, Y, with stationary increments, with $\mathcal{H} > 0$. Then*

$$\left(n^{-\mathcal{H}} S_n(t), t \geq 0\right) \Rightarrow \left(Y(t), t \geq 0\right), \text{ as } n \to \infty.$$

Which demonstrates the rate of scaling of the process, for self-similarity, varies depending on the strength of the correlations with the past.

Similar phase transition behaviour exists for the partial maxima of processes, which are defined as

$$M_n = \max(X_1, \ldots, X_n), n = 1, 2, \ldots.$$

Phase transitions occur in these statistics when the covariances decay at a rate slower than $(\log n)^{-1}$ and the partial maxima grows as a power. Interestingly, this shows that the phase transition behaviour isn't always consistent with power-law decay of the second-order statistics. Examples of this type of behaviour is described Samorodnitsky [151, 152] and in the case of heavy tailed processes with infinite variance [138].

We aim in this thesis to demonstrate that similar phase transition behaviour occurs for LRD processes with respect to the behaviour of the convergence of the conditional entropy to the entropy rate, and finiteness of the mutual information between past and future.

Finally we discuss the problem of distinguishing LRD from that of non-stationarity. Since we are talking about long term local trends of stationary processes it can be difficult using ordinary statistical analysis techniques to determine whether we are analysing data that is stationary or not. A cautionary tale along this line is presented by Samorodnitsky [151, pg. 183] from a paper by Bhattacharya et. al. [17], which details a non-stationary process that behaves like a process with a Hurst parameter greater than $1/2$ indicating that it's an LRD process. We define a non-stationary process

$$X_n = Y_n + (a+n)^{-\beta}, i = 1, 2, \ldots,$$

where $a \geq 0$, $0 < \beta < \frac{1}{2}$, and the random variables Y_n are i.i.d. with finite variance, σ^2. This defines a non-stationary model as the final term is a drift term that decreases in magnitude as $n \to \infty$, however the model becomes stationary in the limit. We will consider the range statistics, *i.e.,* the numerator of the R/S statistic separately for the first and second terms of X_n above. Let

$$r_n = \max_{0 \leq i \leq n} \left(s_i - \frac{i}{n} s_n \right) - \min_{0 \leq i \leq n} \left(s_i - \frac{i}{n} s_n \right), \text{ where } s_m = \sum_{j=1}^{m} (a+j)^{-\beta},$$

$$R_n = \max_{0 \leq i \leq n} \left(T_i - \frac{i}{n} T_n \right) - \min_{0 \leq i \leq n} \left(T_i - \frac{i}{n} T_n \right), \text{ where } T_m = \sum_{j=1}^{m} Y_m.$$

Considering the second term of r_n, we can see that it is 0, since it is a decreasing sequence of positive numbers. Now to get an asymptotic value for the first term, first we can apply Theorem 10.5.6 of Samorodnitsky [151], to s_n which yields

$$s_n \sim \frac{1}{1-\beta} n^{1-\beta}, \text{ as } n \to \infty.$$

Then if we take the maximum value over all of the indices i, in the first term of r_n, denoted i^*, we get the following by substitution and simplifying,

$$\begin{aligned} i^* &= \left\lfloor \left(\frac{s_n}{n} \right)^{-\frac{1}{\beta}} - a \right\rfloor, \\ &\sim (1-\beta)^{\frac{1}{\beta}} n, \text{ as } n \to \infty. \end{aligned}$$

Which after combining and solving gives

$$\begin{aligned} \max_{0 \leq i \leq n} \left(s_i - \frac{i}{n} s_n \right) &= s_{i^*} - \frac{i^*}{n} s_n, \\ &\sim \beta \, (1-\beta)^{\frac{1}{\beta}-2} \, n^{1-\beta}. \end{aligned}$$

Since the Y_n's are i.i.d. with finite variance, the range of the first n observations, R_n, grows as $n^{\frac{1}{2}}$.

For the process

$$r_n - R_n \geq \max_{0 \leq i \leq n} \left(S_i - \frac{i}{n} S_n \right) - \min_{0 \leq i \leq n} \left(S_i - \frac{i}{n} S_n \right) \geq r_n + R_n.$$

Therefore in the limit as $n \to \infty$

$$\max_{0 \leq i \leq n} \left(S_i - \frac{i}{n} S_n \right) - \min_{0 \leq i \leq n} \left(S_i - \frac{i}{n} S_n \right) \sim \beta \, (1-\beta)^{\frac{1}{\beta}-2} \, n^{1-\beta}.$$

A similar bounding argument on the denominator of the R/S statistic shows that

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \frac{S_n}{n}\right)} \to \sigma.$$

Combining all of this we can see that

$$\frac{1}{n^{1-\beta}}\frac{R}{S}\left(X_1, \ldots, X_n\right) \to \frac{\beta\left(1-\beta\right)^{\frac{1}{\beta}-2}}{\sigma}.$$

From this example we can see that the R/S statistic grows at the rate of $n^{1-\beta}$, which is the same as LRD processes with $\mathcal{H} = 1 - \beta$. There are other classes of non-stationary models models that can have this property. Regime switching models have also shown that they can replicate the characteristics of LRD models, such as Diebold and Inoue [53], who showed that only small regime changes are able to induce partial sum growth at the same rate as LRD processes. Markov switching models have also shown this ability, matching the results of scaling with a positively correlated Hurst parameter and with good agreement to real economic data [123].

Not all perspectives on LRD make it difficult to determine whether data is generated from a stationary LRD process or a non-stationary one. Roughan [148] has shown that LRD in data makes anomaly detection easier as stronger correlations mean that large immediate deviations are rarer. This approach can determine differences from LRD data, particularly the case of non-stationarity, more than traditional techniques. We will investigate some approaches to test for robustness to non-stationary for estimation of entropy for LRD processes in this thesis.

C.1 R/S Statistic for FGN

We present a discussion the growth of the R/S statistic of FGN from Samorodnitsky [151, pg. 181]. By the definition of FGN, the partial sums are $S_n = B^{\mathcal{H}}(n)$ for all n. Then we can see

$$\begin{aligned}
&\max_{0\le i\le n}\left(S_i - \frac{i}{n}S_n\right) - \min_{0\le i\le n}\left(S_i - \frac{i}{n}S_n\right)\\
&= \max_{0\le i\le n}\left(B^{\mathcal{H}}(i) - \frac{i}{n}B^{\mathcal{H}}(n)\right) - \min_{0\le i\le n}\left(B^{\mathcal{H}}(i) - \frac{i}{n}B^{\mathcal{H}}(n)\right),\\
&\overset{D}{=} n^{\mathcal{H}}\left[\max_{0\le i\le n}\left(B^{\mathcal{H}}(\frac{i}{n}) - \frac{i}{n}B^{\mathcal{H}}(1)\right) - \min_{0\le i\le n}\left(B^{\mathcal{H}}(\frac{i}{n}) - \frac{i}{n}B^{\mathcal{H}}(1)\right)\right],
\end{aligned}$$

by the self-similarity property of FBM. Then, from the continuity of the sample paths of FBM, we have

$$\begin{aligned}&\max_{0\le i\le n}\left(B^{\mathcal{H}}(\frac{i}{n})-\frac{i}{n}B^{\mathcal{H}}(1)\right)-\min_{0\le i\le n}\left(B^{\mathcal{H}}(\frac{i}{n})-\frac{i}{n}B^{\mathcal{H}}(1)\right)\\ \to&\sup_{0\le t\le 1}\left(B^{\mathcal{H}}(t)-tB^{\mathcal{H}}(1)\right)-\inf_{0\le t\le 1}\left(B^{\mathcal{H}}(t)-tB^{\mathcal{H}}(1)\right),\end{aligned}$$

with probability 1. Which then implies that the R/S statistic of FGN grows like $n^{\mathcal{H}}$, since

$$n^{-\mathcal{H}}\frac{R}{S}(X_1,\ldots,X_n)\Rightarrow\frac{1}{\sigma}\sup_{0\le t\le 1}\left(B^{\mathcal{H}}(t)-tB^{\mathcal{H}}(1)\right)-\inf_{0\le t\le 1}\left(B^{\mathcal{H}}(t)-tB^{\mathcal{H}}(1)\right),$$

as the denominator is same as the previous arguments of the R/S statistic.

