Long range dependent models in information theory

Barlas Oguz

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2012-204
http://www.eecs.berkeley.edu/Pubs/TechRpts/2012/EECS-2012-204.html

October 23, 2012

Long range dependent models in information theory

by

Barlas Oğuz

A dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Electrical Engineering and Computer Sciences

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Venkat Anantharam, Chair
Professor Jean Walrand
Professor Jim Pitman

Fall 2012

Long range dependent models in information theory

by

Barlas Oğuz

Abstract

Long range dependent models in information theory

by

Barlas Oğuz

Doctor of Philosophy in Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Venkat Anantharam, Chair

Long range dependence refers to stochastic processes for which correlations persist at much longer time scales as compared to traditional models. For such processes the central limit theorem does not in general hold, and the smoothing effect of the law of large numbers takes more time to settle in. Such phenomena have been observed in many different fields including financial time series, DNA sequences, network traffic and variable bit-rate video. The bursty nature and persistent correlation structure of long range dependent processes make them tough to control and predict in practice, and tough to analyze in theory. In this thesis we look at the origins of long range dependence through the use of Markov models.

We first introduce a model of long range dependence using countable state Markov chains. A positive recurrent, aperiodic Markov chain is said to be long range dependent (LRD) when the indicator function of a particular state is LRD. This happens if and only if the return time distribution for that state has infinite variance. We investigate the question of whether other instantaneous functions of the Markov chain also inherit this property. We provide conditions under which the function

has the same degree of long range dependence as the chain itself. We illustrate our results through three examples in diverse fields: queuing networks, source compression, and finance. We then prove information-theoretic pointwise lossless source coding theorems for a class of sources constructed from this model. We are able to show that the code length process at the output of an encoder inherits the long range dependent nature of the source irrespective of the coding algorithm chosen. We extend our results to lossy source coding under suitable conditions, demonstrating quite generally the information-theoretic relevance of long range dependence.

Professor Venkat Anantharam
Dissertation Committee Chair

Dedicated to my dear wife.

Contents

List of Figures

Acknowledgements

the mentor

When I came to Berkeley, I had an image of the ideal professor - humble, with extraordinary ability and endless knowledge. I only realized later that I was very fortunate to not only have met such a person, but have him as my advisor. I thank Venkat for being everything you could expect from an advisor, setting high expectations, and giving me something to look up to.

the committee

Jean and Jim are two of the most positive people I have met in my life - their attitude towards their work, students, and life in general just makes me happy to be in academia. It's always a pleasure to work with good people, and I thank both of them for not only improving my work, but also improving my life.

the one

It is not easy for a relationship to outlive a Ph.D. I am indebted to my wife Yasemin for all her effort in lengthening and strengthening our relationship, and in shortening my Ph.D. My years here have been rendered much more memorable due to her existence.

the family

The everpresent support of family is like no other. I thank my mom, dad, and sister for always being a part of my life. I'm so used to their help that it's almost easy to forget that it's there.

special thanks to

Berkeley Büyükşehir Belediyespor

Bukalemun group

BUPS class of 2003

Pofican, the cat

Curriculum Vitæ

Barlas Oğuz

Education

2007	Bilkent University
	B.S., Electrical Engineering
2011	University of California, Berkeley
	M.S., Electrical Engineering and Computer Sciences
2012	University of California, Berkeley
	Ph.D., Electrical Engineering and Computer Sciences

Bio: the campus life

For some, going to college and living on campus is a life changing experience to be remembered with fondness and yearning for the remainder of time. I have been lucky enough to live on campus my entire life. Having spent 20+ years in Bilkent University in Ankara, Turkey, growing up and going to school, I came to Berkeley to continue enjoying the amenities that the campus has to offer: a forever young, vibrant, intellectual community; never ending opportunities for entertainment and personal development, and a convenient displacement from reality that presents one with a sense of timeless good will. Campus life is in some ways an abstraction of reality. Some regard this as an illusion. To me, it is a model. And like most models, it is prettier than the reality that it is aiming to describe. Hopefully, the rest of my life will follow this model, if not literally, at least in spirit.

Chapter 1

Introduction

1.1 Long range dependence

When we refer to *long range dependence*, or a random process that exhibits *long memory* we informally mean that the present behavior of the process is heavily dependent on the preceding values even going back to the distant past of the process. To turn this intuition into a mathematical definition, one needs to resolve two ambiguities. How do we quantify 'dependent', and how do we quantify 'distant'?

Distortion measures abound; answering the first question is a matter of picking the one that suits the application. Various mixing coefficients and information measures have been used. In applications where partial sums of stationary processes are of central interest, or when second order properties are most relevant, the simple covariance function is most common. This approach is also appropriate for our discussion, and so our choice of dependence measure will be the covariance. It might be argued that a better name for this definition would be *long range correlations* instead of *long range dependence* or *long memory*. This convention is indeed used in some places, however we will stick with the more standard terms.

Having picked a measure of dependence, it is now possible to discuss the *long* in *long memory*. For instance one might reasonably argue that a moving average process with window size 5 has longer memory than a moving average process with window size 2, or that an AR(1) process has longer memory than either, since the correlation function is always non-zero. In general one could say that a random process X has longer memory than random process Y whenever the covariance function of X asymptotically dominates that of Y.

While such approaches are feasible, our concern is not to think of long range dependence as an ordering, but as a classification of random processes. We want to divide the space of stationary random processes into two disjoint classes: long range dependent and not long range dependent, and we will refer to the latter more conveniently as *short range dependent* processes. The boundary line separating these two classes should represent a phase transition where a qualitative change in behavior takes place. In a way, the entire class of short range dependent processes should be akin to an i.i.d. process, having qualitatively similar characteristics, where those processes that cross the long range dependent line should exhibit behavior that is completely absent in the other class.

To pinpoint where such a phase transition happens, we turn to the central limit theorem. Let $\mathcal{F}_n = \sigma(X_n, n \leq m)$ and $||Y||_2 = \sqrt{EY^2}$.

Theorem 1.1.1. *(Central limit theorem, (18) thm. 7.6)*

Let (X_n) be a stationary sequence with $EX_0 = \mu$ and

$$\sum_{n \geq 1} ||E[X_0|\mathcal{F}_{-n}]||_2 < \infty.$$

Then

$$var(X_0) + 2\sum_{r=1}^{\infty} cov(X_0, X_r) := \sigma^2 < \infty,$$

and

$$\frac{\sum_{i=1}^{\lfloor nt \rfloor}(X_i - \mu)}{\sigma\sqrt{n}} \xrightarrow{d} B_t$$

where B_t is the standard Brownian motion.

What is interesting about this statement is that on the left hand side we have the centralized partial sums of a somewhat general stationary process with memory, but on the right hand side, we have Brownian motion, which is an *independent increments* process. In other words, the dependence in the original process has disappeared under the limit of scaling. Looking at the process at the level of larger and larger blocks, we see that the dependence between blocks becomes negligible because of the finite covariances condition, and the block-aggregated process looks like an i.i.d. process with variance $n\sigma^2$. We find it appropriate to call such ephemeral dependence as *short range.*

Now let's look at what happens when the finite correlations condition is violated:

$$\limsup_{n\to\infty} \quad \mathrm{var}(X_0) + 2\sum_{r=1}^{n} \mathrm{cov}(X_0, X_r) = \infty.$$

Clearly,

$$\frac{\sum_{i=1}^{\lfloor nt \rfloor} X_i}{\sqrt{n}}$$

does not have a proper distributional limit with finite variance as $n \to \infty$ for any $t > 0$. To see this, take e.g. $t = 1$ and note that

$$\mathrm{var}\left(\frac{\sum_{i=1}^{n} X_i}{\sqrt{n}}\right) = \frac{1}{n}\left(n\ \mathrm{var}(X_0) + \sum_{r=1}^{n} r\ \mathrm{cov}(X_0, X_r)\right) \to \infty.$$

As an example, take a Gaussian process (X_n) with $\mathrm{cov}(X_0, X_r) = r^{-\alpha}$ for $0 < \alpha < 1$, and with zero mean. (We can check that this is a valid covariance function by observing that it has a positive Fourier transform.) Denoting $S_n := \sum_{i=1}^{n} X_i$, we can

write

$$\begin{aligned}\operatorname{cov}(S_{\lfloor nt_1 \rfloor}, S_{\lfloor nt_2 \rfloor}) &= \sum_{i=1}^{\lfloor nt_1 \rfloor} \sum_{j=1}^{\lfloor nt_2 \rfloor} \operatorname{cov}(X_i, X_j) \\ &= \sum_{i=1}^{\lfloor nt_1 \rfloor} \sum_{j=1}^{\lfloor nt_2 \rfloor} |i-j|^{-\alpha} \\ &\sim n^{2-\alpha}(|t_1|^{2-\alpha} + |t_2|^{2-\alpha} - |t_1 - t_2|^{2-\alpha}),\end{aligned}$$

from which we deduce that to get a meaningful limit the proper scaling is

$$\frac{\sum_{i=1}^{\lfloor nt \rfloor}(X_i - \mu)}{(n)^{1-\frac{\alpha}{2}}} \xrightarrow{d} fB_t.$$

where fB_t is a Gaussian process with covariance function equal to $|t_1|^{2-\alpha} + |t_2|^{2-\alpha} - |t_1 - t_2|^{2-\alpha}$. This process is called *fractional Brownian motion* with (Hurst) parameter $1 - \frac{\alpha}{2}$.

Fractional Brownian motion is a Gaussian process with stationary increments. The increments process is called *fractional Brownian noise*, which is a Gaussian process with correlation function

$$R_{fBn}(r) = \frac{1}{2}(|r+1|^{-\alpha} - 2|r|^{-\alpha} + |r+1|^{-\alpha}) \sim r^{-\alpha} \text{ as } r \to \infty.$$

We see that the limiting increments process has a similar covariance function to the original process. In particular, the dependence has not disappeared under scaling. This behavior is in sharp contrast to the finite correlations case, and we deem it appropriate to refer to it as long range dependence.

Definition 1.1.2. *(29) A stationary real valued random process (X_n) is said to be* ***long range dependent*** *whenever*

$$\limsup_{n \to \infty} \quad \sum_{r=1}^{n} cov(X_0, X_r) = \infty.$$

See also (29) for variants on second order definitions of long range dependence. As suggested earlier, we will use *long range dependence* and *long memory* interchangeably.

1.1.1 Regular variation, heavy tails

The most typical correlation function which satisfies 1.1.2 is a *regularly varying* function:

$$R(r) = r^{-\alpha} L(r),$$

where $L(r)$ is a slowly varying function, i.e.

$$\lim_{n\to\infty} \frac{L(cn)}{L(n)} \to 1 \text{ for any } c > 0.$$

In this case $0 < \alpha < 1$ implies long memory. While the treatment in this thesis will be at the level of generality of definition 1.1.2, it is helpful to think about the results in terms of regularly varying functions. Some of the examples will make use of this definition.

We will refer to random variables with regularly varying distributions as *heavy tailed*. This term is used in some places to refer to any distribution which decays slower than an exponential. We adopt a narrower definition which further requires infinite variance.

Definition 1.1.3. *A random variable is* ***heavy tailed*** *if the cumulative distribution can be written as*

$$F_X(t) = 1 - t^{-\alpha+1} L(t),$$

for some slowly varying function L, and $E[X^2] = \infty$.

Here again, $0 < \alpha < 1$ implies infinite variance.

Heavy tailed distributions and long range dependence go hand in hand. For instance, a renewal process with heavy tailed inter-arrival times will be long range dependent (2.1.1). A single server queue with i.i.d. heavy tailed service times will have a long range dependent busy-idle process (section 2.4). Conversely, a long range dependent processes will cause heavy-tailed waiting times at a queue.

1.1.2 Self similarity and the Hurst index

As we saw, the definition of long range dependence is motivated around scaling laws for random processes in the form $X(t) \rightarrow \frac{1}{n^H} X(nt)$. Distributions that are stationary points of such scalings are referred to as *self similar* laws. The parameter H is the index of self similarity.

Brownian motion is the unique process with stationary increments that is self similar with parameter $H = \frac{1}{2}$. Fractional Brownian motion is the unique Gaussian process with stationary increments that is self similar with parameter $0 < H < 1$ ((54), 7.2). Due to their natural appearance in central limit type theorems, fractional Brownian motions have been the single most popular continuous time model for long range dependence in the literature. Other self similar processes with stationary increments are α-stable Lévy processes ((54), 7.5).

Self similarity can also be defined for deterministic functions, when they are referred to as fractals, which are functions that are invariant under similar joint scaling of time and space. For this reason, self similar processes are sometimes also referred to as *fractal* processes.

While it is possible to discuss long range dependence without reference to self similarity, as a result of these connections and historical coupling of their development, the two fields have come to be closely associated with each other.

The self similarity parameter can be alternatively defined in terms of the index of the scaling law which governs the variance of the partial sums of (X_n). Let $S_n = \sum_{i=1}^{n} X_i$. If S_n is self similar, then we know $\frac{S_n}{n^H}$ has a meaningful limit as $n \rightarrow \infty$. In particular, we have that

$$0 < \lim_{n \rightarrow \infty} \frac{\mathrm{var}(S_n)}{n^{2H}} < \infty.$$

For short range dependent processes, it can easily be verified that the variance of

S_n scales at most linearly with n, therefore this limit will exist if we set $H = \frac{1}{2}$. A higher H signifies a faster scaling of $\text{var}(S_n)$, caused by 'longer' correlations in (X_n). Thus the scaling index H can be regarded as a measure of long memory. Properly, we define

Definition 1.1.4. *(9) Let the* ***Hurst index*** H *($0 \leq H \leq 1$) be defined as*

$$H := inf \left\{ h : \limsup_{n\to\infty} \frac{var(\sum_{i=1}^{n} X_i)}{n^{2h}} < \infty \right\}.$$

While short range dependent processes all have Hurst index $\leq \frac{1}{2}$, the converse is not true. This is because the Hurst index only defines the polynomial order of the growth of $\text{var}(S_n)$, i.e. up to slowly varying terms. To avoid border cases, we will sometimes assume $H > \frac{1}{2}$.

A Hurst index lower than $\frac{1}{2}$ is possible, for instance in cases where the sum of the absolute correlations diverge, nevertheless the signed sum remains finite. Take for example a $\{-1, 1\}$ valued process where $X_{n+1} = -X_n$, $P(X_0 = 1) = \frac{1}{2}$. This process has Hurst index 0, since $\text{var}(S_n)$ is always bounded. Again, we will not concern ourselves with such processes in the remainder of this thesis. For our purposes, these negatively correlated processes are short range dependent.

While it may be somewhat restrictive to define Hurst index for only processes of finite variance, this will be adequate for our applications of interest where this is often a natural assumption (e.g. network traffic has a bounded bit-rate). For a more general discussion of self-similarity, the reader is referred to (54).

1.2 History and applications

The history of long range dependence starts with the studies of the hydrologist Harold Edwin Hurst (1880-1978). Hurst investigated historical rainfall data and oc-

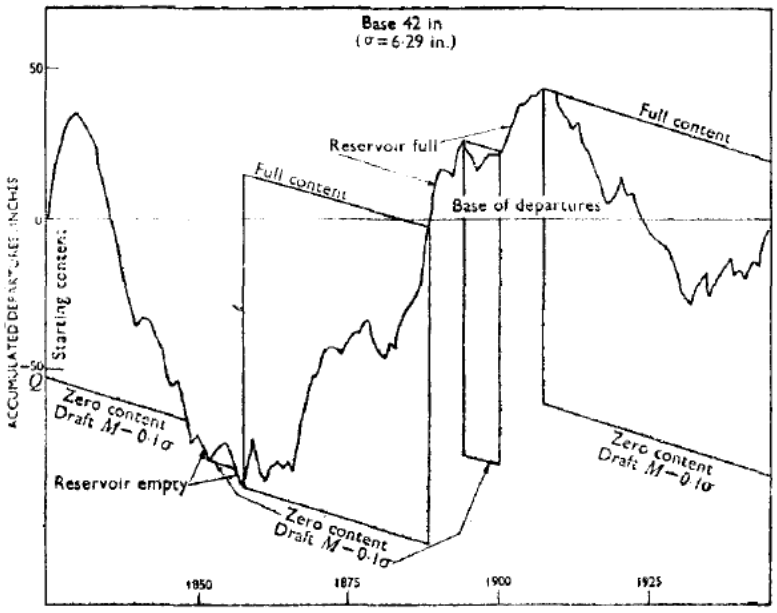


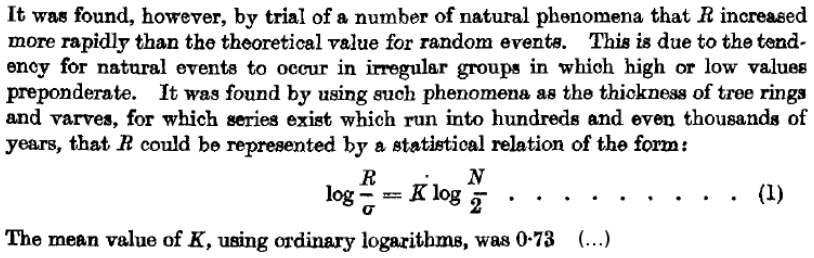
FIG. 1.—NEW YORK RAINFALL. ACCUMULATED DEPARTURES

Figure 1.1. Accumulated rainfall, New York. (30)

cupancy of water reservoirs on rivers in the hope of regulating water storage to avoid droughts and floods. Figure 1.1, taken from his 1956 paper (30) shows the storage levels which resulted from New York rainfall data. Hurst realized that this data series shows much more variability than would be expected if annual rainfall was an independent series.

Denoting by R, the range of this data (the difference of the max and min storage level) over N years, Hurst postulated that $\log R$ and $\log N$ are linearly related with slope K (fig. 1.2).

For a short range dependent series, one would expect R to scale as $\sqrt{N}$, suggesting $K = \frac{1}{2}$. However, Hurst notes that the mean value of K is in fact 0.73. He also noted in this paper that the variances of the accumulated rainfalls is growing faster than what could be explained by a short range dependent model. The observation has dire implications for storage planning, in that the minimum reservoir capacity needed to



It was found, however, by trial of a number of natural phenomena that R increased more rapidly than the theoretical value for random events. This is due to the tendency for natural events to occur in irregular groups in which high or low values preponderate. It was found by using such phenomena as the thickness of tree rings and varves, for which series exist which run into hundreds and even thousands of years, that R could be represented by a statistical relation of the form:

$$\log \frac{R}{\sigma} = K \log \frac{N}{2} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad (1)$$

The mean value of K, using ordinary logarithms, was 0·73 (...)

Figure 1.2. First appearance of a Hurst index. (30)

avoid drought and floods for a given time horizon is many times larger than what would ordinarily be needed.

This was the first of many observations that showed long range dependence occurring in natural time series. Subsequent work demonstrated that this phenomenon is not limited to the field of hydrology, but in fact very common in financial series, network traffic and variable bit-rate multimedia data streams.

1.2.1 Long memory in financial time series

The presence of persistent correlations in financial data first came to light in the work of Granger (25) who noted that low frequency components were typically dominant in the empirical power spectrum of economics data. This was interpreted as the data having a 'trend' in the mean. It was not until the work of Mandelbrot (40)(41)(42) however that the concept of long range dependence was popularized as a modeling tool for financial time series.