Appendix D

Markov Chains

In this section we introduce some additional content that is required to understand the definitions and concepts used in Chapter 5.

We will introduce a common example that will be used to highlight the properties introduced, and to build some intuition on the behaviour. This is the simple random walk on the integers, which we will define as a Markov chain, via its transition probabilities. We define the simple random walk, $\{S_n\}_{n\in\mathbb{Z}^+}$ with support on $\mathbb{Z}$, as the stochastic process $S_0 = 0$, with the following probability transitions to the next state from the current state for $i \in \mathbb{Z}$,

$$p_{i,i+1} = p_{i,i-1} = \frac{1}{2}.$$

That is, we are considering the random walk starting at the origin, that moves to the next or previous integer with equal probability.

We want to be able to describe properties of Markov chains as existing for classes of states, rather than considering the individual state. We will introduce a concept of accessibility and then discuss the classification of different states of a Markov chain.

Definition D.0.1. *For states, $i, j \in \Omega$, we say that j is accessible from i, $i \to j$ if there exists an m, such that the m-step transition probability is positive, i.e.,*

$$p_{i,j}^m > 0.$$

Using Definition D.0.1 we can generalise accessibility to a definition of accessibility in both directions, we call states with this property communicating states. We call all members of a subset that communicate exclusively a communicating class.

Definition D.0.2. *For states, $i, j \in \Omega$, we say that j communicates with i, $i \leftrightarrow j$ if $i \to j$ and $j \to i$.*

Under this definition, communication forms an equivalence relation. They have the reflexive property since for all $i, j, k \in \Omega$, we have that every state communicates with itself, and symmetric since $i \leftrightarrow j$ implies that $j \leftrightarrow i$. A quick argument for transitivity follows, since $i \leftrightarrow j$ and $j \leftrightarrow k$ implies that

$$\exists\ m > 0, \text{ such that } p_{i,j}^m > 0, \text{ and } \exists\ n > 0, \text{ such that } p_{j,k}^n > 0.$$

Then we split up the transition from i to k into portions of m and n steps, which gives

$$p_{i,k}^{n+m} = \sum_{l \in \Omega} p_{i,l}^n p_{l,k}^m \geq p_{i,j}^n p_{j,k}^m > 0,$$

by the communication of $i \leftrightarrow j$ and $j \leftrightarrow k$. Communicating classes partition the states of a Markov chain, and many properties can be assigned to the entire class.

Next we will discuss two important properties, irreducibility and aperiodicity, which are used to establish a property called ergodicity, meaning that the Markov chain eventually "forgets" its initial state. These properties enable some rich analysis and classification of behaviour of Markov chains. The first, irreducibility, is related to the communicating classes defined above.

Definition D.0.3. *A Markov chain, $\{X_n\}_{n \in \mathbb{Z}^+}$, is called irreducible if the Markov chain consists of a single communicating class. That is, for every pair of states, $i, j \in \Omega$, there exists an integer m, such that the m-step transition probability is positive, i.e.,*

$$\forall i, j \in \Omega, \quad p_{i,j}^m > 0.$$

Intuitively this property says that all states are reachable from all states, and hence we can't reduce the chain into disjoint connected smaller chains. We can see that the simple random walk has this property, if we consider the n-step probability from above and then consider a direct path from i to j, where without loss of generality $j > i$ and $|i - j| = n$, then

$$p_{i,j}^n \geq p_{i,i+1} p_{i+1,i+2} \cdots p_{j-1,j} = \left(\frac{1}{2}\right)^n > 0.$$

The next property is aperiodicity, which refers to the absence of periodic behaviour in the state transitions. We define the period, $d(i)$, of a state, i, to be defined as the greatest common divisor of the set $d(i) = \{m \geq 1 : p_{i,i}^m > 0\}$. With the assumption of an irreducible Markov chain the following result shows that the period is a property of the entire irreducible chain.

Theorem D.0.1 (Lemma 1.6 [118]). *For a Markov chain, $\{X_n\}_{n\in\mathbb{Z}^+}$, with i and j in the same communicating class, then the greatest common divisor of $d(i)$ is equal to the greatest common divisor of $d(j)$.*

This result easily generalises to irreducibility, by considering the entire chain as the communicating class. We call properties that apply to all states of a communicating class a class property. This leads to the following definition of aperiodicity, and for irreducible Markov chains the property is the same for the entire chain.

Definition D.0.4. *A Markov chain, $\{X_n\}_{n\in\mathbb{Z}^+}$, is called aperiodic if every state of the chain has period 1.*

The simple random walk is not aperiodic, since the probability of being in a state depends on whether you have taken an odd or even number of steps. For example at step 0 we are in state 0 and then on step 1 we move to an odd numbered state, either 1 or -1. Then in step 2 we move to either -2, 0 or 2, *i.e.,* even numbered states. This informal argument generalises and we see that starting at the origin we must be in an odd numbered state after an odd number of steps and an even numbered state after an even number of steps.

We can make a small change to the simple random walk to make it aperiodic, by adding a positive probability of remaining in the same state. In general, for an irreducible Markov chain adding a self loop, *i.e.,* $p_{i,i} > 0$ for any $i \in \Omega$ is sufficient to ensure that a Markov chain is aperiodic [58, pg. 76]. We call this the lazy random walk, and we define the following transitions

$$p_{i,i+1} = p_{i,i-1} = \frac{1}{4} \quad \text{and} \quad p_{i,i} = \frac{1}{2}.$$

We define an ergodic Markov chain below, these conditions ensure that there is a unique limiting distribution of a Markov chain on a finite number of states.

Definition D.0.5. *We call a Markov chain, $\{X_n\}_{n\in\mathbb{Z}^+}$, ergodic if it is irreducible and aperiodic.*

Using this definition, we can see that the simple random walk is not an ergodic Markov chain, but the lazy random walk is ergodic.

We discuss some properties of the time to reach and return to states. First we define a random variable, called the hitting time, intuitively used to analyse the time from beginning until reaching a certain state.

Definition D.0.6. *The hitting time of a set $A \subset \Omega$, is the random variable, $T_A : \Omega \to \{0, 1, 2, \ldots\} \cup \{\infty\}$ defined as*

$$T_A = \inf\{n \geq 0 : X_n \in A\}.$$

A closely related concept is the return time, similar to the hitting time adding restriction on the starting state.

Definition D.0.7. *The return time to a state $i \in \Omega$, is the random variable, $T_{i,i} : \Omega \to \{1, 2, \ldots\} \cup \{\infty\}$ defined as*

$$T_{i,i} = \inf\{n \geq 1 : X_n = i, X_0 = i\}.$$

We define the probability of the first return to state i occurring on at the nth step, as

$$f_{i,i}^n = \mathbb{P}\left(T_{i,i} = n\right).$$

Then the probability of ever returning to the ith state is

$$f_{i,i} = \mathbb{P}(T_{i,i} < \infty) = \sum_{n=1}^{\infty} f_{i,i}^n.$$

We use these quantities to define the concepts of recurrence and transience, describing whether the Markov chain returns to a particular state.

Definition D.0.8. *A state, i, is called recurrent if $f_{i,i} = 1$, or transient if $f_{i,i} < 1$.*

Since the Markov chain depends on the past through its current state only, any return with probability 1 ensures that the chain returns to the state infinitely often, so each return can be considered independently.

Another characterisation of recurrence and transience can be made by the n-step transition probabilities, where recurrent states have the property

$$\sum_{n=1}^{\infty} p_{i,i}^n = \infty,$$

and transient states having the property

$$\sum_{n=1}^{\infty} p_{i,i}^n < \infty,$$

by Proposition 21.3 [118].

Recurrence is a class property, classifying all states into communication classes, as shown by the following theorem.

Theorem D.0.2 (Theorem 5.3.16 [34]). *For an irreducible Markov chain, $\{X_n\}_{n\in\mathbb{Z}^+}$, all states are either transient or recurrent.*

This is an extremely useful result, as we can use knowledge about an individual state to apply to all states within an irreducible Markov chain.

The simple random walk is recurrent, which we show by the following argument from Bremard [22, Example 7.1.6]. From Theorem D.0.2, we only need to show the recurrence of one state, the origin. For odd numbers of steps the random walk cannot be at the origin, *i.e.*, $p_{0,0}^n = 0$ when $n = 2m+1, m \in \mathbb{Z}^+$. The probability of being at the origin after even steps, $n = 2m$, is given by

$$p_{0,0}^{2m} = \binom{2m}{m}\left(\frac{1}{2}\right)^{2m},$$

since for a path $2m$ steps long, we take m steps in both directions to be at the origin, and any individual path occurs with probability $\left(\frac{1}{2}\right)^{2m}$.

We have by Stirling's approximation, $n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$ [63, pg. 52], that

$$p_{0,0}^{2m} \sim \frac{1}{\sqrt{\pi m}}.$$

Then taking the partial sum of n-step transition probabilities as $n \to \infty$,

$$\sum_{n=1}^{\infty} p_{0,0}^n \sim \sum_{m=1}^{\infty} \frac{1}{\sqrt{\pi m}} = \infty,$$

and by Theorem D.0.2 the random walk is recurrent.

The Markov chains we consider in this section will be irreducible, aperiodic and recurrent. We make a further classification of the recurrent states into positive or null recurrence, by the finiteness of the expected return time to a state i. The expected return time of of a Markov chain to state i is given by

$$\mathbb{E}[T_{i,i}] = \sum_{n=1}^{\infty} n\mathbb{P}(T_{i,i} = n).$$

Definition D.0.9. *An irreducible Markov chain, $\{X_n\}_{n\in\mathbb{Z}^+}$, is called positive-recurrent if*

$$\mathbb{E}[T_{i,i}] < \infty,$$

and called null-recurrent if

$$\mathbb{E}[T_{i,i}] = \infty.$$

Remark. *Theorem D.0.2 also applies to null-recurrence and positive-recurrence, and a recurrent state must be in one of these classes.*

This may seem like a counter-intuitive distinction, however simple random walks demonstrate how recurrence does not ensure that the expected time to return is finite. We use a probability generating function approach adapted from Ash [6, pg. 193], for the probability $f^n_{i,i,}$. We define the probability generating function of the return time probability as

$$G_i(z) = \sum_{n=1}^{\infty} z^n f^n_{i,i} = \mathbb{E}[z^{T_{i,i}}].$$

Note that this function has the property that $\mathbb{E}[T_{i,i}] = G'_i(1)$ [34, pg. 31], which we use to show that this quantity is infinite. We analyse the return to the origin only, since positive and null recurrence are class properties [34, Theorem 5.3.16]. We define a second generating function for the n-step probability transitions, *i.e.,* for the probability $p^n_{0,0} = \mathbb{P}(S_n = 0)$

$$F_0(z) = \sum_{n=0}^{\infty} p^n_{0,0} z^n.$$

We partition the probability of being at the origin after n steps, $p^n_{0,0}$, via the first return time probabilities. Since if a first return occurs at a time, k, between 0 and n, then the probability of being at the origin at n is $p^{n-k}_{0,0} f^k_{0,0}$. Therefore

$$p^n_{0,0} = \sum_{k=0}^{n} p^{n-k}_{0,0} f^k_{0,0}, \quad \forall n \in \mathbb{Z}^+,$$

where $f^0_{0,0} = 0$ and $p^0_{0,0} = 1$. We define another generating function

$$H(z) = F_0(z)G_0(z),$$

which is a convolution of the previous two [6, pg. 192]. However, note that $h_n = p^n_{0,0}$ for $n \geq 1$, and $h_0 = 0 = p^0_{0,0} - 1$. This implies that

$$H(z) = F_0(z) - 1.$$

Combining the two expression and putting in terms of $G_0(z)$, we get

$$G_0(z) = 1 - \frac{1}{F_0(z)}.$$

Given knowledge of $p_{0,0}^{2n}$, we get the following expression

$$F_0(z) = \sum_{n=0}^{\infty} p_{0,0}^n z^n = \sum_{n=0}^{\infty} p_{0,0}^{2n} z^{2n} = \sum_{n=0}^{\infty} \binom{2n}{n} \left(\frac{1}{2}\right)^{2n} z^{2n}.$$

From equation 5.72 of Graham *et. al.* [76], we see

$$\sum_{n=0}^{\infty} \binom{2n}{n} \left(\frac{1}{2}\right)^{2n} z^{2n} = \frac{1}{\sqrt{1-z^2}}.$$

Which implies that the generating function of the return time is

$$G_0(z) = 1 - \sqrt{1-z^2}.$$

Therefore, we calculate the derivative and hence the expected value of the time to return to the origin as

$$G_0'(z) = z\left(1-z^2\right)^{-\frac{1}{2}}.$$

As $z \to 1$, this implies that $G_0'(z) \to \infty$, and therefore the expected hitting time is infinite, *i.e.,* $\mathbb{E}[T_{0,0}] = \infty$ and the simple random walk is null-recurrent.

These properties enable us to classify the long-term behaviour of the Markov chain. We will introduce the stationary distribution, giving the long-term probabilities of the Markov chain, *i.e.*, $\mathbb{P}(X_n = j)$.

Theorem D.0.3 (Theorem 6.2.1 [34])**.** *For an irreducible and aperiodic Markov chain, $\{X_n\}_{n\in\mathbb{Z}^+}$, all states are positive-recurrent if and only if there is a solution to the set of linear equations*

$$\pi_j = \sum_{i\in\Omega} \pi_i p_{i,j}, \quad \forall j \in \Omega,$$

and,

$$\sum_{j\in\Omega} \pi_j = 1.$$

If there exists a solution $\boldsymbol{\pi}$, then it is strictly positive, the solution is unique and,

$$\pi_j = \lim_{n\to\infty} p_{i,j}^n,$$

for all $i, j \in \Omega$.

We call this distribution the stationary distribution, also known as the limiting or invariant distribution, $\pi_j = \mathbb{P}(X_n = j)$. Once the chain becomes stationary, the chain remains stationary as seen from Theorem D.0.3, where the stationary probability of a state j can be decomposed into the probability of being in any state and taking a step to state j.

An example of calculating the stationary distribution is given for the following 2-state Markov chain. We define the Markov chain via the transition probabilities,

$$\begin{aligned} & p_{0,0} = 1 - p, \quad p_{0,1} = p, \\ \text{and,} \quad & p_{1,0} = q, \quad p_{1,1} = 1 - q. \end{aligned}$$

From Theorem D.0.3, we have the following equations for the stationary probabilities, π_0 and π_1,

$$\begin{aligned} \pi_0 &= (1-p)\,\pi_0 + q\pi_1, \\ \pi_1 &= p\pi_0 + (1-q)\,\pi_1, \\ \text{and,} \quad \pi_0 + \pi_1 &= 1. \end{aligned}$$

Substituting the second equation into the first, results in

$$\begin{aligned} \pi_0 &= (1-p)\,\pi_0 + q\left(p\pi_0 + (1-q)\,\pi_1\right), \\ \implies p\,(1-q)\,\pi_0 &= q\,(1-q)\,\pi_1, \\ \implies \pi_0 &= \frac{p}{q}\pi_1. \end{aligned}$$

Which by substitution into the expression, $\pi_0 + \pi_1 = 1$, gives

$$\pi_0 = \frac{q}{p+q}, \quad \pi_1 = \frac{p}{p+q},$$

the long run probabilities of being in the states 0 and 1.