While it is reasonable to expect that the prices of commodities will show strong correlations over time, it is somewhat surprising that the returns on speculative assets would have long memory. In fact, the price of a publicly traded good is assumed to be arbitraged by the market so that the past returns do not have any value in predicting the future returns. As a result, the aggregate returns are well modeled by a martingale

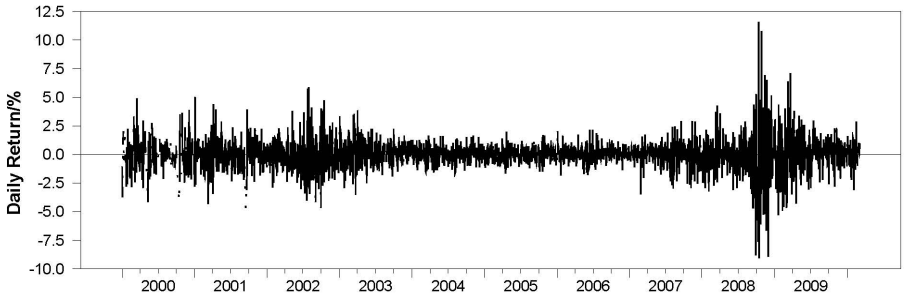


Figure 1.3. Percent returns of S&P 500. (23)

process, with next to zero correlations. The series of price returns is therefore not long range dependent.

Arbitraging therefore erodes correlations in the data, making long range dependence, which we defined solely in terms of correlations, disappear. Disappearing correlations however, does not mean disappearing dependence. In fact the dependence remains and shows up as long range dependence in the series of absolute returns. Figure 1.3 plots the percent daily returns of the stock market (S&P 500) over a decade (23). We see that the series looks like noise, but with varying amplitude. This is a typical martingale sequence, with dependence showing up in the second order statistics.

This kind of behavior can be directly modeled by a generalized autoregressive conditional heteroskedasticity (GARCH, (8)) model, where (X_n) is a zero mean sequence which is independent conditioned on the variance sequence. The variance (σ_n^2) can be based on an autoregressive moving average (ARMA(p,q)) model:

$$\sigma_n^2 = \alpha_0 + \sum_{i=1}^{q} \alpha_i \epsilon_{n-i} + \sum_{j=1}^{p} \beta_j \sigma_{n-j}^2,$$

resulting in a martingale sequence with persistent volatility.

Such parametric models are useful for inference, and have been employed in practice. However, if we want to explain the observed behavior, rather than just model it,

this approach falls short. For this, we can think of the price returns as resulting from an operation (arbitraging by the market) performed on some underlying long range dependent process. Then we can see how long range dependent volatility emerges from this operation, and infer the characteristics of the resulting process from the behavior of the underlying process.

We take this approach using a simple example in section 2.6. Our construction is based on long range dependent Markov chains, the theory for which is developed in chapter 2.

1.2.2 Long memory in network traffic

Note that the water reservoirs studied by Hurst are mathematically similar to a queue. Rainfall corresponds to incoming packets to a queue, while draft corresponds to service rate. A drought and flood correspond to empty or overflowing buffers respectively. While these events may not have the same drastic consequences, they are still undesirable, since overflowing buffers mean lost packets and empty buffers mean lost service capacity. In practice, network engineers aim to minimize the probability of these events happening by picking appropriate buffer sizes, leading to many of the same issues that Hurst faced in looking for the optimum reservoir size. Queuing networks form the basis for modeling communication systems, and interestingly, communication flows in these networks turn out to have many statistical similarities with water flow in rivers.

Interest in LRD processes in communication networks was sparked by several empirical observations that showed such distributions were characteristic of network traffic on the internet (36),(13),(49). Due to the fundamentally different qualities of LRD processes mentioned in the first section, these discoveries have important, and often negative consequences for the modeling and analysis of communication networks.

Among these are different asymptotics for queue sizes and packet drop probabilities (51; 38; 37; 28; 63; 20), and a need for new optimal schedulers (2),(48),(53).

The mostly degrading effect of LRD traffic in networks has led to research efforts for understanding the mechanisms by which such traffic is generated and whether preventive measures are possible (48),(13). For instance, in a network of queues with heterogeneous arrival traffic, one might be interested in scheduling long range dependent traffic differently than short range dependent traffic. The choice of scheduling strategy effects how the different flows get coupled, and to what extent the short range dependent traffic is affected by the presence of long range dependence in the network.

We will again illustrate the use of long range dependent Markov models in the setting of queuing networks. In section 2.4 we discuss a simple queuing network of two parallel queues, one of them being driven by a process with long memory. We will show that under a fixed rate shared server with longest queue first scheduling, long range dependence will spread so that the busy-idle process of both queues will become long range dependent, (see also (43)).

1.2.3 Variable-bit-rate video

Variable-bit-rate traffic (mainly VBR video) is an important component of internet traffic. In the hope of understanding such traffic better, there has been considerable work on analyzing traces of VBR video ((5; 22; 52; 21) to cite a few). The common observation that is the culmination of this work is that long range dependence is omnipresent in VBR traffic, and persists across a wide variety of codecs. Coupled with the discussion in the preceding section, this observation might shed some light on why network traffic exhibits long memory.

Consider the plot in figure 1.4. The plot shows the number of bytes per frame that

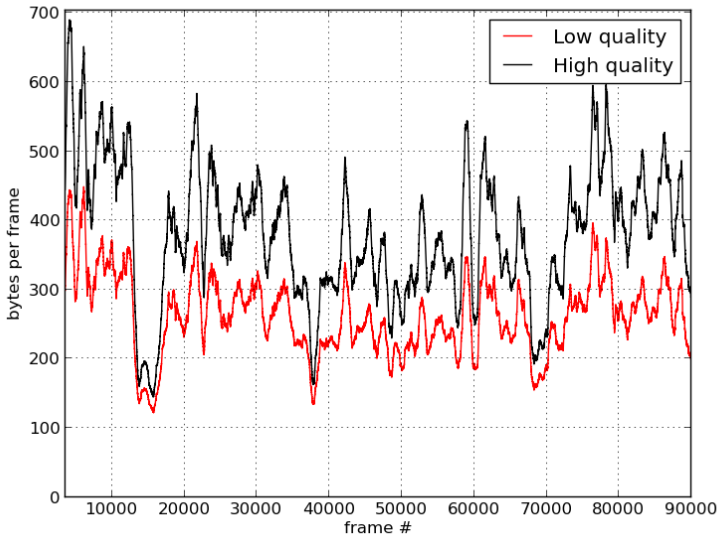


Figure 1.4. Bytes per frame resulting from MPEG4 encoding of *Star Wars IV:A New Hope* (21).

was needed to encode a 1 hour long segment of the movie *Star Wars IV* using `MPEG4` compression at two different distortion (quality) levels [1]. The immediate observation to be made is that the traces at different distortion levels have roughly the same shape. It should not come as a surprise that both of these traces were estimated (using R/S statistics) to have identical Hurst index of around 0.75 (21).

Allowing more distortion does not seem to reduce long range dependence. Is this fact dependent on the choice of encoder, or a universal property of video traces? Are there encoders that can reduce or eliminate long range dependence regardless of the choice of distortion level? These are practical questions which may have implications for encoder design and bandwidth management. They are also fundamental questions that ask whether long range dependence is an intrinsic property of some information sources. We can attempt to answer such questions within the framework of information theory.

[1] Data smoothed over 500 frames. Trace taken from `http://http://www-tkn.ee.tu-berlin.de/research/trace/ltvt.html`

1.3 Markov models and information theory

As we mentioned, in much of the prior work analyzing communication networks, the distribution of network traffic at the source is given a priori. In the continuous time setting the most popular model that is used is the fractional Brownian motion (fBm) (see (54), chapter 7.2). In the discrete time case fractional ARIMA models have been widely adopted (see e.g. (4), chapter 2.5). Although parametric models such as these have their advantages in terms of model fitting and estimation, in many cases they can only provide an approximation to the underlying system. Here we will work with models based on countable state, long range dependent Markov chains, which is a much more flexible class of models. We might want to model network traffic, which is usually created at the output of an algorithm, that involves coding of an information source. The traffic could for example be a stream of variable-bit-rate video, as discussed in the preceding section.

Motivated by examining what effects encoding algorithms might have on the long range dependence of the compressed bit rate process, we will prove source coding theorems about information sources that can be represented in terms of long range dependent Markov chains. The fundamental theorem of source coding, due to Shannon (56), says that the average bit-rate needed to represent an information source cannot be smaller than the entropy rate of that source. Furthermore, optimal source codes achieve on average the entropy rate. The work of Kontoyiannis (33; 35) attempts to find similar fundamental bounds on the bit-rate process, but on the level of second order statistics. In other words, what is the minimum variability the bit-rate process can have, given that the average bit-rate is equal to the entropy rate? Results are known mainly for i.i.d. sources and certain fast mixing sources. We pick up this question for long range dependent processes, also providing partial answers for lossy coding. We will show quite generally that independent of the choice of en-

coder and distortion level, 'optimal' encoders preserve the Hurst index of the original information source.

In one line, this thesis is about developing Markov chain models of long range dependence, with applications in information theory. The models are described in the next chapter, along with several applications to diverse fields. The information theory results are explained in chapter 3.

Chapter 2

Long range dependent Markov models

2.1 Introduction

A stationary random process (X_n) with $E[X_n^2] < \infty$ is said to be long range dependent (LRD) if

$$\limsup_{n\to\infty} \sum_{r=1}^{n} \text{cov}(X_0, X_r) = \infty.$$

The degree of long range dependence is measured by the Hurst index H ($\frac{1}{2} \leq H \leq 1$).

$$H := inf \left\{ h : \limsup_{n\to\infty} \frac{\sum_{r=1}^{n} \text{cov}(X_0, X_r)}{n^{2h-1}} < \infty \right\}.$$

Equivalently, we can write

$$H := inf \left\{ h : \limsup_{n\to\infty} \frac{\text{var}(\sum_{i=1}^{n} X_i)}{n^{2h}} < \infty \right\}.$$

Take (M_n), a positive-recurrent, aperiodic, discrete time, countable state Markov chain. We will take the state space to be the natural numbers $\mathbb{N}$, without loss of generality. The chain is in stationarity with stationary distribution π. We will now

define a notion of long range dependence for such chains. Since the choice of $\mathbb{N}$ as the state space is a mere convenience, to apply the regular definition of long range dependence directly to M_n would be quite arbitrary. For a more usable definition, we first turn to a simpler long range dependent process.

2.1.1 Long range dependent renewal process

Take a discrete time, stationary renewal process $(X_n) \in \{0, 1\}$, characterized by the inter-arrival time distribution $T \sim \mathbf{F}(t)$. Here $\mathbf{F}(t) := P(T \leq t)$. We define the moment index κ of this distribution as

$$\kappa := \sup\{k : E[T^k] < \infty\}.$$

The following theorem of Daley (15) relates the Hurst index of the renewal process to the moment index of the inter-arrival time distribution, in the case when (X_1^n) is long range dependent.

Theorem 2.1.1. *A stationary renewal process with inter-arrival time distribution function* $\mathbf{F}(t) := P(T \leq t)$ *which has* $\sum_{t=1}^{\infty} t(1-\mathbf{F}(t)) = \infty, \sum_{t=1}^{\infty}(1-\mathbf{F}(t)) < \infty$ *and moment index* κ*, is long-range dependent and has Hurst index* $H = \frac{1}{2}(3-\kappa)$*.*

In particular, the renewal process is long range dependent if and only if the inter-arrival time has infinite second moment.

Using the fact that an indicator function of a state of a Markov chain defines a renewal process, we can attempt to define long range dependence for Markov chains through the long range dependence of its indicator functions. Note that the Hurst index of a renewal process has a one-to-one correspondence with the moment index of its inter-arrival distribution. Recalling that, in an irreducible Markov chain, the moment index of the return time to a state is identical for each state in the chain (10), we conclude that the Hurst index of the indicator function $1(M_n = i)$ of state i

of a Markov chain is a class property (9). Morover, the indicator function $1(M_n = i)$ is LRD if and only if indicator functions of every state is LRD (9). Thus we adopt the following natural, consistent definition:

Definition 2.1.2. *A positive-recurrent, aperiodic, discrete time Markov chain $M_n \in \mathbb{N}$ is said to be* long range dependent *iff the indicator function $1(M_n = i)$ is long range dependent for every i. The common Hurst index H of all such indicator functions is said to be the Hurst index of the chain.*

2.1.2 Functions of a Markov chain

In (9) it is proved that a Markov chain is LRD if and only if the return time distribution of any state has infinite variance. It is also argued that finite weighted sums of indicator functions on this chain also inherit this property. It is natural to conjecture that this might be true for all functions of the chain. However, this conjecture is easily disproved, most easily by considering a constant function (also see the two counter examples in (9)). It is then of considerable interest to find which functions of an LRD Markov chain are also LRD.

Let $\varrho_n = \rho(M_n)$ be an L_2 function of M_n. In this chapter, we provide conditions under which one can infer the long range dependence of (ϱ_n) from that of (M_n).

It is instructive to consider the case where $\varrho_n = 1(M_n = i)$, an indicator function. We can write

$$\sum_{r=1}^{n} \mathrm{cov}(\varrho_0, \varrho_r) = \pi_i \sum_{r=1}^{n} (p_{ii}^{(r)} - \pi_i) =: \pi_i Q_{ii}^{(n)}.$$

Here $p_{ii}^{(r)}$ is the r-step return probability to state i.

Note that $p_{ii}^{(r)} \to \pi_i$, since the chain is ergodic, and the difference $(p_{ii}^{(r)} - \pi_i)$ represents how far the chain is from stationarity. In a finite state chain, these differences would decay exponentially to zero, and we would have $\lim_{n\to\infty} Q_{ii}^{(n)} < \infty$. In the

long range dependent case, we have $Q_{ii}^{(n)} \to \infty$ (9). In fact, when the return time distribution satisfies $P(T > t) \sim t^{-\alpha}$, for $1 < \alpha \leq 2$, we will have (see example 1 in (15))

$$\sum_{r=1}^{n} \mathrm{cov}(\varrho_0, \varrho_r) \sim Q_{ii}^{(n)} \sim n^{2-\alpha}.$$

Since $\mathrm{var}(\sum_{r=1}^{n} \varrho_r) = \sum_{r=1}^{n} \sum_{s=1}^{n} \mathrm{cov}(\varrho_r, \varrho_s) - n\mathrm{var}(\varrho_0)$, we can read off the Hurst index in this case easily as being $H = \frac{1}{2}(3 - \alpha)$, recovering the earlier result, since α is equal to the moment index of T in this case.

Now let us consider a slightly more complicated function, composed of a finite sum of indicator functions:

$$\varrho_n = \sum_{i=1}^{K} \rho(i) 1(M_n = i).$$

Then the above expression becomes,

$$\begin{aligned} \sum_{r=1}^{n} \mathrm{cov}(\varrho_0, \varrho_r) &= \sum_{r=1}^{n} \sum_{i=1}^{K} \sum_{j=1}^{K} \rho(i)\rho(j)\pi_i(p_{ij}^{(r)} - \pi_j) \\ &= \sum_{i=1}^{K} \sum_{j=1}^{K} \pi_i \rho(i)\rho(j) Q_{ij}^{(n)}. \end{aligned}$$

where we defined $Q_{ij}^{n} := \sum_{r=1}^{n}(p_{ij}^{(r)} - \pi_j)$. Now dividing both sides by $Q_{11}^{(n)}$,

$$\frac{\sum_{r=1}^{n} \mathrm{cov}(\varrho_0, \varrho_r)}{Q_{11}^{(n)}} = \sum_{i=1}^{K} \sum_{j=1}^{K} \pi_i \rho(i)\rho(j) \frac{Q_{ij}^{(n)}}{Q_{11}^{(n)}}. \tag{2.1}$$

It turns out that, since the quantities $\sum_{r=1}^{n} p_{ij}^{(r)}$ asymptotically behave similarly for each i and j (see (10) corollary 2 to theorem 9.4), $\frac{Q_{ij}^{(n)}}{Q_{11}^{(n)}}$ has a finite, non-zero limit as $n \to \infty$ (9):

$$\lim_{n\to\infty} \frac{Q_{ij}^{(n)}/\pi_j}{Q_{11}^{(n)}/\pi_1} = 1. \tag{2.2}$$

Taking a limit as $n \to \infty$ in 2.1, and comparing with the result for the indicator function, we see that for the two cases, the quantity $\sum_{r=1}^{n} \mathrm{cov}(\varrho_0, \varrho_r)$ is asymptotically equivalent, up to a constant. Thus, in the slightly more general case of compound

indicator functions, the conclusion remains that the Hurst index matches that of the underlying Markov chain. It is tempting to attempt to generalize the above argument for arbitrary functions. However, the difficulty is that the limit in 2.2 is unfortunately not uniform in i and j, and therefore we cannot justify exchanging the double sum in i and j with the limit in n in 2.1, when the double sum has infinitely many terms. In this chapter, we work around this limitation under fairly general conditions on ϱ.

The main result, given in section 2.3, provides a technical condition under which the rate of growth of $\sum_{r=1}^{n} \mathrm{cov}(X_0, X_r)$ is identical for $X_n = \varrho_n$ and $X_n = 1(M_n = i)$. We set up the proof with a collection of lemmas presented in section 2.8. For convenience, most of the notation is collected together in section 2.2.

There are many interesting scenarios where such a theorem might be useful. In the second half of the chapter, we collect three such examples. Section 2.4 discusses a simple queuing network of two parallel queues. One queue is driven by an LRD process, whereas the other one is driven by a short range dependent process. We model the inputs and queue lengths by countable state Markov chains, and show that under longest queue first scheduling both queues are LRD.

An example from information theory is given in section 2.5, where we re-prove a recent result in the source coding of LRD sequences (45). We show that the code length process of any lossless encoder which is compressing an LRD renewal process must dominate an LRD process with the same Hurst index as the source process. This example is a precursor to the more general results that will be presented in chapter 3.

The last example is about long range dependence in financial series. We discuss how the model can explain the LRD behavior observed in some instantaneous functions of the absolute returns of some asset.

2.2 Notation and setup

(M_n) is a positive-recurrent, discrete time, countable state Markov chain with state space $\mathbb{N}$ and stationary distribution π_i, $i \in \mathbb{N}$. Most of the notation we use is borrowed from (10).

$\rho : \mathbb{N} \to \mathbb{R}$ is such that $\sum_{i\in\mathbb{N}} \rho(i)^2\pi_i < \infty$.

$\varrho_n := \rho(M_n)$.

$\mu := \sum_i \rho(i)\pi_i$, is the mean of ρ.

$p_{ij}^{(n)} := P(M_n = j|M_0 = i), n \geq 0$, is the n-step transition probability from i to j.

${}_kp_{ij}^{(n)} := P(M_n = j; M_l \neq k, 0 < l < n|M_0 = i), n > 0$, is the n-step transition probability from i to j with taboo state k.

${}_kp_{ij}^{*} := \sum_{n=1}^{\infty} {}_kp_{ij}^{(n)}$.

${}_{\mathcal{H}}p_{ij}^{(n)} := P(M_n = j; M_l \notin \mathcal{H}, 0 < l < n|M_0 = i), n > 0$, is the n-step transition probability from i to j with taboo set $\mathcal{H}$.

${}_{\mathcal{H}}p_{ij}^{*} := \sum_{n=1}^{\infty} {}_{\mathcal{H}}p_{ij}^{(n)}$.

$f_{ij}^{(n)} := {}_jp_{ij}^{(n)}, n > 0$.

$Q_{ij}^{(n)} := \sum_{r=1}^{n}(p_{ij}^{(r)} - \pi_j), n > 0$.