Informally, we can see that irreducibility and aperiodicity are needed for a unique solution. If a Markov chain had states that didn't communicate then the limit of the n-step probabilities would depend on which communication class the chain started in, and would not be unique, hence the requirement for irreducibility. If there is periodicity then the limits of the n-step probabilities will not exist, as the limit will depend on the particular n.

The limiting behaviour of the n-step transition probabilities is very different in the case of a null-recurrent Markov chain, illustrated by the following theorem.

Theorem D.0.4 (Theorem 21.17 [118]). *For a null-recurrent Markov chain,* $\{X_n\}_{n\in\mathbb{Z}^+}$*, for all* $i, j \in \Omega$*,*

$$\lim_{n\to\infty} p_{i,j}^n = 0.$$

So in contrast to the positive recurrent case, the limit of the n-step transition probabilities tends to 0 as $n \to \infty$, *e.g.*, the simple random walk.

In the next theorem we discuss the link between the limits of the n-step transition probabilities and the expected return time.

Theorem D.0.5 (Theorem 7.4.1 [6]). *For an irreducible and aperiodic Markov chain,* $\{X_n\}_{n\in\mathbb{Z}^+}$*,*

$$p_{i,i}^n \to \frac{1}{\mathbb{E}[T_{i,i}]}, \quad \text{as } n \to \infty.$$

This theorem demonstrates that in the case of the positive-recurrent Markov chains that the stationary probability for a state is the reciprocal of the expected time between visits between the state. Note that Theorem D.0.5 also applies to null-recurrent chains, given the infinite expected time between visits the limiting probability is 0.

In this thesis, we only be consider positive-recurrent Markov chains, due to the existence of a stationary distribution. We investigate the entropy rate of Markov chains with long range dependence, as we saw in Chapter 2 the entropy rate of a Markov chain depends on the stationary distribution. We will introduce a final result in this section, the ergodic theorem for Markov chains.

Theorem D.0.6 (Theorem 4.16 [118]). *Let* g *be a real-valued function defined on a state space,* Ω*. For an irreducible Markov chain, with any starting distribution,* α*,*

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} g(X_k) = \mathbb{E}_\pi\left(f\right)\right) = 1,$$

where $\mathbb{E}_\pi\left(g\right) = \sum_{i\in\Omega} \lim_{n\to\infty} p_{i,i}^n g(i)$*.*

This theorem shows that the time average of a function applied to a Markov chain equals the space average, independent of the starting distribution. This is, a Markov chain "forgets" its initial state.

Appendix E

Estimation Theory

In this section we will introduce some of the key concepts in estimation, and assessing the quality of estimators. These will be used in the discussion of the estimation techniques, to be able to compare the strengths and weaknesses of different approaches.

For some context, we will introduce the estimation problem that we are looking solve generally. That is, given a sample of data, $x_1, \ldots, x_n$ we are aiming to define a function, T_n, such that we generate an estimate of a quantity, θ, as $\hat{\theta} = T(x_1, \ldots, x_n) = T_n$. Since the data is being generated randomly, we view the estimator, $\hat{\theta}$ as a random variable. We will introduce some properties that quantify how the estimator performs with respect to the true value and to describe the behaviour of the estimator. First we define some convergence properties, that are used to describe and compare estimation techniques.

E.1 Convergence

To describe the convergence of random variables we will introduce three different types of convergence used in probability theory, convergence in probability, convergence in distribution and almost sure convergence. First we define convergence in probability [115, pg. 332]

Definition E.1.1. *A sequence of random variables, Y_n converges in probability to a constant, c, if for every $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left\{|Y_n - c| > \epsilon\right\} \to 0.$$

Another common type of convergence is convergence in distribution, commonly called weak convergence [115, pg. 336].

Definition E.1.2. *A sequence of random variables, Y_n with cumulative distribution function,*

$$F_n(y) = \mathbb{P}(Y_n \leq y),$$

and there exists a cumulative distribution function, F, such that

$$\lim_{n\to\infty} F_n(y) \to F(y),$$

for all points y for which F is continuous. Then we say that the distribution functions F_n converge in distribution to F.

A final mode of convergence that is discussed is called almost sure convergence or strong convergence [23, Definition 10.8.1].

Definition E.1.3. *A sequence of random variables, Y_n, is said to converge to a random variable, Y, almost surely if*

$$\mathbb{P}\left(Y_n = Y\right) = 1.$$

All of these modes of convergence are subtly different and results stated in the thesis are with respect to specific types of convergence. Although many results do generalise across different modes of convergence, we have to be careful to ensure we refer to the correct type of convergence in theorems used for the construction of estimators.

E.2 Estimation Properties

We will describe some important properties of estimators. These will be used in the assessment and discussion of the quality of estimation techniques.

An important concept that we introduce here is the consistency of an estimator, which informally states that an estimator will converge to the true value, given n data points, as $n \to \infty$.

Definition E.2.1. *An estimator, T_n, of a quantity, θ, is called consistent if it converges to the true value in probability,* i.e.*,*

$$\lim_{n\to\infty} \mathbb{P}\left\{|T_n - \theta| > \epsilon\right\} \to 0.$$

This definition is often called weak consistency, for strong consistency we replace the mode of convergence of T_n to almost sure convergence.

An important property of an estimator concerns the form of the asymptotic distribution of the estimator. If the asymptotic distribution follows a normal distribution, then we say the estimator has asymptotic normality. The theory is related to the central limit theorem, and forms a natural extension, with some conditions on the data [115, pg. 336]

Definition E.2.2. *An estimator, T_n, of a quantity, θ, is asymptotically normal if*

$$\sqrt{n}\left(T_n - \theta\right) \sim \mathcal{N}\left(0, \sigma^2\right),$$

in distribution, where $\mathcal{N}\left(0, \sigma^2\right)$ is a normally distributed random variable and σ^2 is the variance.

We are going to define a measure of the quality of an estimator, called efficiency, which quantifies the variance of the estimator with respect to the lowest possible variance. First we define a quantity, the Fisher information, and a lower bound for the variance called the Cramer-Rao bound. The Fisher information, $I(\theta)$, quantifies the amount of information that a random variable, X, carries about a parameter, θ, where $f\left(X|\theta\right)$ is the probability mass, or density, that is a parametric model of X.

Definition E.2.3. *The Fisher information, $I(\theta)$, of an estimator, T_n for a quantity, θ, is defined as*

$$I(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\log f\left(X|\theta\right)\right)^2 \middle| \theta\right],$$

$$= \int \left(\frac{\partial}{\partial\theta}\log f\left(x|\theta\right)\right)^2 f(x|\theta)dx.$$

The Cramer-Rao bound for estimators links the variance to a lower bound of the reciprocal of the Fisher information [43, pg. 480].

Theorem E.2.1 (Cramer-Rao Bound)**.** *The variance of an estimator, T_n, of a parameter, θ, has the following bound,*

$$Var\left(T_n\right) \geq \frac{1}{I\left(\theta\right)}.$$

This result gives the lowest possible variance for an estimator. The concept of efficiency measures how an estimator performs against its best possible value, that is equality in Theorem E.2.1.

Definition E.2.4. *The efficiency of an unbiased estimator, T_n is defined as*

$$e(T_n) = \frac{\frac{1}{I(\theta)}}{Var\left(T_n\right)}.$$

We analyse the errors of estimators by considering the deviation of the estimate from its true value, $T_n - \theta$. We define a quantity that measures the deviation from the true value, the mean-squared error, which has large weighting on outliers and is used to assess the quality of estimation.

Definition E.2.5. *The mean-squared error, $MSE(T_n)$, of an estimate is defined as*

$$MSE\left(T_n\right) = E\left[\left(T_n - \theta\right)^2\right].$$

This is related to the variance and bias of the estimator, however this quantifies the squared deviation of the estimates and the deviation from the true value respectively.

Definition E.2.6. *The variance of an estimator, T_n, is given by*

$$Var\left(T_n\right) = E\left[\left(T_n - E[T_n]\right)^2\right].$$

Then we define the bias, the difference between the true value and expected value of the estimator.

Definition E.2.7. *The bias of an estimator, T_n, is defined as*

$$Bias\left(T_n\right) = E[T_n] - \theta.$$

These three are related by the bias-variance tradeoff, or the mean-squared error decomposition [85, pg. 24]. Which is the statement that

$$\begin{aligned} MSE\left(T_n\right) &= E\left[\left(T_n - \theta\right)^2\right], \\ &= E[\left(T_n - E[T_n]\right)^2] + \left(E[T_n] - \theta\right)^2, \\ &= \mathrm{Var}\left(T_n\right) + \mathrm{Bias}\left(T_n\right)^2. \end{aligned}$$

In the context of estimation, this means we can achieve better mean-squared error estimators in some cases by utilising biased estimators, depending on the performance goals.

Bibliography

[1] A. A. Al-Babtain, I. Elbatal, C. Chesneau, and M. Elgarhy. Estimation of different types of entropies for the kumaraswamy distribution. *PLOS ONE*, 16(3):1–21, 03 2021.

[2] R. Alcaraz and J. J. Rieta. A review on sample entropy applications for the non-invasive analysis of atrial fibrillation electrocardiograms. *Biomedical Signal Processing and Control*, 5(1):1–14, 2010.

[3] S.-I. Amari. Differential geometry of curved exponential families-curvatures and information loss. *The Annals of Statistics*, 10(2):357 – 385, 1982.

[4] S.-I. Amari. Information geometry. *Japanese Journal of Mathematics*, 16(1):1–48, 2021.

[5] J. M. Amigó, S. G. Balogh, and S. Hernández. A brief review of generalized entropies. *Entropy*, 20(11):813, 2018.

[6] R. B. Ash. *Basic probability theory*. Wiley New York, 1970.

[7] C. Bandt and B. Pompe. Permutation entropy: a natural complexity measure for time series. *Physical review letters*, 88(17):174102, 2002.

[8] C. Bandt and F. Shiha. Order patterns in time series. *Journal of Time Series Analysis*, 28(5):646–665, 2007.

[9] M. S. Bartlett. On the theoretical specification and sampling properties of autocorrelated time-series. *Supplement to the Journal of the Royal Statistical Society*, 8(1):27–41, 1946.

[10] M. S. Bartlett. Periodogram analysis and continuous spectra. *Biometrika*, 37(1/2):1–16, 1950.

[11] G. P. Basharin. On a statistical estimate for the entropy of a sequence of independent random variables. *Theory of Probability & Its Applications*, 4(3):333–336, 1959.

[12] J. Beirlant, E. J. Dudewicz, L. Györfi, E. C. Van der Meulen, et al. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6(1):17–39, 1997.

[13] J. Beran. Statistical methods for data with long-range dependence. *Statistical Science*, 7(4):404 – 416, 1992.

[14] J. Beran. *Statistics for Long-Memory Processes*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 1994.

[15] A. Berger, S. A. Della Pietra, and V. J. Della Pietra. A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71, 1996.

[16] J. M. Berger and B. Mandelbrot. A new model for error clustering in telephone circuits. *IBM Journal of Research and Development*, 7(3):224–236, July 1963.

[17] R. N. Bhattacharya, V. K. Gupta, and E. Waymire. The Hurst effect under trends. *Journal of Applied Probability*, 20(3):649?662, 1983.

[18] F. Biagini, Y. Hu, B. Øksendal, and T. Zhang. *Stochastic calculus for fractional Brownian motion and applications*. Springer Science & Business Media, 2008.

[19] N. H. Bingham. Szegö's theorem and its probabilistic descendants. *Probability Surveys*, 9:287–324, 2012.

[20] S. Bouzebda and I. Elhattab. Uniform-in-bandwidth consistency for kernel-type estimators of shannon's entropy. *Electronic Journal of Statistics*, 5:440–459, 2011.

[21] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[22] P. Brémaud. *Discrete probability models and methods: probability on graphs and trees, Markov chains and random fields, entropy and coding*. Springer Publishing Company, Incorporated, 1st edition, 2017.

[23] P. J. Brockwell and R. A. Davis. *Time series: theory and methods.* Springer-Verlag, Berlin, Heidelberg, 1986.

[24] J. Burg. Maximum entropy spectral analysis, paper presented at the 37th meeting. *Society of Exploration Geophysics, Oklahoma City*, 1967.

[25] J. P. Burg. *Maximum entropy spectral analysis.* Ph.D. Thesis, Department of Geophysics. Stanford University, 1975.

[26] J. Capon. Maximum-likelihood spectral estimation. In *Nonlinear Methods of Spectral Analysis*, pages 155–179. Springer, Berlin, Heidelberg, 1983.

[27] K. J. E. Carpio and D. Daley. Long-range dependence of Markov chains in discrete time on countable state space. *Journal of Applied Probability*, 44(4):1047–1055, 2007.

[28] C. K. Carter and R. Kohn. Semiparametric Bayesian inference for time series with mixed spectra. *Journal of the Royal Statistical Society Series B*, 59(1):255–268, 1997.

[29] H. S. Chang. On convergence rate of the shannon entropy rate of ergodic markov chains via sample-path simulation. *Statistics & probability letters*, 76(12):1261–1264, 2006.

[30] G. Chavez. Conditional and marginal mutual information in Gaussian and hyperbolic decay time series. *Journal of Time Series Analysis*, 37(6):851–861, 2016.

[31] X. Chen, I. C. Solomon, and K. H. Chon. Comparison of the use of approximate entropy and sample entropy: applications to neural respiratory signal. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 4212–4215. IEEE, 2006.

[32] B. Choi and T. M. Cover. An information-theoretic proof of Burg's maximum entropy spectrum. *Proceedings of the IEEE*, 72(8):1094–1096, 1984.

[33] N. Choudhuri, S. Ghosal, and A. Roy. Bayesian estimation of the spectral density of a time series. *Journal of the American Statistical Association*, 99(468):1050–1059, 2004.

[34] E. Cinlar. *Introduction to Stochastic Processes.* Prentice-Hall, 1975.

[35] G. Ciuperca and V. Girardin. On the estimation of the entropy rate of finite Markov chains. In *Proceedings of the international symposium on applied stochastic models and data analysis, ENST Bretagne. Brest, France*, pages 1109–1117.

[36] G. Ciuperca and V. Girardin. Estimation of the entropy rate of a countable Markov chain. *Communications in Statistics: Theory and Methods*, 36(14):2543–2557, 2007.

[37] G. Ciuperca, V. Girardin, and L. Lhote. Computation and estimation of generalized entropy rates for denumerable Markov chains. *IEEE Transactions on information theory*, 57(7):4026–4034, 2011.

[38] A. Clauset, C. R. Shalizi, and M. E. Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.

[39] R. Cont. Long range dependence in financial markets. In *Fractals in Engineering*, pages 159–179, London, 2005. Springer London.

[40] L. Contreras Rodríguez, E. J. Madarro-Capó, C. M. Legón-Pérez, O. Rojas, and G. Sosa-Gómez. Selecting an effective entropy estimator for short sequences of bits and bytes with maximum entropy. *Entropy*, 23(5):561, 2021.

[41] T. M. Cover and J. A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006.

[42] D. R. Cox. *Principles of Statistical Inference*. Cambridge University Press, 2006.