$R_{ij}^{(n)} := \sum_{r=1}^{n} Q_{ij}^{(r)}, n > 0$.

$T_j := \inf_t\{t > 0 : M_t = j\}$ is the first time to state j at stationarity.

$m_{ij} := E_i[T_j]$ is the mean time to state j starting from i.

$H := inf\left\{h : \limsup_{n\to\infty} \frac{\mathrm{var}(\sum_{i=1}^{n} \mathbf{1}_{(M_i=1)})}{n^{2h}} < \infty\right\}$, the Hurst index of (M_n).

$H_\varrho := inf\left\{h : \limsup_{n\to\infty} \frac{\mathrm{var}(\sum_{i=1}^{n} \varrho_i)}{n^{2h}} < \infty\right\}$, the Hurst index of (ϱ_n).

To understand the results in the next section, it is useful to know the following properties:

Lemma 2.2.1. *For an LRD Markov chain,*

$$\lim_{n\to\infty} Q_{ij}^{(n)} = \infty, \tag{2.3}$$

$$\lim_{n\to\infty} \frac{R_{ij}^{(n)}}{n} = \infty, \tag{2.4}$$

$$\lim_{n\to\infty} \frac{Q_{ij}^{(n)}/\pi_j}{Q_{11}^{(n)}/\pi_1} = 1. \tag{2.5}$$

Proof. (2.5) is eq. 8 in (9). (2.3) follows from eqs. 8 and 5 of (9). (2.4) follows from (2.3).

□

We will assume henceforth that n is large enough s.t. $Q_{11}^{(n)}, R_{11}^{(n)} > 1$.

2.3 Main results

Theorem 2.3.1. *Let*

(condition 1)

$$\lim_{n\to\infty} \frac{1}{Q_{11}^{(n)}/\pi_1} \sum_{r=1}^{n} \sum_{i,j} \pi_i(\rho(i)-c)(\rho(j)-c)_{\mathcal{H}} p_{ij}^{(r)} = 0$$

for some constant c, and non-empty, finite set $\mathcal{H}$, and

(condition 2)

$$\lim_{L\to\infty} \limsup_{n\to\infty} \frac{1}{Q_{11}^{(n)}/\pi_1} \sum_{r=1}^{n} \sum_{i,j} \pi_i|\rho(i)\rho(j)|1(|\rho(i)|>L, |\rho(j)|>L)_{\mathcal{H}} p_{ij}^{(r)} = 0$$

Then,

$$\lim_{n\to\infty} \frac{var(\sum_{r=1}^{n} \varrho_i)}{R_{11}^{(n)}/\pi_1} = (\mu - c)^2.$$

Moreover, if $c \neq \mu$, then $H_\varrho = H$.

Some remarks about the conditions are in order.

1. They fail to hold if $\lim_i \rho(i)$ exists and is not c. This shows that $\lim_i \rho(i)$ is the unique choice for c in this case.

2. They will hold whenever $\lim_i(\rho(i) - c) = 0$. Specifically when $(\rho(i) - c) = 0$ for i greater than some value.

   Both of these can be seen as direct consequences of lemma 2.8.6, which is stated later in section 2.8.

3. They are implied by the considerably stronger condition

$$\frac{1}{Q_{11}^{(n)}/\pi_1} \sum_{r=1}^{n} \sum_{i,j} \pi_i |\rho(i) - c||\rho(j) - c|_1 p_{ij}^{(r)} \to 0.$$

4. Choice of $\mathcal{H}$ is arbitrary, and we will often just pick $\mathcal{H} = \{1\}$. This is due to lemma 2.8.9

Condition 2 is trivially satisfied for bounded functions. When ϱ_n are not bounded, condition 2 ensures that they can be truncated without affecting the long range dependence discussions.

In light of remark 2, c can be interpreted as a 'limiting mean', in a weak sense, of ϱ as the return time to the compact set $\mathcal{H}$ becomes large. The deviance of ϱ from its average behavior in this limiting regime, given by $(\mu - c)$ determines the limiting constant in the statement of the theorem. When $(\mu - c)^2 = 0$, the behavior of ϱ is similar to its average behavior even when M_n takes a long excursion before returning to $\mathcal{H}$. Therefore the long range dependence of M might not exhibit itself in ϱ. ϱ might have a lower Hurst index in this case, or even be short range dependent. What happens exactly depends on the detailed structure of M and ρ, and cannot be captured by our formulation which only investigates the asymptotics at the scale of the Hurst index of M. In this regard, $(\mu - c)^2 > 0$ is necessary for ϱ to be LRD

at the same scale as M, and examples can easily be constructed where ϱ that fail this condition fail to be LRD to the same degree. We give one such non-example in section 2.7.

The following theorem extends the usefulness of the preceding theorem considerably. It describes the case, when the state space of the Markov chain is divided into a finite number of subsets, with communication between the sets happening almost only through a finite set of states $\mathcal{H}$. The canonical example for such a structure would be the Markov chain representation of a semi-Markov process given by the pair (S,T), where S is described by a finite state Markov chain and T is the time since the last transition, having an arbitrary distribution with $E[T] < \infty$. In this case, the state space would be divided into sets $\{S = k\}$, and transition between sets is only possible by visiting $(S, 0)$.

Theorem 2.3.2. *Let $\{\mathcal{A}_k\}$, $1 \le k \le K$, be a finite partition of the state space $\mathbb{N}$. (condition 1) Let $\mathcal{H}$ be a non-empty finite set, and*

$$\lim_{n\to\infty} \frac{1}{Q_{11}^{(n)}/\pi_1} \sum_{r=1}^{n} \sum_{i\in\mathcal{A}_k, j\in\mathcal{A}_l} \pi_i |\rho(i)-\mu||\rho(j)-\mu|_{\mathcal{H}} p_{ij}^{(r)} = 0, \quad \forall k \neq l.$$

Also suppose $\pi_{\mathcal{A}_k}^{\infty} := \lim_{n\to\infty} \frac{\sum_{i,j\in\mathcal{A}_k} \pi_i \sum_{r=1}^{n} {}_1p_{ij}^{(r)}}{\sum_{i,j} \pi_i \sum_{r=1}^{n} {}_1p_{ij}^{(r)}}$ exists $\forall k$. Let there exist constants $c_k, 1 \le k \le K$, such that

(condition 2)

$$\lim_{n\to\infty} \frac{1}{Q_{11}^{(n)}/\pi_1} \sum_{r=1}^{n} \sum_{i,j\in\mathcal{A}_k} \pi_i (\rho(i)-c_k)(\rho(j)-c_k)_{\mathcal{H}} p_{ij}^{(r)} = 0 \quad \forall k,$$

and

(condition 3)

$$\lim_{L\to\infty} \limsup_{n\to\infty} \frac{1}{Q_{11}^{(n)}/\pi_1} \sum_{r=1}^{n} \sum_{i,j\in\mathcal{A}_k} \pi_i |\rho(i)\rho(j)| 1(|\rho(i)| > L, |\rho(j)| > L)_{\mathcal{H}} p_{ij}^{(r)} = 0 \quad \forall k.$$

Then,

$$\lim_{n\to\infty} \frac{var(\sum_{r=1}^{n} \varrho_i)}{R_{11}^{(n)}/\pi_1} = \sum_{k=1}^{K} \pi_{\mathcal{A}_k}^{\infty}(\mu - c_k)^2.$$

Moreover, if $\pi_{\mathcal{A}_k}^{\infty}(c_k - \mu) \neq 0$ *for some* k*, then* $H_\varrho = H$*.*

Remark. If $c_k = c_l$ *for a pair of subsets* $\mathcal{A}_k, \mathcal{A}_l$*, then condition* 1 *is not needed for this particular pair.*

Here condition 2 defines a 'limiting mean' c_k for ϱ in each set $\mathcal{A}_k$, as condition 1 did in theorem 2.3.1. Condition 3 is the analogue of condition 2 in theorem 2.3.1. Condition 1 ensures that transition events across different sets $\mathcal{A}_k$ without visiting $\mathcal{H}$ can be ignored. $\pi_{\mathcal{A}_k}^{\infty}$ can be regarded as the limiting probability of $\mathcal{A}_k$ as the return time back to $\mathcal{H}$ becomes large.

As before $\pi_{\mathcal{A}_k}^{\infty}(c_k - \mu) \neq 0$ for at least one k is necessary for the long range dependence of ϱ to be at the same scale as M.

For a sanity check, consider the trivial example of an indicator function.

Example 2.3.3. *(Indicator functions) Let* ρ *be the indicator function of a finite set. We take* $\mathcal{A}_1 = \mathbb{N}$ *and* $c_1 = 0$*. Condition 1 is vacuous as there is only 1 partition. Condition 2 holds since the inner sum is finite and* $Q_{ij}^{(n)} \to \infty$*. Condition 3 holds because* ρ *is a bounded function. Thus we have that*

$$\lim_{n\to\infty} \frac{var(\sum_{r=1}^{n} \rho_i)}{R_{11}^{(n)}/\pi_1} = \pi(S)^2$$

where S *is the set on which* ρ *is non-zero.*

Now we illustrate the use of these tools with some applications. The first one uses theorem 2.3.1 directly, while the last two examples use theorem 2.3.2.

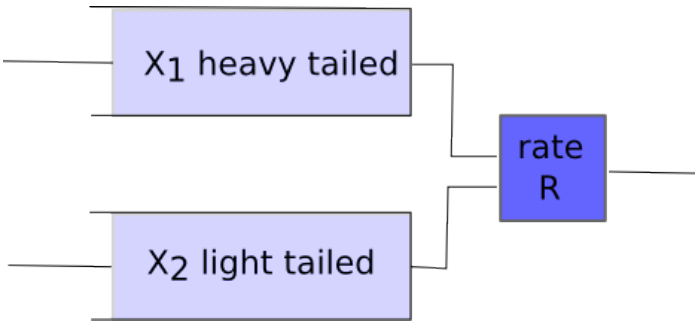


Figure 2.1. Parallel queues with fixed rate server.

2.4 Example 1: Longest queue first with mixed heavy and light tailed inputs

This example replicates the conclusion in (43) that long range dependence might spread under LQF scheduling in a parallel queue setting, using a general technique based on the theorems of the preceding section.

There is a single server of rate $R \in \mathbb{N}$ with 2 parallel queues (fig. 2.1). The queues are fed by independent random processes, each modeled by a discrete time, countable state Markov chain. As an example, we investigate the scenario where X_1 is i.i.d. with heavy tailed ($\text{var}(X_1) = \infty$) arrival distribution on $\mathbb{N}$. $X_2 \in \mathbb{N}$ is either an i.i.d. process with light tailed ($\text{var}(X_2) < \infty$) arrivals or X_2 can be a finite state $\mathbb{N}$-valued Markov chain in stationarity. We assume $E[X_1(0)] + E[X_2(0)] < R$.

Let $Q_1(n), Q_2(n)$ be the stationary queue lengths. We assume that the queue is work conserving, and moreover the scheduling decision at time n (number of packets to be served from each queue at time slot n) is a function of $(Q_1(n), Q_2(n))$, the queue sizes at time n. Given such a scheduling strategy, it is easily verified that $(X_1(n), X_2(n), Q_1(n), Q_2(n))$ is a countable state Markov chain.

Lemma 2.4.1. $(X_1(n), X_2(n), Q_1(n), Q_2(n))$ *is positive recurrent.*

Proof. $E[X_1(0)] + E[X_2(0)] < R$ implies that the queue process $(Q_1(n), Q_2(n))$ is positive recurrent. Pick $M_1 > 0$ and define the set $S_1 = \{Q_1(n) + Q_2(n) < M_1\}$. The return times to this set have finite mean (say ν). Also define $S_2 = \{X_1(n) + X_2(n) < M_2\}$ (or in the case X_2 is a finite state chain, $S_2 = \{X_1(n) < M_2\}$) where M_2 is large enough such that S_2 is nonempty. $S_1 \cap S_2$ is a nonempty compact set. We claim the return times to this set have a finite mean. Since $1_n(S_2)$ is i.i.d, there is a positive probability (say at least p) of visiting S_2 each time there is a visit to S_1 (independent of previous visits). It is easily seen that the mean return time to $S_1 \cap S_2$ is at most ν/p (Expectation of a sum of geometrically many i.i.d variables). □

We will look at long range dependence through the Hurst indices of the busy-idle processes of the queues. Let (X_1, Q_1') be the Markov chain if all the capacity were to be allocated to queue 1. Denote by $1(Q_1'(n) = 0)$, the busy-idle process of this queue. We know that the busy periods of Q_1' have infinite variance (see e.g. (7) theorem 8.10.3). Therefore both the Markov chain (X_1, Q_1') and the function $1(Q_1'(n) = 0)$ are LRD. (X_2, Q_2'), similarly defined, is a short range dependent chain.

Lemma 2.4.2. *$(X_1(n), X_2(n), Q_1(n), Q_2(n))$ is LRD.*

Proof. Consider the chain $(X_1(n), Q_1'(n), X_2(n), Q_2'(n))$. This chain is LRD because it is a combination of two independent chains (X_1, Q_1') and (X_2, Q_2'), one of which we assume to be LRD. Let t_1 be the return time to a nonempty compact set $S_1 = \{X_1(n), Q_1(n), X_2(n), Q_2(n) < M\}$. Similarly t_2 is the return time to the set $S_2 = \{X_1(n), Q_1'(n), X_2(n), Q_2'(n) < M\}$. Since $Q_1'(n) \leq Q_1(n)$ and $Q_2'(n) \leq Q_2(n)$, t_1 stochastically dominates t_2, and therefore $(X_1(n), X_2(n), Q_1(n), Q_2(n))$ is also LRD. □

The question we want to ask then is whether $1(Q_2(n) = 0)$, the busy-idle process of the second queue (fed by short range dependent traffic), is also long range dependent.

$\varrho_n := 1(Q_2(n) = 0)$ is an L_2 function of the chain $(X_1(n), X_2(n), Q_1(n), Q_2(n))$. Take $c = 0$ in theorem 2.3.1. $\mathcal{H} = \{X_1(n), X_2(n), Q_1(n), Q_2(n) \leq R\}$. Condition 2 holds trivially for bounded functions. Thus we are left with having to check the condition

$$\lim_{n\to\infty} \frac{1}{Q_{11}^{(n)}/\pi_1} \sum_{i,j:Q_{2,j}=0,Q_{2,i}=0} \pi_i \sum_{r=1}^{n} {}_{\mathcal{H}}p_{ij}^{(r)} = 0.$$

To see why this is true, note that $\sum_{i,j:Q_{2,j}=0Q_{2,i}=0} \pi_i \sum_{r=1}^{\infty} {}_{\mathcal{H}}p_{ij}^{(r)}$ is bounded above by 1 plus the stationary time spent in the states $\{Q_2 = 0\}$ before the chain visits $\mathcal{H}$. Note that the length of an idle period for Q_2 has finite expectation. Also note, if an idle period begins at time $n+1$, this implies due to the LQF policy that $Q_1(n) \leq R$, $Q_2(n) \leq R$, $X_1(n) \leq R$, and $X_2(n) \leq R$. Thus between successive idle periods of Q_2, the chain must visit $\mathcal{H}$. The stationary expected time spent in $\{Q_2 = 0\}$ without visiting $\mathcal{H}$ is therefore finite. Since $Q_{11}^{(n)} \to \infty$ (by (2.3)), the above limit holds. Using theorem 2.3.1, we conclude that $1(Q_2(n) = 0)$ has the same Hurst index as the chain $(X_1(n), X_2(n), Q_1(n), Q_2(n))$.

The advantage of this approach is that in general the input processes need not be i.i.d. Dependencies can easily be modeled, as long as the sources can be represented as countable state Markov.

2.5 Example 2: Compressing a long range dependent renewal process

This section provides an alternative proof for the result in (45).

Let $(X_n) \in \{0, 1\}$ be a discrete, stationary, ergodic renewal process. Denote by τ_1, τ_2 the times of the first two arrivals. Then we denote by $T \stackrel{d}{=} \tau_2 - \tau_1$, a random variable having the inter-arrival distribution. We assume $E[T] < \infty$ and $E[T^2] = \infty$.

As discussed in section 2.1.1, this is equivalent to stating that the renewal process is LRD . We begin by introducing the function

$$\varrho_n(X_{-\infty}^n) = -\log P(X_n|X_{-\infty}^{n-1}),$$

which is of central importance to coding theory. The behavior of (ϱ_n) restricts the minimum code length of lossless compression algorithms by the following lemma, (3), which is also proved in (33).

Lemma 2.5.1 (Barron's Lemma)**.** *Given* $\{c(n), n \geq 1\}$*, positive constants with* $\sum_n 2^{-c(n)} < \infty$*, we have*

$$l_n(X_1^n) \geq -\log P(X_1^n|X_{-\infty}^0) - c(n), \ \textit{eventually, a.s.} \ . \tag{2.6}$$

Here $l_n(X_1^n)$ is the code length for the first n symbols of the source for some lossless coding algorithm that produces bit strings. (i.e. let $l_n(X_1^n)$ be the length of $\phi(X_1^n)$ where $\phi(x_1^n) : \{0,1\}^n \to \{0,1\}^*$ is a one to one mapping.) $c(n)$ can be made logarithmic in n.

By the ergodic theorem, the limit of $\frac{1}{n}\sum_{i=1}^n \varrho_i$ as $n \to \infty$ exists a.s. and equals $\eta := E[-\log P(X_1|X_{-\infty}^0)]$, i.e. the entropy rate of (X_n). This implies the following well known first order converse source coding theorem for such sources.

Theorem 2.5.2.

$$\liminf_n \frac{1}{n} l_n(X_1^n) \geq \eta, \ \textit{a.s.} \ .$$

Lemma 2.5.1 is strong enough to permit second order refinements to theorem 2.5.2 once we know more about the process (ϱ_n). For example, in (33), it is shown that for certain short range dependent classes of sources (e.g. finite state Markov chains), and appropriate coding schemes (e.g. Shannon codebooks, Huffman coding etc.), $(l_n - n\eta)$ satisfies a central limit theorem.

Here, we will prove a second order converse source coding theorem, stating that the bit length process (l_n) will eventually dominate a long range dependent process the growth of whose variance is identical to that of (X_n), so that, in particular, it has the same Hurst index as (X_n). The proof relies on our general theorem 2.3.2. This result provides partial theoretical justification to existing empirical work in the field of variable bit-rate (VBR) video traffic ((5; 22; 52; 21) to cite a few). A conclusion resulting from this work is that long range dependence is omnipresent in VBR video traffic, and persists across a wide variety of codecs. Combined with these observations, the result backs the intuition that for many information sources long range dependence persists under compression. We generalize this result considerably in the next chapter.

Theorem 2.5.3. *Let (X_n) be an aperiodic, long range dependent, stationary, ergodic renewal process. Then, there exists a long range dependent random process (γ_n) such that*

$$L_n(X_1^n) \geq \gamma_n, \text{ eventually, a.s.}$$

for all uniquely decodable source codes. Moreover, (γ_n) has the same Hurst index as (X_n).

Proof. This immediately follows from Barron's lemma once we show (ϱ_n) are LRD with the same Hurst index as (X_n). This will follow from theorem 2.3.2 if we can set up (ϱ_n) as a function of a Markov chain.

We construct the following Markov chain (M_n) from the renewal process (X_n) (fig. 2.2):

- $M_n \in \{0, 1, 2, 3, \ldots\}$.
- $\{M_n = 0\} = \{X_{n-1}^n = 11\}$.
- For $k \in \{1, 2, \ldots\}$

– $\{M_n = 2k-1\} =$

$\{X_n = 0$ and k zeros since last arrival $\}$,

– $\{M_n = 2k\} =$

$\{X_n = 1$ and k zeros since last arrival in $X_n\}$.

Note that this Markov chain is equivalent to the characterization (X_n, t_n) (where t_n is the time since the last transition), only states are numbered such that the state space is $\mathbb{N}$.

We establish some notation:

$$\begin{aligned}
(X_n), &\quad \text{stationary renewal process,} \\
&\quad \text{interval-arrival lengths having the law of } T+1; \\
f_T(k) &:= P(T = k); \\
F_T(k) &:= P(T \leq k); \\
\varrho_n(X_{-\infty}^n) &:= -\log P(X_n | X_{-\infty}^{n-1}); \\
\eta &:= E[\log P(X_1 | X_{-\infty}^0)].
\end{aligned}$$

One can easily check $\varrho_n = \rho(M_n)$, with

- $\rho(0) = -\log f_T(0)$,
- $\rho(2k-1) = -\log P(T > k-1 | T \geq k-1)$,
- $\rho(2k) = -\log P(T = k | T \geq k)$.

We verify:

Lemma 2.5.4. *ϱ_n is an L_2 function of M_n.*

Proof. Let π_i be the stationary distribution of (M_n). Note that $\pi_i > 0 \implies \rho(i) < \infty$. We want to prove

$$\sum \rho(i)^2 \pi_i < \infty.$$

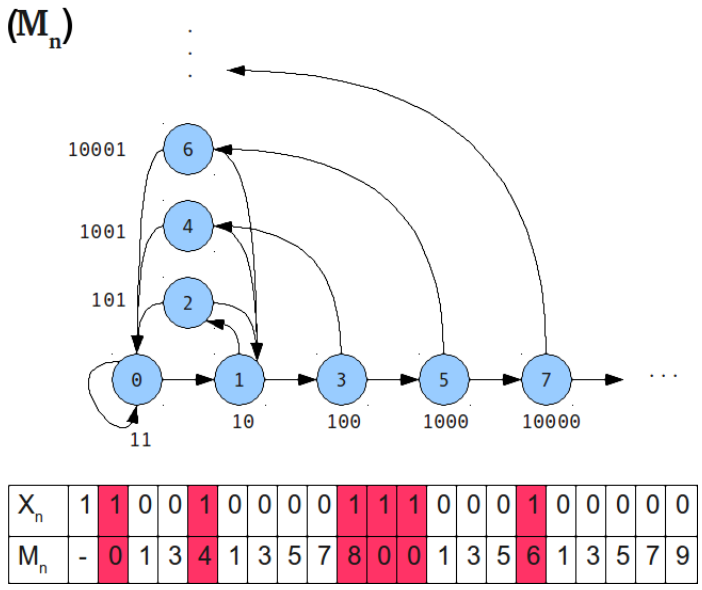


Figure 2.2. Construction of the Markov chain, with an example sequence showing the correspondence with X_n

Note that $\pi_{2k+1} = \pi_{2k-1}P(T > k|T \geq k)$, and $\pi_{2k} = \pi_{2k-1}P(T = k|T \geq k)$ for $k = 1, 2, \ldots$. This gives

$$\begin{aligned}
\sum \rho(i)^2\pi_i &= \pi_0\rho(0)^2 + \pi_1\rho(1)^2 \\
&\quad + \sum_{k=1}^{\infty} \pi_{2k-1}P(T = k|T \geq k)\log^2 P(T = k|T \geq k) \\
&\quad + \sum_{k=1}^{\infty} \pi_{2k-1}P(T > k|T \geq k)\log^2 P(T > k|T \geq k), \\
\pi_0\rho(0)^2 &= (\sum_{k=1}^{\infty} \pi_{2k}) f_T(0)\log^2 f_T(0), \\
\pi_1\rho(1)^2 &= (\sum_{k=1}^{\infty} \pi_{2k})(1 - f_T(0))\log^2(1 - f_T(0)).
\end{aligned}$$

Since the $p\log^2 p$ terms are bounded above by 1, $\sum \rho(i)^2\pi_i \leq 4$. □

Now, to apply theorem 2.3.2 we partition the state space into 3 sets as follows: $\mathcal{A}_1 = \{i > 0, i \text{ even}\}$, $\mathcal{A}_2 = \{0\} \cup \{i \text{ odd} : \rho(i) \leq -\log(1 - \epsilon_i)\}$, and $\mathcal{A}_3 = \{i \text{ odd} : \rho(i) > -\log(1 - \epsilon_i)\}$. Here we will will choose $\epsilon_i \downarrow 0$ later. Take $c_1 = c_2 = c_3 = 0$ and

$\mathcal{H} = \{1\}$ in that theorem. By the remark to the theorem, we don't need condition 1. We will check conditions 2 and 3 of theorem 2.3.2 for each of the sets.

When $i, j \in \mathcal{A}_1$ notice ${}_1p_{ij}^{(r)} = 0$, so both conditions hold automatically. For $i, j \in \mathcal{A}_2$, condition 2 holds due to remark no. 2 because the limit of $\rho(i)$ as $i \to \infty$ is zero, and condition 3 holds because ρ is bounded on this set. Thus we focus on $i, j \in \mathcal{A}_3$. Define $\rho(i) =: -\log(1-\tilde{\epsilon}_i)$. Let subsequence $\{i_k\} = \mathcal{A}_3$. We have $\tilde{\epsilon}_{i_k} \geq \epsilon_{i_k}$. $\pi_{i_k} \leq \pi_1 \prod_{l=1}^{k}(1-\tilde{\epsilon}_{i_k})$, and $\sum_1^\infty {}_1p_{i_k i_j}^{(r)} = \pi_{i_j}/\pi_{i_k}$. We have

$$
\begin{aligned}
&\sum_i \rho(i)\pi_i \sum_j \rho(j) \sum_{r=1}^{n} {}_1p_{ij}^{(r)} \\
\leq &\sum_k \prod_{l=1}^{k}(1-\tilde{\epsilon}_{i_l})(-\log(1-\tilde{\epsilon}_{i_k})) \sum_{j>k} -\log(1-\tilde{\epsilon}_{i_j}) \prod_{l=k+1}^{j}(1-\tilde{\epsilon}_{i_l}) \\
= &\sum_j \sum_{k<j} (1-\tilde{\epsilon}_{i_k}) \log(1-\tilde{\epsilon}_{i_k})(1-\tilde{\epsilon}_{i_j}) \log(1-\tilde{\epsilon}_{i_j}) \prod_{l=1, l\neq k,j}^{j}(1-\tilde{\epsilon}_{i_l}) \\
< &\sum_j j \prod_{l=3}^{j}(1-\tilde{\epsilon}_{i_l}).
\end{aligned}
$$

We can easily choose $\epsilon_i \downarrow 0$ such that this is finite. Dividing by $Q_{11}^{(n)}$, both conditions in theorem 2.3.2 will be satisfied.

□

2.6 Example 3: Long range dependence in financial time series

Let $(P_n, -\infty < n < \infty)$ be the price of some financial asset, and $X_n = \log P_n$. It is an established assumption that the log returns, $r_n = X_n - X_{n-1}$ is well modeled by a martingale difference process. Such a model accounts for the fact that the log returns exhibit little correlation. Nevertheless, it is also a widely observed fact that

some instantaneous functions of the log returns, such as $|r_n|^d$, exhibit long memory. (see e.g. (11))

The popular approach to modeling this behavior has been to explicitly write the dependence of the absolute log returns into the statistical description of the model. The result is the various long-memory autoregressive conditional heteroskedasticity (ARCH) process models of financial time series. ((24) for an example)

We want to show in this example that, given a martingale difference sequence (r_n) that can be represented as a function of a long range dependent Markov chain, the outcome that $|r_n|^d$ will exhibit long range dependence should not be considered surprising.

We want to illustrate this with a very simple example based on Mandelbrot's model for wheat prices ((40)). We should note that this simple model is for purposes of illustration only, and does not account for all known properties of financial time series. For instance, it has been observed in many situations that (r_n) has a finite variance, despite having a polynomially decaying marginal distribution. The (r_n) in this example has infinite variance. Nevertheless, the proof scheme used here to establish the long range dependence of $|r_n|^d$ should be applicable much more generally.

Let (W_n) be a stationary random process which models the weather. (W_n) can take on 3 values: good, bad, and neutral $\{g, b, n\}$. The length of a good period, T, (number of consecutive good days) has the same distribution as the length of a bad, or a neutral period. Let $P(T \geq t) = t^{-\alpha}$. T has finite mean but infinite variance (i.e. $1 < \alpha \leq 2$). A good or bad period is followed necessarily by a neutral period. A neutral period is followed by a good or bad period with equal probabilities.

Let $\hat{X}_n$ be the fundamental (log) price of the asset (which can be thought of as summarizing exogenous variables that affect the real price). $\hat{X}_n$ varies as follows: increases by 1 for every good day, decreases by 1 for every bad day, and stays the

same for every neutral day. The market calculates the real (log) price by projecting the expected future fundamental price: $X_n = \lim_{t\to\infty} E[\hat{X}_{n+t}|\hat{X}^n_{-\infty}]$.

By construction, (r_n) itself is a martingale difference sequence. We will now show that $\varrho_n = |r_n|^d$ is LRD with Hurst index $\frac{1}{2}(3-\alpha)$. ($0 < d < \alpha/2$ for $\text{var}(\varrho_0)$ to be finite.)

It can be verified that (also see the calculations in Mandelbrot's original paper (40)) X_n changes as follows: jumps by $E[T]$ on the first good day. Jumps by $-E[T]$ on the first bad day. Increases by $E[T|T \geq t] - E[T|T \geq t-1]$ on the t^{th} good day ($t \geq 2$). Decreases by $E[T|T \geq t] - E[T|T \geq t-1]$ on the t^{th} bad day. The first neutral following t good days decreases X_n by $E[T|T \geq t] - t$. The first neutral following t bad days increases X_n by $E[T|T \geq t] - t$.

Let $J_n = \mathbf{1}(\text{there is a transition at time } n)$. Let $T_n := \inf_t\{t \geq 0 : W_{n-t-1} \neq W_{n-t-2}\}$ be the number of days since the last transition (0 on the first day following).

Then $M_n = (W_n, J_n, T_n)$ is a countable state, long range dependent Markov chain, with Hurst index $\frac{1}{2}(3-\alpha)$. Moreover, $\varrho_n = |r_n|^d$ is a function of M_n:

- $\rho(\{g,b\}, 0, t) = (E[T|T \geq t+2] - E[T|T \geq t+1])^d$
- $\rho(\{n\}, 0, \cdot) = 0$
- $\rho(\{g,b\}, 1, \cdot) = (E[T])^d$
- $\rho(\{n\}, 1, t) = (E[T|T \geq t+1] - (t+1))^d$

Lemma 2.6.1.

$$E[T|T \geq t+2] - E[T|T \geq t+1] \to \frac{\alpha}{\alpha-1}, \qquad t \to \infty.$$

Proof.

$$P(T \geq s|T \geq t) = \frac{s^{-\alpha}}{t^{-\alpha}}, \qquad s \geq t$$

$$E[T|T \geq t+1] - E[T|T \geq t] = \sum_{s=t+1}^{\infty} P(T \geq s|T \geq t+1) - P(T \geq s|T \geq t)$$
$$= ((t+1)^{\alpha} - t^{\alpha}) \sum_{s=t+1}^{\infty} s^{-\alpha} \to \frac{\alpha}{\alpha - 1}$$

since $\frac{1}{\alpha-1}(t+2)^{-\alpha+1} = \int_{t+2}^{\infty} s^{-\alpha} ds < \sum_{s=t+1}^{\infty} s^{-\alpha} < \int_{t+1}^{\infty} s^{-\alpha} ds = \frac{1}{\alpha-1}(t+1)^{-\alpha+1}$ and $((t+1)^{\alpha} - t^{\alpha}) / t^{\alpha-1} \to \alpha$. □

Lemma 2.6.2.

$$E[T|T \geq t] - t \leq \frac{t}{\alpha - 1}.$$

Proof.

$$E[T|T \geq t] - t = \sum_{s=t}^{\infty} \frac{s^{-\alpha}}{t^{-\alpha}} \leq \int_{t}^{\infty} s^{-\alpha} ds = \frac{t}{\alpha - 1}.$$

□

We will utilize theorem 2.3.2 with $\mathcal{A}_1 = (\{g, b\}, 0, \cdot), \mathcal{A}_2 = (\{n\}, 0, \cdot), \mathcal{A}_3 = (\{g, b\}, 1, \cdot)$, $\mathcal{A}_4 = (\{n\}, 1, \cdot)$. $c_1 = c_4 = \left(\frac{\alpha}{\alpha-1}\right)^d, c_2 = c_3 = 0$. $\mathcal{H} = (\cdot, \cdot, 0)$. We have

$$\text{var}(\varrho_0) \leq E\varrho_0^2 = \sum_i \pi_i \rho(i)^2$$
$$= \sum_{i \notin \mathcal{A}_4} \pi_i \rho(i)^2 + \sum_{i \in \mathcal{A}_4} \pi_i \rho(i)^2 \qquad \leq C + \sum_{t=1}^{\infty} \frac{1}{2} P(T = t) (\frac{t}{\alpha - 1})^{2d} < \infty$$

by lemma 2.6.2. As $\rho(i)$ is bounded when $i \notin \mathcal{A}_4$, the contribution to the sum is a constant C. We also used the fact that if $i = (\{n\}, 1, t-1)$, then $\pi_i = P(W_{-t} = n)P(T = t) = \frac{1}{2}P(T = t)$.

We need to first show that condition 1 holds:

$$\lim_{n \to \infty} \frac{1}{Q_{11}^{(n)}} \sum_{r=1}^{n} \sum_{i \in \mathcal{A}_k, j \in \mathcal{A}_l} \pi_i |\rho(i) - \mu||\rho(j) - \mu|_{\mathcal{H}} p_{ij}^{(r)} \to 0 \quad \forall k \neq l.$$

By inspection, the following transitions require visiting $\mathcal{H}$:

(k,l) or $(l,k) = (1,2),(1,3),(2,4),(3,4)$. The sum is zero for these pairs. For (k,l) or $(l,k) = (1,4),(2,3)$, the condition is not needed due to the remark to theorem 2.3.2.

Condition 2 reads

$$\lim_{n\to\infty} \frac{1}{Q_{11}^{(n)}} \sum_{i,j\in\mathcal{A}_k} \pi_i(\rho(i)-c_k)(\rho(j)-c_k) \sum_{r=1}^{n} {}_{\mathcal{H}}p_{ij}^{(r)} = 0 \quad \forall k.$$

For $k = 3,4$, ${}_{\mathcal{H}}p_{ij}^{(r)} = 0$ because these states must go to $\mathcal{H}$ in one step. For $k = 1,2$, we have chosen c_k such that $(\rho(i) - c_k) \to 0$ by lemma 2.6.1. The condition holds by remark no. 2.

Condition 3 also holds for $\mathcal{A}_1$, $\mathcal{A}_2$, and $\mathcal{A}_3$ because ρ is bounded on these sets. On $\mathcal{A}_4$, it holds because ${}_{\mathcal{H}}p_{ij}^{(r)} = 0$ as argued earlier. We finally have the conclusion:

$$\lim_{n\to\infty} \frac{\sum_{r=1}^{n} \operatorname{cov}(|r_0|^d, |r_n|^d)}{Q_{11}^{(n)}/\pi_1} = \sum_{k=1}^{K} \pi_k^{\infty}(\mu - c_k)^2 > 0.$$

2.7 A non-example

Consider an LRD Markov chain with $p_{12}^{(1)} = 1$ and $p_{i2}^{(1)} = 0$ for $i > 1$. Set $\rho(1) = 1$, $\rho(2) = -1$ and $\rho(i) = 0$ for $i > 2$. We have for this chain $\pi_1 = \pi_2$ and $\mu = 0$. Since $\rho(i) = 0$ for $i > 2$, the conditions of theorem 2.3.1 hold with $c = 0$ and $\mathcal{H} = \{1\}$. However, since $\mu = c$, the conclusion about the equality of Hurst indices does not follow. In fact we can show ϱ to be short range dependent. From (10), section 3 we know

$$\sum_{r=1}^{n} \operatorname{cov}(\varrho_0, \varrho_r) = \sum_{i,j} \rho(i)\rho(j)\pi_i Q_{ij}^{(n)}.$$

The RHS is a finite sum, giving

$$\sum_{r=1}^{n} \operatorname{cov}(\varrho_0, \varrho_r) = \pi_1(Q_{11}^{(n)} + Q_{22}^{(n)}) - \pi_1(Q_{12}^{(n)} + Q_{21}^{(n)}),$$

where we used $\pi_1 = \pi_2$. Since $p_{12}^{(1)} = 1$ and $p_{i2}^{(1)} = 0$ for $i > 1$, we also know $p_{12}^{(r+1)} = p_{11}^{(r)}$ and $p_{22}^{(r+1)} = p_{21}^{(r)}$. Expanding the $Q^{(n)}$ as sums, we get

$$\begin{aligned}\sum_{r=1}^{n} \operatorname{cov}(\varrho_0, \varrho_r) &= \pi_1 \sum_{r=1}^{n} (p_{12}^{(r+1)} - \pi_1) - (p_{12}^{(r)} - \pi_1) + \pi_1 \sum_{r=1}^{n} (p_{22}^{(r)} - \pi_1) - (p_{22}^{(r+1)} - \pi_1) \\ &= \pi_1[(p_{12}^{(n+1)} - \pi_1) + (p_{22}^{(1)} - \pi_1) - (p_{12}^{(1)} - \pi_1) - (p_{22}^{(n+1)} - \pi_1)]\end{aligned}$$

which remains bounded, demonstrating that (ϱ_n) is a short range dependent process.

2.8 Proof of theorems

For the proofs, we will rely on several lemmas, most of which are already known.