[43] H. Cramér. *Mathematical methods of statistics*. Princeton University Press, 1946.

[44] J. P. Crutchfield and D. P. Feldman. Regularities unseen, randomness observed: levels of entropy convergence. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 13(1):25–54, 2003.

[45] D. J. Daley. The Hurst index of long-range dependent renewal processes. *The Annals of Probability*, 27(4):2035 – 2041, 1999.

[46] D. J. Daley and R. Vesilo. Long range dependence of point processes, with queueing examples. *Stochastic Processes and Their Applications*, 70(2):265–282, 1997.

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

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

[49] R. B. Davies and D. Harte. Tests for Hurst effect. *Biometrika*, 74(1):95–101, 1987.

[50] Ł. Dębowski. *Własności entropii nadwyżkowej dla procesów stochastycznych nad różnymi alfabetami*. PhD thesis, Institute of Computer Science, Polish Academy of Sciences, 2005.

[51] Ł. Dębowski. *Information Theory and Statistics*. Citeseer, 2013.

[52] A. Delgado-Bonal and A. Marshak. Approximate entropy and sample entropy: A comprehensive tutorial. *Entropy*, 21(6):541, 2019.

[53] F. X. Diebold and A. Inoue. Long memory and regime switching. *Journal of econometrics*, 105(1):131–159, 2001.

[54] Y. Ding and X. Xiang. An entropic characterization of long memory stationary process, 2016. arXiv preprint http://arxiv.org/abs/1604.05453.

[55] R. Dobrushin. A simplified method of experimental estimation of the entropy of a stationary distribution. *Theory of Probability and its Applications*, 3:462–464, 1958.

[56] M. E. Dumitrescu. Some informational properties of Markov pure-jump processes. *Časopis pro pěstování matematiky*, 113(4):429–434, 1988.

[57] R. Durrett. *Probability: Theory and Examples*. Cambridge University Press, New York, NY, USA, 4th edition, 2010.

[58] R. Durrett. *Essentials of Stochastic Processes*. Springer Texts in Statistics. Springer New York, 2012.

[59] J. P. Eckmann and D. Ruelle. Ergodic theory of chaos and strange attractors. *Rev. Mod. Phys.*, 57:617–656, 1985.

[60] M. C. Edwards, R. Meyer, and N. Christensen. Bayesian nonparametric spectral density estimation using b-spline priors. *Statistics and Computing*, 29(1):67–78, 2019.

[61] S. Egner, V. Balakirsky, L. Tolhuizen, S. Baggen, and H. Hollmann. On the entropy rate of a hidden Markov model. In *International Symposium on Information Theory, 2004. ISIT 2004. Proceedings.*, page 12, 2004.

[62] Y. Ephraim and N. Merhav. Hidden Markov processes. *IEEE Transactions on Information Theory*, 48(6):1518–1569, 2002.

[63] W. Feller. *An Introduction to Probability Theory and Its Applications, Volume 1.* A Wiley publication in mathematical statistics. Wiley, 1968.

[64] A. Feutrill and M. Roughan. Differential entropy rate characterisations of long range dependent processes, 2021. arXiv preprint arXiv:2102.05306.

[65] A. Feutrill and M. Roughan. A review of Shannon and differential entropy rate estimation. *Entropy*, 23(8):1046, 2021.

[66] C. Flynn. fbm. `https://github.com/crflynn/fbm`, 2018.

[67] S. Foss, D. Korshunov, and S. Zachary. *An Introduction to Heavy-Tailed and Subexponential Distributions*. Springer Series in Operations Research and Financial Engineering. Springer New York, 2011.

[68] J. Franke. ARMA processes have maximal entropy among time series with prescribed autocovariances and impulse responses. *Advances in Applied Probability*, 17(4):810?840, 1985.

[69] J. Franke. A Levinson-Durbin recursion for autoregressive-moving average processes. *Biometrika*, 72(3):573–581, 1985.

[70] A. Gangopadhyay, B. Mallick, and D. Denison. Estimation of spectral density of a stationary time series via an asymptotic representation of the periodogram. *Journal of Statistical Planning and Inference*, 75(2):281–290, 1999.

[71] Y. Gao, I. Kontoyiannis, and E. Bienenstock. Estimating the entropy of binary time series: methodology, some theory and a simulation study. *Entropy*, 10(2):71?99, Jun 2008.

[72] A. Gefferth, D. Veitch, I. Maricza, S. Molnár, and I. Ruzsa. The nature of discrete second-order self-similarity. *Advances in Applied Probability*, 35(2):395–416, 2003.

[73] J. Gibbons and S. Chakraborti. *Nonparametric Statistical Inference*. Taylor & Francis, 4th edition, 2014.

[74] V. Girardin and A. Sesboüé. Asymptotic study of an estimator of the entropy rate of a two-state Markov chain for one long trajectory. In *AIP Conference Proceedings*, volume 872, pages 403–410. American Institute of Physics, 2006.

[75] V. Girardin and A. Sesboüé. Comparative construction of plug-in estimators of the entropy rate of two-state Markov chains. *Methodology and Computing in Applied Probability*, 11(2):181–200, 2009.

[76] R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics: A Foundation for Computer Science*. Addison-Wesley Longman Publishing Co., Inc., USA, 2nd edition, 1994.

[77] C. W. J. Granger and R. Joyeux. An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis*, 1(1):15–29, 1980.

[78] P. Grassberger. Toward a quantitative theory of self-generated complexity. *International Journal of Theoretical Physics*, 25(9):907–938, 1986.

[79] P. Grassberger. Estimating the information content of symbol sequences and efficient codes. *IEEE Transactions on Information Theory*, 35(3):669–675, 1989.

[80] T. Graves, R. Gramacy, N. Watkins, and C. Franzke. A brief history of long memory: Hurst, Mandelbrot and the road to ARFIMA, 1951–1980. *Entropy*, 19(9):437, 2017.

[81] U. Grenander. On empirical spectral analysis of stochastic processes. *Arkiv för Matematik*, 1(6):503–531, 1952.

[82] G. Han and B. Marcus. Analyticity of entropy rate of hidden Markov chains. *IEEE Transactions on Information Theory*, 52(12):5251–5266, 2006.

[83] Y. Han, J. Jiao, C.-Z. Lee, T. Weissman, Y. Wu, and T. Yu. Entropy rate estimation for Markov chains with large state space. *arXiv preprint arXiv:1802.07889*, 2018.

[84] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del R'ıo, M. Wiebe, P. Peterson, P. G'erard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020.

[85] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition.* Springer Series in Statistics. Springer New York, 2009.

[86] C. Heyde and Y. Yang. On defining long-range dependence. *Journal of Applied probability*, 34(4):939–944, 1997.

[87] J. R. M. Hosking. Fractional differencing. *Biometrika*, 68(1):165–176, 04 1981.

[88] J. Hunter. *Mathematical techniques of applied probability.* Mathematical techniques of applied probability. Acad. Press, 1983.

[89] H. E. Hurst. Long-term storage capacity of reservoirs. *Trans. Amer. Soc. Civil Eng.*, 116:770–799, 1951.

[90] I. A. Ibragimov and I. A. Rozanov. *Gaussian random processes.* Springer-Verlag New York, 1978.

[91] S. Ihara. Maximum entropy spectral analysis and ARMA processes. *IEEE transactions on information theory*, 30(2):377–380, 1984.

[92] S. Ihara. *Information Theory for Continuous Systems.* Series on Probability and Statistics. World Scientific, 1993.

[93] A. Inoue. Asymptotic behavior for partial autocorrelation functions of fractional ARIMA processes. *The Annals of Applied Probability*, 12(4):1471 – 1491, 2002.

[94] A. Inoue. AR and MA representation of partial autocorrelation functions, with applications. *Probability theory and related fields*, 140(3):523–551, 2008.