Lemma 2.8.1. *(Chung (10), chapter 11, Corollary 1.) For $p \geq 0$,*

$$E_1 T_1^p = \infty \iff E_i T_i^p = \infty, \qquad \forall i \in \mathbb{N}.$$

Lemma 2.8.2. *Let (a_n) be an arbitrary sequence and $b_n \to \infty$. c is a finite real number. If*

$$\frac{a_n}{b_n} \to c,$$

then

$$\frac{\sum_{r=1}^{n} a_r}{\sum_{r=1}^{n} b_r} \to c.$$

Proof. This elementary result follows from the discrete analogue of l'Hôpital's rule, referred to as the Stolz-Cesàro theorem. See e.g. 3.1.7 in (44). □

Lemma 2.8.3. *(i)*

$$cov(\varrho_0, \varrho_r) = \sum_{i,j} \pi_i p_{ij}^{(r)} (\rho(i) - \mu)(\rho(j) - \mu).$$

(ii)

$$\sum_{r=1}^{n} cov(\varrho_0, \varrho_r) = \sum_{i,j} \rho(i)\rho(j)\pi_i Q_{ij}^{(n)}.$$

(iii)

$$var(\varrho_0 + \ldots + \varrho_n) - (n+1)var(\varrho_0) = 2\sum_{i,j} \rho(i)\rho(j)\pi_i R_{ij}^{(n)}.$$

Proof. (i) is a simple expansion. (ii) is derived from (i), and (iii) can be found in (9), section 3. □

Lemma 2.8.4. *(Eq. (1) in Chung (10), theorem 9.1.)*

$$p_{ij}^{(r)} = {}_1p_{ij}^{(r)} + \sum_{m=1}^{r-1} {}_1p_{i1}^{(m)} p_{1j}^{(r-m)}, \quad r \geq 1. \tag{2.7}$$

Lemma 2.8.5. *(Carpio & Daley (9), 2.12.)*

$$\begin{aligned} Q_{11}^{(n)} &\sim (\pi_1)^2 \sum_{u=1}^{\infty} \min(u,n) \sum_{s=u+1}^{\infty} f_{11}^{(s)} \\ &= (\pi_1)^2 \sum_{u=1}^{\infty} \sum_{r=1}^{\min(u,n)} \sum_{s=u+1}^{\infty} f_{11}^{(s)} \\ &= (\pi_1)^2 \sum_{r=1}^{n} \sum_{u=r}^{\infty} \sum_{s=u+1}^{\infty} f_{11}^{(s)}. \end{aligned}$$

Lemma 2.8.6.

$$\lim_{n\to\infty} \frac{1}{Q_{11}^{(n)}/\pi_1} \sum_{i,j} \pi_i \sum_{r=1}^{n} {}_1p_{ij}^{(r)} = 1.$$

Proof.

$$\begin{aligned} \sum_{i,j} \pi_i \sum_{r=1}^{n} {}_1p_{ij}^{(r)} &= \sum_{r=1}^{n} \sum_{i,j} \pi_i \, {}_1p_{ij}^{(r)} \\ &=^a \sum_{r=1}^{n} \sum_{i} \pi_i \sum_{u=r}^{\infty} f_{i1}^{(r)} \\ &=^b \sum_{r=1}^{n} \sum_{u=r}^{\infty} \frac{1}{m_{11}} \sum_{s=u}^{\infty} f_{11}^{(s)} \\ &= \frac{1}{m_{11}} \sum_{r=1}^{n} \sum_{u=r}^{\infty} f_{11}^{u} + \sum_{r=1}^{n} \sum_{u=r}^{\infty} \frac{1}{m_{11}} \sum_{s=u+1}^{\infty} f_{11}^{(s)} \\ &= \frac{1}{m_{11}} \sum_{r=1}^{n} P_1(T_1 \geq r) + \sum_{r=1}^{n} \sum_{u=r}^{\infty} \frac{1}{m_{11}} \sum_{s=u+1}^{\infty} f_{11}^{(s)} \\ &\sim \frac{1}{\pi_1^2 m_{11}} Q_{11}^{(n)} = \frac{Q_{11}^{(n)}}{\pi_1} \end{aligned}$$

since $\sum_{r=1}^{n} P_1(T_1 \geq r) \leq m_{11}$ and by lemma (2.8.5). Here (a) uses $\sum_j {}_1p_{ij}^{(r)} = \sum_r^{\infty} f_{i1}^{(r)}$, which are equivalent ways of expressing the probability of going from i to any other state without going to 1 in r steps. This expression also appears chapter 9 of (10) (proof of thm. 6). (b) uses the fact $P_\pi(T_1 = r) = \frac{P_1(T_1 \geq r)}{m_{11}}$, where T_1 is the first return time to 1 at stationarity.

□

Lemma 2.8.7. *Let $M > 0$ be a finite number,*

$$\lim_{n\to\infty} \frac{1}{Q_{11}^{(n)}/\pi_1} \sum_{\{i<M\}\cup\{j<M\}} \pi_i \sum_{r=1}^{n} {}_1p_{ij}^{(r)} = 0.$$

Proof. Pick m s.t. ${}_1p_{1i}^{(m)} > 0$, then

$${}_1p_{1i}^{(m)}\,{}_1p_{ij}^{*} \leq {}_1p_{1j}^{*} = \pi_j/\pi_1.$$

Thus, there exists a finite constant C_M s.t. ${}_1p_{ij}^{*} < C_M\pi_j$ for all $i < M$. Conclude

$$\sum_{i<M,j} \pi_i \sum_{r=1}^{n} {}_1p_{ij}^{(r)} \leq C_M \sum_{i<M,j} \pi_i\pi_j \leq C_M.$$

Similarly, there exists a finite constant D_M s.t. ${}_1p_{ij}^{*} \leq 1 + {}_1p_{jj}^{*} \leq D_M$ for all $j < M$.

$$\sum_{j<M,i} \pi_i \sum_{r=1}^{n} {}_1p_{ij}^{(r)} \leq D_M \sum_{j<M,i} \pi_i \leq MD_M.$$

Using (2.3) we conclude the proof. □

Lemma 2.8.8. *((9), pg 1051.)*

$$\left| \frac{Q_{1j}^{(n)}/\pi_j}{Q_{11}^{(n)}/\pi_1} \right| \leq 1.$$

Lemma 2.8.9.

$$\left| \sum_{r=1}^{n} \sum_{i,j} \pi_i |\rho(i)\rho(j)|\,{}_1p_{ij}^{(r)} - \sum_{r=1}^{n} \sum_{i,j} \pi_i |\rho(i)\rho(j)|\,{}_{\mathcal{H}}p_{ij}^{(r)} \right| \leq (|\mathcal{H}|+1)C_{\mathcal{H}} \sum_{i,j} \pi_i\pi_j |\rho(i)\rho(j)|,$$

where $\mathcal{H}$ is any non-empty set with a finite number of states and $C_{\mathcal{H}}$ is a constant that depends only on $\mathcal{H}$.

Proof. Let $\mathcal{H}' = \mathcal{H} \cup \{k\}$, $k \notin \mathcal{H}$. We will argue by induction. We write

$$\sum_{r=1}^{n} {}_{\mathcal{H}}p_{ij}^{(r)} - {}_{\mathcal{H}'}p_{ij}^{(r)} = \sum_{r=1}^{n} P(M_r = j; M_l \notin \mathcal{H}, 1 \le l < r; M_l = k, \text{for some } 1 \le l < r | M_0 = i)$$

$$\begin{aligned} &= \sum_{r=1}^{n} \sum_{m=1}^{r-1} {}_{\mathcal{H}'}p_{ik}^{(m)} {}_{\mathcal{H}}p_{kj}^{(r-m)} \\ &= \sum_{m=1}^{n-1} {}_{\mathcal{H}'}p_{ik}^{(m)} \sum_{r=m+1}^{n} {}_{\mathcal{H}}p_{kj}^{(r-m)} \\ &\le \underbrace{\left(\sum_{m=1}^{\infty} {}_{\mathcal{H}'}p_{ik}^{(m)} \right)}_{C_1} \underbrace{\left(\sum_{r=1}^{\infty} {}_{\mathcal{H}}p_{kj}^{(r)} \right)}_{{}_{\mathcal{H}}p_{kj}^{*}}. \end{aligned}$$

C_1 is bounded above by 1 since

$$\sum_{m=1}^{\infty} {}_{\mathcal{H}'}p_{ik}^{(m)} \le \sum_{m=1}^{\infty} {}_{k}p_{ik}^{(m)} = 1.$$

Let $h \in \mathcal{H}$. m is s.t. ${}_{h}p_{hk}^{(m)} > 0$.

$${}_{h}p_{hk}^{(m)} {}_{\mathcal{H}}p_{kj}^{*} \le {}_{h}p_{hj}^{*} = \pi_j / \pi_h.$$

Thus ${}_{\mathcal{H}}p_{kj}^{*} \le \pi_j / ({}_{h}p_{hk}^{(m)} \pi_h) = C_{\mathcal{H}'}$.

$$\sum_{r=1}^{n} \sum_{i,j} \pi_i |\rho(i)\rho(j)| {}_{\mathcal{H}}p_{ij}^{(r)} - \sum_{r=1}^{n} \sum_{i,j} \pi_i |\rho(i)\rho(j)| {}_{\mathcal{H}'}p_{ij}^{(r)} \le C_{\mathcal{H}'} \sum_{i,j} \pi_i \pi_j |\rho(i)\rho(j)|.$$

Therefore adding or subtracting a state from the set $\mathcal{H}$ (as long as the resulting set is non-empty) only affects the sum in question by a bounded amount. As a result, replacing $\mathcal{H}$ by $\{1\}$ can change the sum by at most $(1 + |\mathcal{H}|) C_{\mathcal{H}} \sum_{i,j} \pi_i \pi_j |\rho(i)\rho(j)|$. (Add state 1 if it is not already in set $\mathcal{H}$. Then subtract all other states until only state 1 is left.) □

2.8.1 Proof of theorem 2.3.1

Proof. By (2.3) and lemma (2.8.9) the conditions are equivalent to

(condition 1)

$$\lim_{n\to\infty} \frac{1}{Q_{11}^{(n)}/\pi_1} \sum_{r=1}^{n} \sum_{i,j} \pi_i(\rho(i)-c)(\rho(j)-c)_1 p_{ij}^{(r)} = 0$$

for some constant c, and

(condition 2)

$$\lim_{L\to\infty} \limsup_{n\to\infty} \frac{1}{Q_{11}^{(n)}/\pi_1} \sum_{r=1}^{n} \sum_{i,j} \pi_i |\rho(i)\rho(j)| 1(|\rho(i)|, |\rho(j)| > L)_1 p_{ij}^{(r)} = 0.$$

Define

$$\overline{\rho}^M(i) = \begin{cases} \rho(i) & , i \le M \\ c & , i > M \end{cases}.$$

$\overline{\mu}^M = E[\overline{\varrho}_n^M]$, $\underline{\rho}^M(i) = \rho(i) - \overline{\rho}^M(i)$, and $\underline{\mu}^M = E[\underline{\varrho}_n^M]$. We adopt the shorthand notation:

$$\phi_n = \frac{(\varrho_0 + \ldots + \varrho_n) - (n+1)\mu}{\sqrt{2R_{11}^{(n)}/\pi_1}},$$

$$\overline{\phi}_n^M = \frac{(\bar{\varrho}_0^M + \ldots + \bar{\varrho}_n^M) - (n+1)\overline{\mu}^M}{\sqrt{2R_{11}^{(n)}/\pi_1}},$$

$$\underline{\phi}_n^M = \phi_n - \overline{\phi}_n^M.$$

We will be referring to the reverse triangle inequality for random variables:

$$\left| \sqrt{\mathrm{var}(\phi_n)} - \sqrt{\mathrm{var}(\overline{\phi}_n^M)} \right| \le \sqrt{\mathrm{var}(\underline{\phi}_n^M)}. \tag{2.8}$$

This follows directly from the triangle inequlity. Using lemma 2.8.4, write 2.8.3(i) as

$$\sum_{r=1}^{n} \mathrm{cov}(\varrho_0, \varrho_r) = \sum_{i,j} \pi_i(\rho(i)-\mu)(\rho(j)-\mu) \sum_{r=1}^{n} {}_1p_{ij}^{(r)} + \tag{2.9}$$

$$\sum_{i,j} \pi_i \sum_{r=1}^{n} \sum_{m=1}^{r-1} {}_1p_{i1}^{(m)} p_{1j}^{(r-m)} (\rho(i)-\mu)(\rho(j)-\mu).$$

The second term can be rewritten

$$\sum_{i,j} \pi_i \sum_{r=1}^{n} \sum_{m=1}^{r-1} {}_1p_{i1}^{(m)} p_{1j}^{(r-m)} (\rho(i)-\mu)(\rho(j)-\mu) =$$

$$\sum_{i,j} \pi_i \sum_{m=1}^{n-1} {}_1p_{i1}^{(m)} \sum_{r=m+1}^{n} p_{1j}^{(r-m)} (\rho(i)-\mu)(\rho(j)-\mu)$$

$$= \sum_{m=1}^{n-1} \Bigg(\sum_{r=m+1}^{n} \sum_{i,j} {}_1p_{i1}^{(m)} \pi_i (p_{1j}^{(r-m)} - \pi_j)(\rho(i)-\mu)(\rho(j)-\mu) + \underbrace{\sum_{r=m+1}^{n} \sum_{i,j} {}_1p_{i1}^{(m)} \pi_i \pi_j (\rho(i)-\mu)(\rho(j)-\mu)}_{0} \Bigg)$$

$$= \sum_{m=1}^{n-1} \sum_{i,j} \pi_{i1} p_{i1}^{(m)} Q_{1j}^{(n-m)} (\rho(i)-\mu)(\rho(j)-\mu).$$

Dividing by $Q_{11}^{(n)}/\pi_1$ we get

$$= \sum_{m=1}^{n-1} \sum_{i,j} \pi_{i1} p_{i1}^{(m)} \pi_j \frac{Q_{1j}^{(n-m)}/\pi_j}{Q_{11}^{(n)}/\pi_1} (\rho(i)-\mu)(\rho(j)-\mu).$$

By lemma 2.8.8 we have

$$\sum_j \pi_j | \frac{Q_{1j}^{(n-m)}/\pi_j}{Q_{11}^{(n)}/\pi_1} || (\rho(j)-\mu)| < \infty.$$

We also know $\sum_i \pi_i \sum_{m=1}^{n-1} {}_1p_{i1}^{(m)} (\rho(i)-\mu) \to 0$. Therefore

$$\lim_{n\to\infty} \sum_{m=1}^{n-1} \sum_{i,j} \pi_{i1} p_{i1}^{(m)} \pi_j \frac{Q_{1j}^{(n-m)}/\pi_j}{Q_{11}^{(n)}/\pi_1} (\rho(i)-\mu)(\rho(j)-\mu) = 0.$$

(Dominated convergence) The result has the interpretation that the sum of the covariances between ϱ_0 and ϱ_n on the event that the chain visits state 1 at least once before time n, is negligible compared to $Q_{11}^{(n)}$.

We want to use these results to conclude $\mathrm{var}(\underline{\phi}_n^M) \to 0$. For this we write eq. 2.9 for $\underline{\rho}^M$, $c = 0$. The first term in eq. 2.9 reads after a little manipulation

$$\sum_{i,j} \pi_i [\underline{\rho}^M(i)\underline{\rho}^M(j) - \underline{\mu}^M(\underline{\rho}^M(i) + \underline{\rho}^M(j)) + (\underline{\mu}^M)^2] \sum_{r=1}^{n} {}_1p_{ij}^{(r)}. \tag{2.10}$$

Now assume ρ is bounded. After dividing by $Q_{11}^{(n)}/\pi_1$, the second and third terms are $O(\underline{\mu}^M)$ as $\underline{\mu}^M \to 0$ by lemma 2.8.6. Since $\underline{\mu}^M \to 0$ with M, these terms go to 0 as $M \to \infty$ uniformly in n.

For the first term in (2.10), write condition 1 as follows for comparison:

$$\lim_{n\to\infty}\frac{1}{Q_{11}^{(n)}/\pi_1}\left(\sum_{r=1}^{n}\sum_{i\le M,j\le M}\pi_i(\rho(i)-c)(\rho(j)-c)_1p_{ij}^{(r)}\right.$$
$$+\sum_{r=1}^{n}\sum_{i\le M,j>M}\pi_i(\rho(i)-c)(\rho(j)-c)_1p_{ij}^{(r)}$$
$$+\sum_{r=1}^{n}\sum_{i>M,j\le M}\pi_i(\rho(i)-c)(\rho(j)-c)_1p_{ij}^{(r)}$$
$$\left.+\sum_{r=1}^{n}\sum_{i>M,j>M}\pi_i(\rho(i)-c)(\rho(j)-c)_1p_{ij}^{(r)}\right)=0.$$

The first three sums have limit 0 because ρ is assumed to be bounded, and by lemma 2.8.7. The last sum is identical to the first term in (2.10). Therefore dividing eq. 2.9 by $Q_{11}^{(n)}/\pi_1$ and applying lemma 2.8.2 while observing lemma 2.8.3 (ii) and (iii), we conclude that $\lim_{M\to\infty}\lim_{n\to\infty}\operatorname{var}(\underset{\sim}{\phi}_n^M)=0$, and by eq. (2.8), also that $\lim_{M\to\infty}\lim_{n\to\infty}\operatorname{var}(\overline{\phi}_n^M)=\lim_{n\to\infty}\operatorname{var}(\phi_n)$.

To calculate $\operatorname{var}(\overline{\phi}_n^M)$, rewrite eq. 2.9 for $\overline{\rho}^M$:

$$\sum_{i,j}\pi_i[(\overline{\rho}^M(i)-c)(\overline{\rho}^M(j)-c)-(\overline{\mu}^M-c)(\overline{\rho}^M(i)+\overline{\rho}^M(j)-2c)+(\overline{\mu}^M-c)^2]\sum_{r=1}^{n}{}_1p_{ij}^{(r)}.$$

The first two sums will go to zero when dividing by $Q_{11}^{(n)}/\pi_1$, by the boundedness of ρ and lemma 2.8.7 because of truncation. The last term will read:

$$(\overline{\mu}^M-c)^2\frac{1}{Q_{11}^{(n)}/\pi_1}\sum_{i,j}\pi_i\sum_{r=1}^{n}{}_1p_{ij}^{(r)}\to(\overline{\mu}^M-c)^2, n\to\infty$$

by lemma 2.8.6. By lemma 2.8.3 (ii) and (iii), and lemma 2.8.2 this concludes the proof when (ϱ_n) is bounded.

When (ϱ_n) is not bounded, we truncate by value, i.e. $\tilde{\rho}^L(i)=\rho(i)1(\rho(i)\le L)$, $\tilde{\mu}^L=E[\tilde{\varrho}_n^L]$, $\underset{\sim}{\rho}^L(i)=\rho(i)-\tilde{\rho}^L(i)$, and $\underset{\sim}{\mu}^L=E[\underset{\sim}{\varrho}_n^L]$. Also define:

$$\tilde{\phi}_n^L=\frac{(\tilde{\varrho}_0^L+\ldots+\tilde{\varrho}_n^L)-(n+1)\tilde{\mu}^L}{\sqrt{2R_{11}^{(n)}/\pi_1}},$$

$$\underset{\sim}{\phi}_n^L = \phi_n - \tilde{\phi}_n^L.$$

We can express $\sum_{r=1}^n \mathrm{cov}(\underset{\sim}{\varrho}_0^L, \underset{\sim}{\varrho}_r^L)$ as in eq. 2.9, and argue there that the second term has limit 0 as $n \to \infty$ when divided by $Q_{11}^{(n)}/\pi_1$. The first term also has limit 0 due to the assumed condition 2. We appeal again to lemma 2.8.3 (ii) and (iii), and lemma 2.8.2 to argue that $\lim_{L\to\infty} \lim_{n\to\infty} \mathrm{var}(\underset{\sim}{\phi}_n^L) = 0$. By eq. (2.8), we also get $\lim_{L\to\infty} \lim_{n\to\infty} \mathrm{var}(\tilde{\phi}_n^L) = \lim_{n\to\infty} \mathrm{var}(\phi_n)$. We conclude:

$$\lim_{n\to\infty} \frac{\mathrm{var}(\sum_{r=1}^n \varrho_i)}{R_{11}^{(n)}/\pi_1} = \lim_{L\to\infty} \lim_{n\to\infty} \mathrm{var}(\tilde{\phi}_n^L) = \lim_{L\to\infty} (\tilde{\mu} - c)^2 = (\mu - c)^2.$$

The claim about the Hurst indices can be argued as follows. Consider the expression in lemma 2.8.3 (ii) for $\varrho_n = 1(M_n = 1)$. Dividing by $Q_{11}^{(n)}/\pi_1$, we see that the right hand side has limit $\pi_1^2 > 0$. From the above argument it follows that $\left(\sum_{r=1}^n \mathrm{cov}(1(M_0 = 1), 1(M_r = 1))\right) / \left(\sum_{r=1}^n \mathrm{cov}(\varrho_0, \varrho_r)\right)$ has a finite, non-zero limit if $\mu \neq c$. It is easily seen from the definition of H that ρ has the same Hurst index as the indicator function $1(M_n = 1)$. □

2.8.2 Proof of theorem 2.3.2

Proof. By (2.3) and lemma (2.8.9) the conditions are equivalent to

(condition 1)

$$\lim_{n\to\infty} \frac{1}{Q_{11}^{(n)}/\pi_1} \sum_{r=1}^n \sum_{i\in\mathcal{A}_k, j\in\mathcal{A}_l} \pi_i |\rho(i) - \mu||\rho(j) - \mu|_1 p_{ij}^{(r)} = 0, \quad \forall k \neq l,$$

(condition 2)

$$\lim_{n\to\infty} \frac{1}{Q_{11}^{(n)}/\pi_1} \sum_{r=1}^n \sum_{i,j\in\mathcal{A}_k} \pi_i (\rho(i) - c_k)(\rho(j) - c_k)_1 p_{ij}^{(r)} = 0, \quad \forall k,$$

(condition 3)

$$\lim_{L\to\infty} \limsup_{n\to\infty} \frac{1}{Q_{11}^{(n)}/\pi_1} \sum_{r=1}^n \sum_{i,j\in\mathcal{A}_k} \pi_i |\rho(i)\rho(j)| 1(|\rho(i)|, |\rho(j)| > L)_1 p_{ij}^{(r)} = 0, \quad \forall k.$$

We truncate as follows:

$$\overline{\rho}^M(i) = \begin{cases} \rho(i) & , i < M \\ c_k & , i \geq M, i \in \mathcal{A}_k \end{cases}.$$

$\underline{\rho}^M$, $\overline{\mu}^M$, $\underline{\mu}^M$, $\overline{\phi}^M$, and $\underline{\phi}^M$ are defined as before.

The first sum in eq. 2.9 can be decomposed as:

$$\begin{aligned} &\sum_{i,j} \pi_i(\rho(i)-\mu)(\rho(j)-\mu) \sum_{r=1}^{n} {}_1p_{ij}^{(r)} = \\ &\quad \sum_{k=1}^{K} \sum_{i,j \in \mathcal{A}_k} \pi_i(\rho(i)-\mu)(\rho(j)-\mu) \sum_{r=1}^{n} {}_1p_{ij}^{(r)} \\ &\quad + \sum_{k,l \in \{1,\ldots,K\}, k \neq l} \sum_{i \in \mathcal{A}_k, j \in \mathcal{A}_l} \pi_i(\rho(i)-\mu)(\rho(j)-\mu) \sum_{r=1}^{n} {}_1p_{ij}^{(r)}. \end{aligned} \tag{2.11}$$

The first condition ensures that the cross terms on the right are insignificant. Therefore we can work with each subset separately.

We will argue as in the proof of theorem 2.3.1 to show $\operatorname{var}(\underline{\phi}_n^M) \to 0$. The analogue of eq. 2.10 for each of the remaining sums reads

$$\sum_{i,j \in \mathcal{A}_k} \pi_i [\underline{\rho}^M(i)\underline{\rho}^M(j) - \underline{\mu}^M(\underline{\rho}^M(i) + \underline{\rho}^M(j)) + (\underline{\mu}^M)^2] \sum_{r=1}^{n} {}_1p_{ij}^{(r)}.$$

Assume ρ is bounded. After dividing by $Q_{11}^{(n)}/\pi_1$, the second and third terms are $O(\underline{\mu}^M)$ as $\underline{\mu}^M \to 0$ by lemma 2.8.6. Since $\underline{\mu}^M \to 0$ as $M \to \infty$, these terms tend to 0 as $M \to \infty$ uniformly in n.

For the first term, we argue exactly as in the proof of theorem 2.3.1 that condition 1, together with lemma 2.8.7 implies that this term, when divided by $Q_{11}^{(n)}/\pi_1$ goes to 0 as $n \to \infty$. Applying lemma 2.8.2 while observing lemma 2.8.3 (ii) and (iii), we conclude that $\lim_{M\to\infty} \lim_{n\to\infty} \operatorname{var}(\underline{\phi}_n^M) = 0$, and by eq. (2.8), also that $\lim_{M\to\infty} \lim_{n\to\infty} \operatorname{var}(\overline{\phi}_n^M) = \lim_{n\to\infty} \operatorname{var}(\phi_n)$.

To calculate $\text{var}(\overline{\phi}_n^M)$, rewrite eq. 2.9 for $\overline{\rho}^M$. We again omit the cross sums:

$$\sum_{i,j\in\mathcal{A}_k} \pi_i[(\overline{\rho}^M(i)-c_k)(\overline{\rho}^M(j)-c_k)-(\overline{\mu}^M-c_k)(\overline{\rho}^M(i)+\overline{\rho}^M(j)-2c_k)+(\overline{\mu}^M-c_k)^2]\sum_{r=1}^{n} {}_1p_{ij}^{(r)}.$$

The first two sums will go to zero due to truncation, boundedness of ρ, and by lemma 2.8.7, when dividing by $Q_{11}^{(n)}/\pi_1$. The last term will read:

$$(\overline{\mu}^M - c_k)^2 \frac{1}{Q_{11}/\pi_1} \sum_{i,j\in\mathcal{A}_k} \pi_i \sum_{r=1}^{n} {}_1p_{ij}^{(r)} \to \pi_{\mathcal{A}_k}^{\infty}(\overline{\mu}^M - c_k)^2$$

by lemma 2.8.6 and the definition of $\pi_{\mathcal{A}_k}^{\infty}$. This concludes the proof when (ϱ_n) is bounded.

When (ϱ_n) is not bounded, we truncate by value, i.e. $\tilde{\rho}^L(i) = \rho(i)1(\rho(i) \le L)$, $\tilde{\mu}^L = E[\tilde{\varrho}_n^L]$, $\underset{\sim}{\rho}^L(i) = \rho(i) - \tilde{\rho}^L(i)$, and $\underset{\sim}{\mu}^L = E[\underset{\sim}{\varrho}_n^L]$. Also define:

$$\tilde{\phi}_n^L = \frac{(\tilde{\varrho}_0^L + \ldots + \tilde{\varrho}_n^L) - (n+1)\tilde{\mu}^L}{\sqrt{2R_{11}^{(n)}/\pi_1}},$$

$$\underset{\sim}{\phi}_n^L = \phi_n - \tilde{\phi}_n^L.$$

We also partition $\underset{\sim}{\varrho}_n^L$ as $\sum_{k=1}^K \underset{\sim}{\varrho}_n^L 1(\underset{\sim}{\varrho}_n^L \in \mathcal{A}_k)$. Define:

$${}_k\underset{\sim}{\phi}_n^L = \frac{\underset{\sim}{\varrho}_0^L 1(\underset{\sim}{\varrho}_0^L \in \mathcal{A}_k) + \ldots + \underset{\sim}{\varrho}_n^L 1(\underset{\sim}{\varrho}_n^L \in \mathcal{A}_k) - (n+1)E(\underset{\sim}{\varrho}_0^L 1(\underset{\sim}{\varrho}_0^L \in \mathcal{A}_k))}{\sqrt{2R_{11}^{(n)}/\pi_1}}.$$

We can express $\sum_{r=1}^n \text{cov}(\underset{\sim}{\varrho}_0^L 1(\underset{\sim}{\varrho}_0^L \in \mathcal{A}_k), \underset{\sim}{\varrho}_r^L 1(\underset{\sim}{\varrho}_r^L \in \mathcal{A}_k))$ by writing eq. 2.9 for $\underset{\sim}{\rho}^L(i)1(\underset{\sim}{\rho}^L(i) \in \mathcal{A}_k)$, and argue there that the second term has limit 0 as $n \to \infty$ when divided by $Q_{11}^{(n)}/\pi_1$. The first term also has limit 0 due to the assumed condition 3. We appeal again to lemma 2.8.3 (ii) and (iii), and lemma 2.8.2 to argue that $\lim_{L\to\infty}\lim_{n\to\infty}\text{var}({}_k\underset{\sim}{\phi}_n^L) = 0$. Applying eq. (2.8), we conclude that:

$$\lim_{L\to\infty}\lim_{n\to\infty}\text{var}(\underset{\sim}{\phi}_n^L) = \lim_{L\to\infty}\lim_{n\to\infty}\text{var}\left(\sum_{k=1}^{K} {}_k\underset{\sim}{\phi}_n^L\right) = 0.$$

One more application of eq. (2.8) gives $\lim_{L\to\infty}\lim_{n\to\infty}\text{var}(\tilde{\phi}_n^L) = \lim_{n\to\infty}\text{var}(\phi_n)$. We conclude:

$$\lim_{n\to\infty}\frac{\text{var}(\sum_{r=1}^n \varrho_i)}{R_{11}^{(n)}/\pi_1} = \lim_{M\to\infty}\lim_{n\to\infty}\text{var}(\overline{\phi}_n^M) =^a \lim_{M\to\infty}\sum_{k=1}^{K}\pi_{\mathcal{A}_k}^{\infty}(\overline{\mu}^M - c_k)^2 = \sum_{k=1}^{K}\pi_{\mathcal{A}_k}^{\infty}(\mu - c_k)^2,$$

where (a) follows from the bounded version of the theorem proved above.

To prove the remark, consider $\mathcal{A}_k \cup \mathcal{A}_l$ as one subset. We can safely ignore the cross terms in eq. 2.11, without needing to use condition 1 for the pair $\mathcal{A}_k, \mathcal{A}_l$. We do not use condition 1 in the remaining part of the proof.

All that remains is to note:

$$(\overline{\mu}^M - c_k)^2 \frac{1}{Q_{11}/\pi_1} \sum_{i,j \in \mathcal{A}_k \cup \mathcal{A}_l} \pi_i \sum_{r=1}^{n} {}_1p_{ij}^{(r)} \to \pi^{\infty}_{\mathcal{A}_k \cup \mathcal{A}_l} (\overline{\mu}^M - c_k)^2$$

where $\pi^{\infty}_{\mathcal{A}_k \cup \mathcal{A}_l} = \pi^{\infty}_{\mathcal{A}_k} + \pi^{\infty}_{\mathcal{A}_l}$. □

Chapter 3

Source coding

3.1 Introduction

We are interested in both lossless and lossy source coding. Let us first consider the lossless case.

Let (X_n) be a discrete, ergodic source taking values in a finite set $\mathbb{K}$. For each n, we consider the problem of efficiently representing (X_n) using variable length block codes $\psi(x_1^n) : \{0,1\}^n \to \{0,1\}^*/\{\emptyset\}$, which map X_1^n to a variable length binary string. Let $l_n(X_1^n)$ be the length of $\psi(X_1^n)$ (i.e. the description length at block size n). We allow any mapping that constitutes a valid code, i.e. any invertible mapping ψ. The source coding theorem in Shannon's original paper (57) states that the average coding rate $\frac{1}{n}E[l_n(X_1^n)]$ of a memoryless information source cannot be made smaller than its entropy. A stronger, pointwise version of this theorem (32) can be stated as:

Theorem 3.1.1. *For a memoryless source (X_n),*

$$\liminf \frac{1}{n} l_n(X_1^n) \geq H(X_1) \quad a.s.$$

where the entropy is defined by $H(X_1) = E[-\log P(X_1)]$.

Despite its simplicity, this theorem can be made remarkably general by only replacing the entropy $H(X)$ with the entropy rate $\nu = \lim \frac{1}{n} H(X_1^n) := \lim \frac{1}{n} E[-\log P(X_1^n)]$. Then the result holds for any source for which the limit exists:

Theorem 3.1.2.

$$\liminf \frac{1}{n} l_n(X_1^n) \geq \nu \quad a.s.$$

whenever $\nu = \lim \frac{1}{n} H(X_1^n)$ *is well defined.*

The Shannon-McMillan-Breiman theorem ensures that this is true for all ergodic sources (1):

Theorem 3.1.3.

$$\lim -\frac{1}{n} \log P(X_1^n) = E[-\log P(X_n | X_{-\infty}^n)] \quad a.s.$$

for all ergodic information sources (X_n).

Therefore, the literature around theorem 3.1.1 is fairly complete. Unfortunately, the same cannot be said about rates of convergence questions regarding theorem 3.1.1. The behavior of the quantity $l_n(X_1^n) - n\nu$ has been discussed in the work of Kontoyiannis (33) under the title 'pointwise redundancy' in source coding. This work is limited mostly to the memoryless case, where it has been demonstrated that $l_n(X_1^n) - n\nu$ behaves like a random walk, and a one sided central limit theorem can be stated for the code length sequence:

Theorem 3.1.4.

$$\frac{l_n(X_1^n) - n\nu}{\sqrt{n}} \geq \sigma^2 g_n,$$

where g_n *is a sequence of random variables with* $g_n \to^d N(0,1)$.

This theorem is a result of 'Barron's lemma', which gives a lower bound on l_n for any encoding sequence:

Lemma 3.1.5 (Barron's Lemma)**.** *For any sequence* $\{c(n)\}$ *of positive constants with* $\sum 2^{-c(n)} < \infty$ *we have*

$$l_n(X_1^n) \geq -\log P(X_1^n | X_{-\infty}^0) - c(n), \ \textit{eventually, a.s.} \tag{3.1}$$

Here $l_n(X_1^n)$ is a code length sequence for the first n symbols of the source, the distribution of which is conditioned on the infinite past, for some lossless coding algorithm. $c(n)$ is a deficit term that can be made logarithmic in n (e.g. $c(n) = 2\log n$).

The first term on the right hand side in 3.1.5 can be written as:

$$-\log P(X_1^n | X_{-\infty}^0) = \sum_{i=1}^{n} -\log P(X_i | X_{-\infty}^{i-1}) := \sum_{i=1}^{n} \rho_i$$

where we refer to the process (ρ_n) as the information density.

In the memoryless case, clearly (ρ_n) is an i.i.d. sequence, from which the result in theorem 3.1.4 follows immediately. When the process (X_n) has memory, the result is less immediate, as one needs to deduce properties of the function (ρ_n) from the process (X_n). It has been shown that the redundancy process exhibits similar asymptotics in the case of a fast mixing Markov chain (33). Here, we take up this question for a class of slower mixing processes, where memory effects are much more prominent. This is the class of long range dependent sources.

3.1.1 The lossy case

We now discuss lossy source coding. Needless to say, the lossy case has deeper technical challenges, mostly due to the fact that the reconstruction distribution is different than the source distribution and is now an optimization parameter. As a result, even though analogues of lossless theorems exist, they are much less useful and difficult to evaluate.

Consider a mapping $\phi_n(X_1^n) : \mathbb{K}^n \to \hat{\mathbb{K}}^n$, where $\hat{\mathbb{K}}$ denotes an output alphabet. We define a distortion function $d_n(x_1^n; y_1^n) = \frac{1}{n}\sum_1^n d(x_i; y_i)$ with $x_i \in \mathbb{K}$ and $y_i \in \hat{\mathbb{K}}$. We consider the problem of finding efficient mappings $\phi_n(X_1^n)$ with the property $d_n(X_1^n; \phi_n(X_1^n)) \leq D$. Let $\psi_n(\phi_n(x_1^n)) : \{0,1\}^n \to \{0,1\}^*/\{\emptyset\}$ map $\phi_n(X_1^n)$ to a variable length binary string. Let $l_n(X_1^n)$ be the length of $\psi_n(\phi_n(X_1^n))$ (i.e. the description length at block size n). We allow any mapping that constitutes a valid code, i.e. any invertible mapping ψ_n.

Consider a random code construction where codewords are picked from an infinite i.i.d. codebook drawn from a n-symbol codebook distribution Q_n on $\hat{\mathbb{K}}^n$. ϕ_n maps X_1^n to the first codeword in the codebook which is within distortion D of X_1^n. The index of this match is subsequently sent to the receiver and ψ_n is the Elias encoding of this index. It is shown in (35) that for this code construction:

Lemma 3.1.6. *For any sequence $\{c(n)\}$ of positive constants with $\sum 2^{-c(n)} < \infty$ we have*

$$l_n(X_1^n) \geq -\log Q_n(B(X_1^n, D)) - c(n), \text{ eventually, a.s.} \tag{3.2}$$

Here $Q_n(B(X_1^n, D))$ is the probability of a distortion ball $B(X_1^n, D) := \{y_1^n \in \hat{\mathbb{K}}^n : d_n(X_1^n; y_1^n) \leq D\}$ of radius D around X_1^n under the n-letter codebook distribution Q_n. Picking the codebook distribution to optimize expected code length, we get $Q_n^* := \arg\max_{Q_n} E[-\log Q_n(B(X_1^n, D))]$, giving us a tighter bound for the above class of codes. It can also be shown, however, that this class of random codes are optimal, and so the optimized bound is valid not just for this class of codes, but in general (35):

Lemma 3.1.7. *For any sequence $\{c(n)\}$ of positive constants with $\sum 2^{-c(n)} < \infty$ we have*

$$l_n(X_1^n) \geq -\log Q_n^*(B(X_1^n, D)) - c(n), \text{ eventually, a.s.} \tag{3.3}$$

for any code which operates under fixed distortion D.

In the memoryless case, the output distribution with minimal $E[l_n]$ is a product distribution $Q_n^* = (Q^*)^n$, where Q^* is the single letter optimal output distribution that minimizes the information between the input and output subject to the distortion constraint. In this case, $-\log Q_n(B(X_1^n, D))$ can again be approximated as a sum of i.i.d. variables, and analogues of theorem 3.1.4 can be proved (35). Unfortunately, for sources with memory, very little is known about the optimal output distribution. In fact, rate distortion functions can be calculated exactly only in a few special cases (strongly connected finite state Markov chains (27), Gaussian autoregressive processes (26)) and even for those, only for a small range of low distortion. For all other processes with memory, one needs to work with bounds, which is useless for second order discussions.

Consequently we will only be able to extend our results to the lossy case in a limited manner. In section 3.3, we prove an alternative lossy Barron's lemma which is more usable and intuitive than (3.1.7), but only tight when the rate distortion function equals the Shannon lower bound (47). This leads to a second order converse lossy coding theorem for this class of information sources. We demonstrate a class of LRD processes for which the rate distortion function matches the Shannon lower bound, and thus can be calculated exactly, for a non-zero range of distortions. For this class of sources, we are able to conclude that the bit length process at the output of any lossy coder operating at the rate distortion function must exhibit LRD behavior.

3.1.2 Summary of results

Recall that a stationary random process (X_n) with $E[X_n^2] < \infty$ is said to be long range dependent (LRD) if

$$\limsup_{n\to\infty} \sum_{r=1}^{n} \operatorname{cov}(X_0, X_r) = \infty. \tag{3.4}$$

The degree of long range dependence is measured by the Hurst index H ($\frac{1}{2} \leq H \leq 1$).

$$H := inf \left\{ h : \limsup_{n \to \infty} \frac{\sum_{r=1}^{n} \mathrm{cov}(X_0, X_r)}{n^{2h-1}} < \infty \right\}.$$

Equivalently, we can write:

$$H := inf \left\{ h : \limsup_{n \to \infty} \frac{\mathrm{var}(\sum_{i=1}^{n} X_i)}{n^{2h}} < \infty \right\}.$$

A process that is not LRD is said to be short range dependent (SRD). The justification for this division is as follows. Although SRD processes may have memory, the effect of this memory can be ignored in asymptotic discussions by taking long blocks of the original source and treating these blocks as a meta process. If the process is SRD, for many practical purposes, the block process can be well approximated by an i.i.d. process. As a result, SRD processes behave similarly to an i.i.d. process in many asymptotic settings. Indeed, the complement of the condition in (3.4) is necessary for a central limit theorem to hold.