[95] P. Jacquet, G. Seroussi, and W. Szpankowski. On the entropy of a hidden Markov process. *Theoretical Computer Science*, 395(2):203–219, 2008.

[96] E. T. Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

[97] A. Jeffrey and H. Dai. *Handbook of Mathematical Formulas and Integrals*. Elsevier Science, 2008.

[98] A. Kaltchenko and N. Timofeeva. Rate of convergence of the nearest neighbor entropy estimator. *AEU-International Journal of Electronics and Communications*, 64(1):75–79, 2010.

[99] A. Kaltchenko, E.-h. Yang, and N. Timofeeva. Entropy estimators with almost sure convergence and an $O(n^{-1})$ variance. In *2007 IEEE Information Theory Workshop*, pages 644–649, 2007.

[100] S. Kamath and S. Verdú. Estimation of entropy rate and Rényi entropy rate for Markov chains. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 685–689. IEEE, 2016.

[101] I. Karatzas and S. Shreve. *Brownian Motion and Stochastic Calculus*. Graduate Texts in Mathematics (113) (Book 113). Springer New York, 1991.

[102] D. G. Kendall. Unitary dilations of Markov transition operators and the corresponding integral representation for transition-probability matrices. In U. Grenander, editor, *Probability and Statistics*, pages 139–161. Almqvist and Wiksell, Stockholm, 1959.

[103] J. M. Keynes. *A treatise on probability*. Macmillan and Company, limited, 1921.

[104] Y. M. Kim, S. N. Lahiri, and D. J. Nordman. Non-parametric spectral density estimation under long-range dependence. *Journal of Time Series Analysis*, 39(3):380–401, 2018.

[105] K. Knopp. *Infinite sequences and series*. Courier Corporation, 1956.

[106] I. Kontoyiannis, P. H. Algoet, Y. M. Suhov, and A. J. Wyner. Nonparametric entropy estimation for stationary processes and random fields, with applications to english text. *IEEE Transactions on Information Theory*, 44(3):1319–1327, 1998.

[107] I. Kontoyiannis and Y. Suhov. Prefixes and the entropy rate for long-range sources. In *Proceedings of 1994 IEEE International Symposium on Information Theory*, page 194, 1994.

[108] S.-Y. Koyama and N. Kurokawa. Euler's integrals and multiple sine functions. *Proceedings of the American Mathematical Society*, 133(5):1257–1265, 2005.

[109] S. Kullback. *Information Theory and Statistics*. A Wiley publication in mathematical statistics. Dover Publications, 1997.

[110] D. E. Lake. Renyi entropy measures of heart rate Gaussianity. *IEEE Transactions on Biomedical Engineering*, 53(1):21–27, 2005.

[111] D. E. Lake, J. S. Richman, M. P. Griffin, and J. R. Moorman. Sample entropy analysis of neonatal heart rate variability. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 283(3):R789–R797, 2002.

[112] S. K. Lam, A. Pitrou, and S. Seibert. Numba: A llvm-based python jit compiler. In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, LLVM '15, pages 1–6, New York, NY, USA, 2015. Association for Computing Machinery.

[113] H. J. Landau. Maximum entropy and maximum likelihood in spectral estimation. *IEEE Transactions on Information Theory*, 44(3):1332–1336, 1998.

[114] A. J. Lawrance and N. T. Kottegoda. Stochastic modelling of riverflow time series. *Journal of the Royal Statistical Society. Series A (General)*, 140(1):1–47, 1977.

[115] E. L. Lehmann and G. Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

[116] W. E. Leland, M. S. Taqqu, W. Willinger, and D. V. Wilson. On the self-similar nature of ethernet traffic. In *Conference Proceedings on Communications Architectures, Protocols and Applications*, SIGCOMM'93, pages 183–193, New York, NY, USA, 1993. ACM.

[117] P. J. Lenk. Towards a practicable Bayesian nonparametric density estimator. *Biometrika*, 78(3):531–543, 1991.

[118] D. Levin and Y. Peres. *Markov Chains and Mixing Times*. American Mathematical Society, 2017.

[119] L. Li and Z. Xie. Model selection and order determination for time series by information between the past and the future. *Journal of time series analysis*, 17(1):65–84, 1996.

[120] L. M. Li. Some notes on mutual information between past and future. *Journal of Time Series Analysis*, 27(2):309–322, 2006.

[121] K. Lindgren and M. G. Nordahl. Complexity measures and cellular automata. *Complex Systems*, 2(4):409–440, 1988.

[122] B. Liseo, D. Marinucci, and L. Petrella. Bayesian semiparametric inference on long-range dependence. *Biometrika*, 88(4):1089–1104, 2001.

[123] R. Liu, T. Di Matteo, and T. Lux. True and apparent scaling: The proximity of the Markov-switching multifractal model to long-range dependence. *Physica A: Statistical Mechanics and its Applications*, 383(1):35–42, 2007.

[124] S. Lowen and M. Teich. *Fractal-Based Point Processes*. Wiley Series in Probability and Statistics. Wiley, 2005.

[125] J. Luo and D. Guo. On the entropy rate of hidden Markov processes observed through arbitrary memoryless channels. *IEEE transactions on information theory*, 55(4):1460–1467, 2009.

[126] B. Mandelbrot. Self-similar error clusters in communication systems and the concept of conditional stationarity. *IEEE Transactions on Communication Technology*, 13(1):71–90, 1965.

[127] B. B. Mandelbrot and J. W. van Ness. Fractional Brownian motions, fractional noises and applications. *SIAM Review*, 10(4):422–437, 1968.

[128] B. B. Mandelbrot and J. R. Wallis. Noah, joseph, and operational hydrology. *Water resources research*, 4(5):909–918, 1968.

[129] D. M. Mason. The Hurst phenomenon and the rescaled range statistic. *Stochastic Processes and their Applications*, 126(12):3790–3807, 2016.

[130] A. I. McLeod, H. Yu, and Z. L. Krougly. Algorithms for linear time series analysis: With r package. *Journal of Statistical Software*, 23(5):1–26, 2007.

[131] S. Meyn, R. Tweedie, and P. Glynn. *Markov Chains and Stochastic Stability*. Cambridge Mathematical Library. Cambridge University Press, 2009.

[132] C. Nair, E. Ordentlich, and T. Weissman. Asymptotic filtering and entropy rate of a hidden Markov process in the rare transitions regime. In *Proceedings. International Symposium on Information Theory, 2005. ISIT 2005.*, pages 1838–1842. IEEE, 2005.

[133] I. Nemenman. Information theory and learning: a physical approach. 2000. arXiv preprint arXiv:0009032.

[134] B. Oğuz and V. Anantharam. Hurst index of functions of long-range-dependent Markov chains. *Journal of Applied Probability*, 49(2):451–471, 2012.

[135] E. Ordentlich and T. Weissman. Approximations for the entropy rate of a hidden Markov process. In *Proceedings. International Symposium on Information Theory, 2005. ISIT 2005.*, pages 2198–2202. IEEE, 2005.

[136] O. Ordentlich. Novel lower bounds on the entropy rate of binary hidden Markov processes. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 690–694. IEEE, 2016.

[137] D. S. Ornstein and B. Weiss. Entropy and data compression schemes. *IEEE Transactions on Information Theory*, 39(1):78–83, 1993.

[138] T. Owada and G. Samorodnitsky. Maxima of long memory stationary symmetric α-stable processes, and self-similar processes with stationary max-increments. *Bernoulli*, 21(3):1575–1599, 2015.

[139] E. Parzen. On choosing an estimate of the spectral density function of a stationary time series. *The Annals of Mathematical Statistics*, 28(4):921–932, 1957.

[140] E. Parzen. On consistent estimates of the spectrum of a stationary time series. *The Annals of Mathematical Statistics*, 28(2):329–348, 1957.