On the other hand, the effect of memory in an LRD process does not disappear even asymptotically under any scaling. LRD processes do not satisfy the central limit theorem, and the limit of their scaled sums is not in general Gaussian. In fact, as the definitions suggest, the scaling required to obtain a meaningful limit is different than $\sqrt{n}$, and is related to the Hurst index of the process. The limiting processes, when they exist, are described by stable distributions and self similar processes. For a detailed description of these results and other properties of LRD processes, the reader is referred to the references (54)(17).

Interest in LRD processes was sparked by several empirical observations that showed such distributions were characteristic of network traffic on the internet (36)(13)(49). Due to the fundamentally different qualities of LRD processes mentioned above, these discoveries have important, and often negative consequences for the modeling and analysis of communication networks. Among these are different

asymptotics for queue sizes and packet drop probabilities (51; 38; 37; 28; 63; 20), and a need for new optimal schedulers (2)(48)(53). The mostly degrading effect of LRD traffic in networks has led to research efforts for understanding the mechanisms by which such traffic is generated and whether preventive measures are possible (48)(13).

In section 3.2, first we describe a general model of an LRD information source, which can be written as an instantaneous function of an LRD Markov chain. Our source model includes renewal processes and semi-Markov processes as special instances. We then seek to prove second order converse lossless source coding theorems for these sources. This is done by first interpreting the information density of the source as a function of a Markov chain that is related to that which underlies the source, and then applying theorems stated in chapter 2 to characterize the Hurst index of the information density process. An application of Barron's lemma leads to the main result in this section, which is that the code length process of an LRD information source (one which can be described in our model) necessarily dominates an LRD process with the same Hurst index as the original source, under any lossless coding scheme. Moreover, this second order asymptotics is achievable to within $O(\log n)$.

As mentioned below, the application of Barron's lemma in the lossy case requires the characterization of the optimal output distribution. This problem is notoriously difficult for sources with memory, even in the simplest case of a binary symmetric Markov chain (31). Nevertheless, in section 3.3, we are able to find a class of LRD sources, for which we are able do this for a constrained range of distortions. We prove second order pointwise lossy source coding theorems for this class of sources. We also demonstrate achievability of this bound within $O(\sqrt{n}\log n)$, which is sufficient to prove long range dependence at the output of an optimal encoder.

3.2 Lossless coding

$(X_n) \in \mathbb{K}$ is a discrete, ergodic source. $\psi(x_1^n) : \{0,1\}^n \to \{0,1\}^*/\{\emptyset\}$ maps X_1^n to a variable length binary string in an invertible fashion. Let $l_n(X_1^n)$ be the length of $\psi(X_1^n)$. Source coding is primarily concerned with obtaining the 'shortest possible' $l_n(X_1^n)$. We know from 3.1.3 that the optimal description length in the mean sense is given by average entropy density $\nu = E[\rho_n] = E[-\log P(X_n|X_{-\infty}^n)]$. Here we are concerned with the asymptotic pointwise redundancy $l_n(X_1^n) - \nu n$.

When (X_n) are i.i.d. or sufficiently short range dependent, this difference process is well approximated by a random walk, exhibiting fluctuations of order $O(\sqrt{n})$ (33). When (X_n) is an LRD process, one would expect that the difference process would be at least as variable as the source process, exhibiting fluctuations at the scale of $O(n^H)$, where H is the Hurst index of (X_n). The first such result for LRD renewal processes was given in (45). This result states the following:

Theorem 3.2.1. *Let (X_n) be a long range dependent ergodic renewal process. Assume*

$$\kappa = \sup\{k : E[|T|^k] < \infty\} > 1 \tag{3.5}$$

Then, there exists a long range dependent random process (ξ_n) such that

$$l_n(X_1^n) \geq \sum_{i=1}^{n} \xi_i, \ \textit{eventually, a.s.}$$

for all uniquely decodable source codes. Moreover, (ξ_n) has the same Hurst index as (X_n).

The theorem is derived from the following result for the entropy density process ρ of an LRD renewal process:

Lemma 3.2.2. *Let (X_n) be a long range dependent ergodic renewal process and $\rho_n =$*

$-\log P(X_n|X_{-\infty}^{n-1})$. *Then*

$$\lim_{n\to\infty} \frac{var(\rho_0 + \ldots + \rho_n)}{var(X_0 + \ldots + X_n)} = C$$

for some $0 < C < \infty$.

Asymptotically, the behavior of the aggregate entropy density is identical to the original process, up to a constant factor. Applying Barron's lemma, we get the theorem. In fact, we also know that codes exist which can achieve within $O(\log n)$ of $\sum_{i=1}^{n} \rho_i$ (33). Therefore, from lemma 3.2.2, we can also deduce that the bound in theorem 3.2.1 is achievable, giving us a complete second order source coding theorem for an LRD information source.

We will now replicate the same arguments to generalize these theorems further.

Definition 3.2.3. *A process* (X_n) *is said to have countable memory if* $\mathcal{L}(X_n^{\infty}|X_{-\infty}^{n-1}) = \mathcal{L}(X_n^{\infty}|g(X_{-\infty}^{n-1}))$ *where* $\mathcal{L}(X_n^{\infty}|X_{-\infty}^{n-1})$ *denotes the regular conditional distribution of* X_n^{∞} *given* $X_{-\infty}^{n-1}$, g *is a deterministic function that maps to the integers and* $g(X_{-\infty}^{n}) = f(X_n, g(X_{-\infty}^{n-1}))$ *for some deterministic function* f.

Let (X_n) be a discrete, ergodic, finite state information source with countable memory. Pick $\tilde{M}_n$, a stationary, ergodic, countable state Markov chain with $\mathcal{F}(\tilde{M}_n) \supset \mathcal{F}(X_n, g(X_{-\infty}^{n}))$. The pair $(X_n, g(X_{-\infty}^{n}))$ itself is one such Markov chain. Define $M_n = (\tilde{M}_{n-1}, \tilde{M}_n)$ to be an extended Markov chain on the product space. Then it is easy to show

Lemma 3.2.4. $\rho_n = -\log P(X_n|X_{-\infty}^{n-1})$ *is an instantaneous function of* M_n. *i.e.* $\rho_n = h(M_n)$ *where* h *is a deterministic function.*

Proof. $\rho_n = -\log P(X_n|X_{-\infty}^{n-1}) = -\log P(X_n|g(X_{-\infty}^{n-1}))$, which is clearly a function of M_n. □

We are now ready to state our first theorem.

Theorem 3.2.5. *Let (X_n) be a discrete, ergodic, finite state information source with countable memory with associated extended Markov chain M_n and information density ρ_n. If $\rho_n = \rho(M_n)$ satisfies the conditions of theorem 2.3.2 for a suitable numbering of the state space, then there exists an LRD process (ξ_n) with mean ν such that the code length process of any lossless code satisfies*

$$l_n(X_1^n) \geq \sum_{i=1}^{n} \xi_i - O(\log n) \quad \textit{eventually a.s.}$$

Moreover, (ξ_n) has the same Hurst index as (X_n).

Proof. From lemma 3.1.5 we know that

$$l_n(X_1^n) \geq -\log P(X_1^n | X_{-\infty}^0) - O(\log n), \text{ eventually, a.s.}$$

Since $-\log P(X_1^n|X_{-\infty}^0) = \sum_{i=1}^n \rho_i$, it remains to show that (ρ_n) is LRD with the same Hurst index as M_n. We get this by applying theorem 2.3.2, as we already assumed that the conditions for this theorem are satisfied. □

At this point, with the countable memory assumption and the rather cryptic conditions of theorem 2.3.2, it is not clear whether this theorem is useful. Therefore, we now demonstrate that fairly general classes of processes with wide applicability can be regarded as special cases of this formulation.

3.2.1 Semi-Markov processes

A semi-Markov process (X_n) is defined in terms of a transition probability matrix $q(k, l)$ and renewal process A_n. X_n is equal to X_{n-1} when $A_n = 0$. Transitions in (X_n) occur when there is an arrival in A_n according to the transition probability matrix.

Formally, let $k, l \in \mathbb{K}$ for some finite set $\mathbb{K}$ and $A_n \in \{0, 1\}$. Then (X_n) is a semi-Markov process (SMP) if $P(X_n = l|X_{n-1} = k, A_n = 0) = \delta(k = l)$, $P(X_n = l|X_{n-1} = k, A_n = 1) = q(k, l)$. We assume $q(k, k) = 0$.

Define a Markov chain in terms of the pair (X_n, T_n), where $X_n \in \mathbb{K}$ and T_n denotes the time since the last transition in A_n (i.e. $\{T_n = j\} = \{\inf_i\{i \geq 0, A_{n-i} = 1\} = j\}$). We will say that a semi-Markov process is LRD whenever the associated Markov chain (X_n, T_n) is LRD. Define $M_n = (X_n, X_{n-1}, T_{n-1})$ to be another Markov chain. We have that $\rho_n = -\log P(X_n|X_{-\infty}^{n-1}) = -\log P(X_n|X_{n-1}, T_{n-1})$ is a function of M_n. Assume a numbering of the states (X_n, X_{n-1}, T_{n-1}) on $\mathbb{N}$. We will be using the index i to refer to this numbering. Given an LRD SMP, we will attempt to show that ρ_n is also LRD with the same Hurst index.

To make this statement meaningful, we first show

Lemma 3.2.6. *$E[\rho_n^2] < \infty$.*

Proof. Denote $P_{m|kl} = P(X_n = m|X_{n-1} = k, T_{n-1} = l)$ and $P_{kl} = P(X_{n-1} = k, T_{n-1} = l)$.

$$E[\rho_n^2] = \sum_{k,l,m} P_{kl} P_{m|kl} \log^2 P_{m|kl} \leq \sum_{k,l,m} 2P_{kl} = 2K$$

since $P \log^2 P$ terms are bounded above by 2. □

Lemma 3.2.7.

$$\lim_{n\to\infty} \frac{var(\sum_{r=1}^{n} \rho_r)}{R_{11}^{(n)}/\pi_1} = C$$

where C is a finite, non-zero constant.

Proof. We will apply theorem 2.3.2 with the following partitions , $\mathcal{A}_1 = \{T_n = 0\}$, $\mathcal{A}_2 = \{T_n > 0, \rho(i) \leq -\log \epsilon_i\}$, and $\mathcal{A}_{3,m} = \{T_n > 0, \rho(i) > -\log(\epsilon_i), X_n = m\}, m \in \mathbb{K}$. Here we will choose $\epsilon_i \uparrow 1$ later. Take $c_1 = c_2 = 0$, $c_{3,m} = 0, \forall m$ and $\mathcal{H} = \{T_{n-1} =$

0} in the theorem. By the remark to the theorem, we don't need condition 1. We will check conditions 2 and 3 of theorem 2.3.2 for each of the sets.

When $i, j \in \mathcal{A}_1$ notice ${}_{\mathcal{H}}p_{ij}^{(r)} = 0$, so both conditions hold automatically. For $i, j \in \mathcal{A}_2$, condition 2 holds because the limit of $\rho(i)$ as $i \to \infty$ is zero, (see the remarks to theorem 3.1 in (46)) and condition 3 holds because ρ is bounded on this set. Thus we focus on $i, j \in \mathcal{A}_{3,m}$. Define $\rho(i) =: -\log(\tilde{\epsilon}_i)$. Let subsequence $\{i_k\} = \mathcal{A}_{3,m}$, ordered such that if $i_k = (m, m, T_1)$ and $i_j = (m, m, T_2)$, $T_1 > T_2 \iff k > j$. We have $\tilde{\epsilon}_{i_k} \leq \epsilon_{i_k}$. $\pi_{i_k} \leq \prod_{l=1}^{k} \tilde{\epsilon}_{i_k}$, and $\sum_1^\infty {}_1p_{i_k i_j}^{(r)} = \pi_{i_j}/\pi_{i_k}$. We have

$$\begin{aligned}
&\sum_{i \in \mathcal{A}_{3,m}} \rho(i)\pi_i \sum_{j \in \mathcal{A}_{3,m}} \rho(j) \sum_{r=1}^{n} {}_1p_{ij}^{(r)} \\
\leq &\sum_k -(\prod_{l=1}^{k} \tilde{\epsilon}_{i_l}) \log \tilde{\epsilon}_{i_k} \sum_{j>k} (-\log \tilde{\epsilon}_{i_j}) \prod_{l=k+1}^{j} \tilde{\epsilon}_{i_l} \\
\leq &\sum_j \sum_{k<j} \prod_{l=1}^{j} \tilde{\epsilon}_{i_l}(1 - \log \tilde{\epsilon}_{i_l}) \\
< &\sum_j j \prod_{l=1}^{j} \tilde{\epsilon}_{i_l}(1 - \log \tilde{\epsilon}_{i_l}).
\end{aligned}$$

We can easily choose $\epsilon_i \uparrow 1$ such that this is finite. Dividing by $Q_{11}^{(n)}$, we will obtain a 0 limit (since $Q_{11}^{(n)} \to \infty$, 2.3) satisfying both conditions in theorem 2.3.2. □

Now we state our theorem for LRD SMPs.

Theorem 3.2.8. *Let (X_n) be a long range dependent semi-Markov process. Then, there exists a long range dependent random process (ξ_n) such that*

$$l_n(X_1^n) \geq \sum_{i=1}^{n} \xi_i - O(\log n), \text{ eventually, a.s.}$$

for all uniquely decodable source codes. Moreover, (ξ_n) has the same Hurst index as (X_n).

Proof. The LRD process (ξ_n) in the theorem is essentially (ρ_n). This is seen directly from lemma 3.1.5. The fact that (ρ_n) is LRD with the same Hurst index as (X_n) follows from lemma 3.2.7. □

3.2.2 Generalized semi-Markov processes

The definition of the SMP can be generalized substantially by allowing the transition matrix $q(k,l)$ to depend on T_n. This construction is equivalent to a class of generalized semi-Markov processes (GSMPs), for which the set of living events is disjoint for each state (also see e.g. (55) for a description of a GSMP). In this construction, with each state k in the finite set $\mathbb{K}$ is associated a set of events $\mathcal{E}_k$ each with its own timer which expires after a time T_e for every $e \in \mathcal{E}_k$. Assume each timer has a distribution $T_e \sim F_e$, which is continuous, so that no two timers expire at the same time. Let e be the event for which the first timer expires. Then the process (X_n) will jump at time $n = \lceil T_e \rceil$ according to a transition matrix $q_e(k,l)$ which is allowed to depend on e. We again assume that $q_e(k,k) = 0$. At this point all timers are reset, and new timers are generated for each event in $\mathcal{E}_l$ corresponding to the new state l.

We can reframe this construction as follows. Let (X_n) be a semi-Markov process, where the transition probability matrix $q(k,l,T)$ depends on the time since the last transition. Set T to have the (integer valued) distribution defined as follows: $P(T = n) = P(n-1 \leq \min T_e < n)$. Also define $q(k,l,n) = \sum_{e \in \mathcal{E}_k} q_e(k,l) P(\text{argmin} T_e = e | T = n)$.

Note again that the tuple (X_n, T_n) is a Markov chain. We say that the GSMP is long range dependent, when this Markov chain is long range dependent. We prove the following theorem for this construction.

Theorem 3.2.9. *Let (X_n) be a long range dependent GSMP. Then, there exists a*

long range dependent random process (ξ_n) *such that*

$$l_n(X_1^n) \geq \sum_{i=1}^{n} \xi_i - O(\log n), \ \textit{eventually, a.s.}$$

for all uniquely decodable source codes. Moreover, (ξ_n) *has the same Hurst index as* (X_n)*.*

Proof. Simply use the same partitioning of the state space as was done in the proof of lemma 3.2.7, and observe that the proof does not depend on the transition probabilities. □

3.2.3 Achievability and Wyner-Ziv waiting times

It is well known that the above lower bounds are achievable by many classes of optimum lossless codes. e.g. Huffman coding achieves $l_n(X_1^n) \leq -\log P(X_1^n) + O(1)$. It is interesting to also consider the performance of algorithms based on the popular Lempel-Ziv compressor.

Lempel-Ziv type lossless compression schemes (62) have been immensely popular due to their practicality and universality. The central idea in these schemes is to use the past realization of the random process as a codebook for compression. The input string is incrementally partitioned into phrases, each phrase corresponding to the shortest substring that has not so far occurred in the past phrases. The encoder then sends an index of the matched phrase to describe the new phrase. Matched phrases are likely to get longer as more of the string is partitioned in this way, since more phrases are continually added to the codebook.

In an idealized version of this scheme, we may imagine that a two sided stationary process is being compressed, with the infinite past of the process already decoded and available at the receiver (58; 59). To communicate the string X_1^n to the receiver, the

encoder looks for the closest index i in the past where this string appears exactly, i.e.

$$\min\{i : X_{-i}^{n-i+1} = X_1^n\}.$$

This index is then sent to the receiver, using the Elias encoding for integers (19) using only $\log i + \log\log i + 1$ bits.

The index i is referred to as the 'recurrence time' of the string X_1^n. The recurrence time is related to the probability $P(X_1^n)$ using the following result.

Theorem 3.2.10 ((34), theorem 1(i))**.** *Let (X_n) be a finite-valued stationary ergodic process, and c_n an arbitrary sequence of non-negative constants such that $\sum n2^{-c_n} < \infty$. For the recurrence times R_n we have*

$$\log R_n P(X_1^n) \leq c_n \text{ ev. as.}$$

Pick e.g. $c_n = 3\log n$ and note that the code lengths satisfy $l_n(X_1^n) \leq \log R_n(X_1^n) + \log\log R_n(X_1^n) + 1$. We can easily modify this scheme to transmit (X_1^n) exactly using at most $n\log|\mathbb{K}|$ bits whenever $\log R_n(X_1^n) > n\log|\mathbb{K}|$. This ensures that the code length is at most $\log R_n(X_1^n) + O(\log n)$. Combining this with the last theorem, we conclude that for the idealized Lempel-Ziv scheme,

$$l_n(X_1^n) = -\log P(X_1^n) + O(\log n).$$

3.3 Lossy coding

We consider the problem of representing X_1^n within distortion D_n, i.e. $\frac{1}{n}\sum_{i=1}^n d(X_i; Y_i) \leq D_n$. Define $\phi_n(X_1^n) = Y_1^n$ to be the code at block length n, and $l_n(X_1^n)$ be the corresponding representation length defined by the invertible mapping $\psi : \hat{\mathbb{K}}^n \to \{0,1\}^*/\{\emptyset\}$. We can show that

Lemma 3.3.1. *Let c_n satisfy $\sum 2^{-c_n} < \infty$.*

$$l_n(X_1^n) \geq -\log P(\phi_n(X_1^n)) - c_n \text{ ev. a.s.}$$

Proof.

$$\begin{aligned} P(l_n < -\log P(\phi_n(X_1^n)) - c_n) &= P(\frac{2^{-l_n}}{P(\phi_n(X_1^n))} > 2^{c_n}) \\ &\leq 2^{-c_n} E\left[\frac{2^{-l_n}}{P(\phi_n(X_1^n))}\right] \\ &= 2^{-c_n} \sum_{x_1^n} 2^{-l_n(x_1^n)} \\ &\leq 2^{-c_n}. \end{aligned}$$

Here the second equality follows simply by expanding the expectation, and the last inequality is a result of Kraft's inequality. The lemma follows through Borel-Cantelli lemma using the assumption on c_n. □

Rewriting the first term on the RHS,

$$-\log P(\phi_n(X_1^n)) = -\log \frac{P(X_1^n, \phi_n(X_1^n))}{P(X_1^n|Y_1^n)} \tag{3.6}$$

$$= -\log P(X_1^n) + \log P(X_1^n|Y_1^n). \tag{3.7}$$

This equation relates the aggregate entropy density process $-\log P(X_1^n)$ to two other quantities. The first one, $-\log P(\phi_n(X_1^n))$, is closely related to the bit-length process as we just argued. The second one, $-\log P(X_1^n|Y_1^n)$, will turn out to be related to the distortion process $d_n = d(X_n; Y_n)$. Consider for instance a distortion function such that X_n is recoverable from d_n and Y_n. Then, we would have $-\log P(X_1^n|Y_1^n) = -\log P(d_1^n|Y_1^n)$, which we can interpret as 'the information lost in distortion'.