[141] Y. Peres and A. Quas. Entropy rate for hidden Markov chains with rare transitions. In *Entropy of Hidden Markov Processes and Connections to Dynamical Systems: Papers from the Banff International Research Station Workshop*, London Mathematical Society Lecture Note Series, pages 172?–178. Cambridge University Press, 2011.

[142] S. M. Pincus. Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences*, 88(6):2297–2301, 1991.

[143] A. N. Quas. An entropy estimator for a class of infinite alphabet processes. *Theory of Probability & Its Applications*, 43(3):496–507, 1999.

[144] P. Regnault. Plug-in estimator of the entropy rate of a pure-jump two-state Markov process. In *AIP Conference Proceedings*, volume 1193, pages 153–160. American Institute of Physics, 2009.

[145] M. Rezaeian. Hidden Markov process: a new representation, entropy rate and estimation entropy, 2006. arXiv preprint arXiv:0606114.

[146] J. A. Rice. *Mathematical Statistics and Data Analysis.* Belmont, CA: Duxbury Press., third edition, 2006.

[147] J. S. Richman and J. R. Moorman. Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology*, 278(6):H2039–H2049, 2000.

[148] M. Roughan. On the beneficial impact of strong correlations for anomaly detection. *Stochastic models*, 25(1):1–27, 2009.

[149] W. Rudin et al. *Principles of mathematical analysis*, volume 3. McGraw-hill New York, 1976.

[150] G. Samorodnitsky. Long range dependence. *Foundations and Trends in Stochastic Systems*, 1(3):163–257, 2006.

[151] G. Samorodnitsky. *Stochastic Processes and Long Range Dependence.* Springer Series in Operations Research and Financial Engineering. Springer International Publishing, 2016.

[152] G. Samorodnitsky et al. Extreme value theory, ergodic theory and the boundary between short memory and long memory for stationary stable processes. *The Annals of Probability*, 32(2):1438–1468, 2004.

[153] R. Schilling. *Measures, Integrals and Martingales.* Measures, Integrals and Martingales. Cambridge University Press, 2017.

[154] C. E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3):379–423, 1948.

[155] R. Shaw. *The Dripping Faucet as a Model Chaotic System.* Science frontier express series. Aerial Press, 1984.

[156] P. C. Shields. Entropy and prefixes. *Ann. Probab.*, 20(1):403–409, 1992.

[157] Y. G. Sinai. Self-similar probability distributions. *Theory of Probability & Its Applications*, 21(1):64–80, 1976.

[158] E. M. Stein and R. Shakarchi. *Complex analysis.* Princeton lectures in analysis. Princeton University Press, Princeton, N.J., 2003.

[159] P. Stoica and T. Sundin. On nonparametric spectral estimation. *Circuits, Systems and Signal Processing*, 18(2):169–181, 1999.

[160] C. C. Strelioff, J. P. Crutchfield, and A. W. Hübler. Inferring Markov chains: Bayesian estimation, model comparison, entropy rate, and out-of-class modeling. *Physical Review E*, 76(1):011106, 2007.

[161] M. S. Taqqu, W. Willinger, and R. Sherman. Proof of a fundamental result in self-similar traffic modeling. *ACM SIGCOMM Computer Communication Review*, 27(2):5–23, 1997.

[162] E. Timofeev. Statistical estimation of measure invariants. *St. Petersburg Mathematical Journal*, 17(3):527–551, 2006.

[163] F. Tobar. Bayesian nonparametric spectral estimation, 2018. arXiv preprint arXiv:1809.02196.

[164] F. Tobar, T. D. Bui, and R. E. Turner. Design of covariance functions using inter-domain inducing variables. In *NIPS 2015-Time Series Workshop*. Citeseer, 2015.

[165] F. Tobar, T. D. Bui, and R. E. Turner. Learning stationary time series using Gaussian processes with nonparametric kernels. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[166] N. F. Travers. Exponential bounds for convergence of entropy rate approximations in hidden markov models satisfying a path-mergeability condition. *Stochastic Processes and their Applications*, 124(12):4149–4170, 2014.

[167] F. G. Tricomi, A. Erdélyi, et al. The asymptotic expansion of a ratio of gamma functions. *Pacific Journal of Mathematics*, 1(1):133–142, 1951.

[168] J. Tukey. The sampling theory of power spectrum estimates. In *Symposium on Applications of Autocorrelation Analysis to Physical Problems. US Office of Naval Research*, pages 47–67, 1950.

[169] P. Tuominen and R. L. Tweedie. Subgeometric rates of convergence of f-ergodic Markov chains. *Advances in Applied Probability*, 26(3):775?798, 1994.

[170] R. K. Udhayakumar, C. Karmakar, and M. Palaniswami. Approximate entropy profile: a novel approach to comprehend irregularity of short-term hrv signal. *Nonlinear dynamics*, 88(2):823–837, 2017.

[171] C. Varotsos and D. Kirk-Davidoff. Long-memory processes in ozone and temperature variations at the region 60 s–60 n. *Atmospheric Chemistry and Physics*, 6(12):4093–4100, 2006.

[172] V. Vatutin and V. Mikhailov. Statistical estimation of the entropy of discrete random variables with a large number of outcomes. *Russian Mathematical Surveys*, 50(5):121–134, 1995.

[173] J. Veenstra. arfima. `https://github.com/JQVeenstra/arfima`, 2018.

[174] D. Veitch, A. Gorst-Rasmussen, and A. Gefferth. Why FARIMA models are brittle. *Fractals*, 21(02):1350012, 2013.

[175] S. Verdú. Empirical estimation of information measures: A literature guide. *Entropy*, 21(8):720, 2019.

[176] R. Vesilo. Long-range dependence of Markov renewal processes. *Australian & New Zealand Journal of Statistics*, 46(1):155–171, 2004.

[177] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[178] W. Willinger, M. S. Taqqu, W. E. Leland, and D. V. Wilson. Self-similarity in high-speed packet traffic: Analysis and modeling of ethernet traffic measurements. *Statist. Sci.*, 10(1):67–85, 1995.

[179] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson. Self-similarity through high-variability: statistical analysis of ethernet lan traffic at the source level. *IEEE/ACM Transactions on Networking*, 5(1):71–86, 1997.

[180] W. Willinger, M. S. Taqqu, and V. Teverovsky. Stock market prices and long-range dependence. *Finance and Stochastics*, 3(1):1–13, 1999.

[181] A. D. Wyner and J. Ziv. Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression. *IEEE Transactions on Information Theory*, 35(6):1250–1258, 1989.

[182] I. Yanovsky. *Real Analysis and Multivariable Calculus: Graduate Level Problems and Solutions*. UCLA, 2005.

[183] H. Yar and Z. Nikooravesh. Taylor expansion for the entropy rate of hidden Markov chains. *Journal of Statistical Research of Iran*, 7(2):103–120, 2011.

[184] G. H. Yari and Z. Nikooravesh. Estimation of the entropy rate of ergodic markov chains. *Journal of the Iranian Statistical Society*, 11(1):75–85, 2012.

[185] J. M. Yentes, N. Hunt, K. K. Schmid, J. P. Kaipust, D. McGrath, and N. Stergiou. The appropriate use of approximate entropy and sample entropy with short data sets. *Annals of biomedical engineering*, 41(2):349–365, 2013.

[186] J. Ziv and A. Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on information theory*, 23(3):337–343, 1977.

[187] O. Zuk, E. Domany, I. Kanter, and M. Aizenman. Taylor series expansions for the entropy rate of hidden Markov processes. In *2006 IEEE International Conference on Communications*, volume 4, pages 1598–1604, 2006.

[188] O. Zuk, I. Kanter, and E. Domany. The entropy of a binary hidden Markov process. *Journal of Statistical Physics*, 121(3):343–360, 2005.

[189] A. Zygmund and R. Fefferman. *Trigonometric Series*. Cambridge Mathematical Library. Cambridge University Press, 3rd edition, 2003.