Now, assuming that the source has LRD entropy density (e.g. it belongs to the class of sources described in section 3.2), this induces long range dependence in at least one of these two objects. Qualitatively, a tradeoff is revealed between rate (as

tied to the bit-length process) and distortion in the context of long range dependence. In this work, we will investigate a fixed distortion scenario, and show that the code length process must exhibit LRD.

To make this more concrete, assume $d(k;j)$ is a balanced distortion measure. Let (X_n) be a discrete, stationary, ergodic source taking values in a finite set $\mathbb{K}$. For each n consider a mapping $\phi_n(X_1^n) : \mathbb{K}^n \to \hat{\mathbb{K}}^n$ where the output alphabet $\hat{\mathbb{K}}$ is also finite. We consider balanced distortion measures $d_n(x_1^n; y_1^n) = \frac{1}{n}\sum_1^n d(x_i; y_i)$ with $x_i \in \mathbb{K}$, $y_i \in \hat{\mathbb{K}}$.

Definition 3.3.2. *$d(x;y)$ is said to be a balanced distortion measure whenever the set of possible values $d(\cdot;y)$ takes is identical for each $y \in \hat{\mathbb{K}}$.*

We are concerned with the problem of finding "minimum length" mappings $\phi_n(X_1^n)$ with the property $d_n(X_1^n; \phi_n(X_1^n)) \leq D$ for each n. Let $\psi_n : \hat{\mathbb{K}}^n \to \{0,1\}^*/\{\emptyset\}$, which maps $\phi_n(X_1^n)$ to a variable length binary string. Let $l_n(X_1^n)$ be the length of $\psi_n(\phi_n(X_1^n))$ (i.e. the description length at block size n). We allow any mapping that constitutes a valid code, i.e. any invertible mapping ψ_n.

It is well known (32) that the average behavior of $l_n(X_1^n)$ is bounded by the rate distortion function.

Definition 3.3.3. *(Rate distortion function)*

$$R_n(D) := \min_{P(X_1^n, Y_1^n): Ed(X_1^n; Y_1^n) \leq D} \frac{1}{n} I(X_1^n; Y_1^n),$$

$$R(D) := \lim_{n\to\infty} R_n(D).$$

Theorem 3.3.4. *((32), prop. 4)*

$$\liminf \frac{1}{n} l_n(X_1^n) \geq R(D) \quad a.s. \ .$$

Going beyond average behavior, one might be interested in the more fine grained problem of how close the code lengths can get to the rate distortion function. This

problem is referred to as the redundancy problem of lossy source coding. The average redundancy of code lengths $E[l_n(X_1^n)] - nR(D)$ has been studied in the works of (61; 60). There, the minimum average redundancy has been shown quite generally to be $O(\log n)$.

Here we are concerned with the pointwise redundancy $l_n(X_1^n) - nR(D)$. This problem was considered in the work (35), where it was proved that:

Theorem 3.3.5. *((35), theorem 6(ii)) For any sequence $\{c_n\}$ of positive constants with $\sum 2^{-c_n} < \infty$,*

$$l_n(X_1^n) \geq -\log \tilde{Q}_n(B(X_1^n, D)) - c_n \quad \textit{eventually a.s. .}$$

Here $B(X_1^n, D)$ is the distortion ball defined by

$$B(X_1^n, D) := \{y_1^n \in \hat{\mathbb{K}}^n : d_n(X_1^n; y_1^n) \leq D\},$$

and $\tilde{Q}_n$ is the probability measure that minimizes $E[-\log Q_n(B(X_1^n, D))]$ under all probability measures Q_n on $\hat{\mathbb{K}}^n$.

Unfortunately, very little can be said about the measures $\tilde{Q}_n$, except when (X_n) is i.i.d, in which case $-\log \tilde{Q}_n(B(X_1^n, D)) = -\log Q_n^*(B(X_1^n, D)) - O(\log n)$ a.s. where Q_n^* is a product distribution. We aim to produce a more workable lower bound to $l_n(X_1^n) - nR(D)$.

Our main result will be the following.

Theorem 3.3.6. *Let $\nu = E[-\log P(X_1|X_{-\infty}^0)]$ be the entropy rate of (X_n). Then*

$$l_n(X_1^n) \geq -\log P(X_1^n|X_{-\infty}^0) - n(\nu - R_l(D)) - O(\log n) \quad \textit{ev. a.s.}$$

Here $R_l(D)$ is the Shannon lower bound to the rate-distortion function, to be defined in the next section. The advantage of this bound over that in theorem 3.3.5 is

that the quantity $-\log P(X_1^n|X_{-\infty}^0)$ can be written as a running sum of a stationary random process as

$$-\log P(X_1^n|X_{-\infty}^0) = \sum_{i=1}^{n} -\log P(X_i|X_{-\infty}^{i-1}). \tag{3.8}$$

The random process $\rho_n = -\log P(X_n|X_{-\infty}^{n-1})$ is referred to as the *entropy density* process. Consequently, the asymptotic (second order) behavior of $l_n(X_1^n)$ can generally be inferred from limit theorems on $\sum \rho_i$, as the stationary ergodic process (ρ_n) typically inherits the mixing properties of the source (X_n). The caveat is that the RHS of 3.3.6 has mean $nR_l(D) - O(\log n)$, meaning that if the Shannon lower bound is not tight, the bound is of little interest.

To put this restriction in context, we point out that in the literature of rate-distortion theory for sources with memory, complete results are rare even in first order discussions (i.e. calculation of $R(D)$). In fact, rate distortion functions can be calculated exactly only in a few special cases (finite state Markov chains with strictly positive transition matrices (27), Gaussian autoregressive processes (26)) and even for those, only for a small range of low distortion. These examples have the property that the Shannon bound to the rate-distortion function is tight. For all other processes with memory, one needs to work with bounds on the rate-distortion function, which is useless for second order discussions.

In the next section we will define the Shannon lower bound to the rate distortion function for balanced distortion measures, and discuss the conditions under which it is tight. In section 3.5, we present the proof of our main theorem 3.3.6. Then we discuss applications of this theorem to fast mixing sources in section 3.6. A one sided central limit theorem for such sources is given, as well as a discussion of minimum coding variance. In section 3.7, we proceed to discuss long range dependent sources. We show through an example that there exist information sources which exhibit long

range dependent code lengths under any coding scheme operating at the Shannon lower bound with fixed distortion.

3.4 Shannon lower bound

The Shannon lower bound (SLB) to the rate-distortion function is defined as follows, (see e.g. (12), problem 10.6.):

Definition 3.4.1 (Shannon lower bound)**.**

$$R_l(D) := \nu - \max_{X:Ed(X;0)\leq D} H(X).$$

Lemma 3.4.2.

$$R_n(D) \geq nR_l(D)$$

Proof. Let $X_1^n \sim P_{x_1^n}$.

$$\begin{aligned}
\min_{\substack{X_1^n \sim P_{x_1^n} \\ Ed_n(X_1^n;Y_1^n)\leq D}} I(X_1^n;Y_1^n) &= H(X_1^n) - \max_{\substack{X_1^n \sim P_{x_1^n}, \\ Ed_n(X_1^n;Y_1^n)\leq D}} H(X_1^n|Y_1^n) \\
&\geq n\nu - \max_{Ed_n(X_1^n;Y_1^n)\leq D} H(X_1^n|Y_1^n) \\
&\overset{(a)}{=} n\nu - \max_{Ed_n(X_1^n;y_1^n)\leq D} H(X_1^n|Y_1^n = y_1^n) \\
&= n\nu - n \max_{X:Ed(X;y)\leq D} H(X)
\end{aligned}$$

Where the min and max are over joint distributions $P(X_1^n, Y_1^n)$. (a) follows because the distortion is balanced. The last equality follows because $H(X_1^n)$ is maximized by a product distribution on X_1^n, and by the concavity of entropy. □

3.4.1 Tightness of the SLB

Let $x, y \in \mathbb{K}$ where $\mathbb{K}$ is an additive group. If the distortion can be written as $d(x;y) = d(x-y)$ for some function $d : \mathbb{K} \to \mathbb{R}$, then d is referred to as a difference

distortion measure. For difference distortion measures, the case in which the SLB is tight is characterized by the following theorem:

Theorem 3.4.3. *(Theorem 4.3.1 in (6))*

$R_l(D) = R(D)$ iff the source r.v. X can be expressed as the sum of two statistically independent random variables one of which is distributed according to the probability distribution that maximizes the expression in 3.4.1. i.e $X_n = Y_n + Z_n$ where $H(Z_n)$ $= \max_{X:Ed(X,0)\leq D} H(X)$.

Proof. We will produce a summary of the proof in (6) as it will be useful later. We find the rate distortion function by maximizing $I(X_1^n, Y_1^n)$ subject to $X_1^n \sim P_n$ over all joint distributions $Q_n(Y_1^n|X_1^n)$. For a given $Q_n(Y_1^n|X_1^n)$, we can also define $Q_n(Y_1^n) = \sum_{x_1^n} P_n(x_1^n)Q_n(Y_1^n|x_1^n)$. For most of the proof, we regard the sequences X_1^n and Y_1^n as discrete random variables. To keep notation uncluttered, we will simply write X for X_1^n and Y for Y_1^n. We will use the time indices when they are necessary.

We are given the following optimization problem:

$$\max_{Q(y|x)} I(X;Y) = \sum_{x,y} P(x)Q(y|x)\log\frac{Q(y|x)}{Q(y)} \tag{3.9}$$

$$\text{s.t. } Q(y|x) \geq 0 \tag{3.10}$$

$$\sum_y Q(y|x) = 1 \quad \forall x \tag{3.11}$$

$$\sum_{x,y} P(x)Q(y|x)d(x;y) \leq D. \tag{3.12}$$

This is a convex optimization problem. To solve it analytically, we ignore the first set of constraints, and introduce Lagrange multipliers μ_x and $s < 0$ for the next two, giving the Lagrangian,

$$J(Q) = \sum_{x,y} P(x)Q(y|x)\log\frac{Q(y|x)}{Q(y)} - \sum_x \mu_x \sum_y Q(y|x) - s\sum_{x,y} P(x)Q(y|x)d(x;y). \tag{3.13}$$

At this point, we make the substitution $\log \lambda_x = \frac{\mu_x}{P(x)}$. Now taking derivatives with respect to the $Q(y|x)$ and setting them equal to zero, we find that the optimal solution should satisfy

$$Q(y|x) = \lambda_x Q(y) e^{sd(x;y)} \tag{3.14}$$

$$\lambda_x = (\sum_y Q(y) e^{sd(x;y)})^{-1} \tag{3.15}$$

$$Q(y|x) = \frac{Q(y) e^{sd(x;y)}}{\sum_y Q(y) e^{sd(x;y)}}. \tag{3.16}$$

Assuming for the moment that the solution to the optimization problem gives a vector $Q(y|x) > 0, \forall x, y$ (we skip here the argument that this entails no loss of generality, see (6) chapter 2, lemma 1), we can eliminate $Q(y|x)$ and write the $R(D)$ curve parametrically in terms of $Q(y) > 0, \lambda_x$, and s as

$$D = \sum_{x,y} \lambda_x P(x) Q(y) e^{sd(x;y)} d(x;y), \tag{3.17}$$

$$R = sD + \sum_x P(x) \log \lambda_x \tag{3.18}$$

with

$$\sum_x \lambda_x P(x) e^{sd(x;y)} = 1. \tag{3.19}$$

The multiplier s turns out to have a natural interpretation as the slope of the $R(D_s)$ curve which it parametrizes ((6), theorem 2.5.1). Note that the existence of a solution $Q(y|x) > 0, \forall x, y$ is necessary and sufficient for this formulation.

Now let $d(x;y) = d(x_1^n; y_1^n) = \sum_1^n d(x_i - y_i) := \sum_1^n d(z_i) := d(z)$ be a difference distortion measure. We will pick $\lambda_x = \frac{1}{KP(x)}$ to give a bound on $R(D)$:

$$R(D) \geq sD - \sum_x P(x) \log P(x) - \log K. \tag{3.20}$$

We set $K = \sum_x e^{sd(x;y)} = \sum_z e^{sd(z)}$ so that (3.19) is satisfied. Note that we are able to do this because the sum is now independent of y. Maximizing the bound with

respect to s, we see that

$$D = \sum_z d(z) \frac{e^{sd(z)}}{\sum_{z'} e^{sd(z')}}. \tag{3.21}$$

Combining (3.20) and (3.21) we can rewrite the bound as

$$R(D) \geq H(X) + \sum_z \frac{e^{sd(z)}}{\sum_{z'} e^{sd(z')}} \log \frac{e^{sd(z)}}{\sum_{z'} e^{sd(z')}} \tag{3.22}$$

$$= H(X) - H(Z) \tag{3.23}$$

where $Z = Z_1^n$ is the random variable having distribution $g(z) = g_n(z_1^n) = \frac{e^{s\sum d(z_i)}}{K}$. We note that this is the maximum entropy distribution subject to $Ed(Z) < D$, giving the expression for the Shannon lower bound.

To summarize, we observe that the SLB is tight if and only if there is a positive vector $Q(y)$, summing up to 1 and satisfying 3.14 with the given choice of λ_x, which can now be written as

$$Q_n(y_1^n|x_1^n)P_n(x_1^n) = Q_n(y_1^n)\frac{1}{K}\prod_1^n e^{sd(x_i-y_i)}. \tag{3.24}$$

We recognize this as the construction described in the statement of the theorem, namely that X_1^n can be constructed from $Y_1^n \sim Q_n$ by passing it through an i.i.d. channel with transition probability $\frac{e^{sd(x_i-y_i)}}{K}$.

□

Although the theorem is stated for difference distortion measures, the proof generalizes to balanced distortion measures without alteration (also see (27) for a partial discussion). To state the general version, let Φ_y, $y \in \hat{\mathbb{K}}$, be the permutation function with $d(x;y) = d(\Phi_y(x);0), \quad \forall x \in \mathbb{K}$, for a balanced distortion d. $\mathbb{K}$ and $\hat{\mathbb{K}}$ are now arbitrary finite sets. Then we have:

Theorem 3.4.4. *$R_l(D) = R(D)$ iff the source r.v. X admits the following characterization.*

$$X_n = \Phi_{Y_n}(Z_n)$$

where $Z_n \in \mathbb{K}$ are i.i.d and independent from Y_n with $H(Z_n) = \max_{X:Ed(X;0)\leq D} H(X)$.

Proof. Following the preceding proof where for difference distortion measures it is defined $d(z) := d(x-y) = d(x;y)$, for balanced distortions, we similarly define $d(z) := d(\Phi_y^{-1}(x)) = d(x;y)$. Equation 3.24 becomes:

$$Q_n(y_1^n|x_1^n)P_n(x_1^n) = Q_n(y_1^n)\frac{1}{K}\prod_1^n e^{sd(\Phi_{y_i}^{-1}(x_i))}.$$

This is equivalent to requiring that there exist a random variable Y_1^n such that the above construction is possible - $X_n = \Phi_{Y_n}(Z_n)$, with Z_n i.i.d. distributed according to $\frac{e^{sd(z)}}{\sum_z e^{sd(z)}}$. But this is the distribution which results from the maximization $\max_{X:Ed(X;0)\leq D} H(X)$ (with D parametrized by the value of s), proving the theorem. □

Immediate examples of information sources which admit such a characterization are explicit constructions where an underlying process is observed through a memoryless, time invariant channel (e.g. hidden Markov models). There also exist more surprising examples however, for instance some finite state Markov chains (27) and autoregressive processes (26).

While the Shannon lower bound is known to be asymptotically tight for small distortions quite generally (39), it is in general a difficult question as to when such a decomposition will exist.

3.5 Pointwise lower bound

Once the mapping ϕ_n has been chosen, the following lemma (33) provides a pointwise lower bound on the code length process.

Recall that by (3.3.1), for any sequence $\{c(n)\}$ of positive constants with

$\sum 2^{-c(n)} < \infty$ we have:

$$l_n(X_1^n) \geq -\log P(\phi_n(X_1^n)) - c(n), \text{ eventually, a.s. .} \tag{3.25}$$

Rewriting the first term on the RHS,

$$-\log P(\phi_n(X_1^n)) = -\log \frac{P(X_1^n, \phi_n(X_1^n))}{P(X_1^n|\phi_n(X_1^n))} \tag{3.26}$$

$$= -\log P(X_1^n) + \log P(X_1^n|\phi_n(X_1^n)). \tag{3.27}$$

Theorem 3.5.1. *Let ϕ_n be a series of codes operating at fixed distortion level $D_n \leq D, \forall n$ for some balanced distortion measure d. Then*

$$l_n(X_1^n) \geq -\log P(X_1^n) - n(\nu - R_l(D)) - O(\log n) \text{ ev. a.s.}$$

Proof. Combining (3.25) and equation (3.27), we have

$$l_n(X_1^n) \geq -\log P(X_1^n) + \log P(X_1^n|\phi_n(X_1^n)) - O(\log n), \text{ ev. a.s.} \tag{3.28}$$

Define $S(y_1^n) = \{x_1^n : d_n(x_1^n; y_1^n) \leq D\}$. For balanced distortion measures, $|S|_n := |S(y_1^n)|$ does not depend on y_1^n. We will argue that:

Lemma 3.5.2.

$$\log |S|_n \geq -\log P(X_1^n|\phi_n(X_1^n)) - O(\log n) \text{ eventually, a.s. .}$$

Proof.

$$P(-\log P(X_1^n|\phi_n(X_1^n)) \geq \log |S|_n + c_n) \tag{3.29}$$

$$= P(\frac{1}{|S|_n P(X_1^n|\phi_n(X_1^n))} \geq 2^{c_n}) \tag{3.30}$$

$$\leq 2^{-c_n} E\left[\frac{1}{|S|_n P(X_1^n|\phi_n(X_1^n))}\right]. \tag{3.31}$$

For any pair of random variables X_1^n, Y_1^n with $X_1^n \in S(Y_1^n)$ we have

$$\begin{aligned} E\left[\frac{1}{P(X_1^n|Y_1^n)}\right] &= \sum_{y_1^n} \sum_{x_1^n \in S(y_1^n)} \frac{P(x_1^n, y_1^n)}{P(x_1^n|y_1^n)} \\ &= \sum_{y_1^n} P(y_1^n) \sum_{x_1^n \in S(y_1^n)} \frac{P(x_1^n|y_1^n)}{P(x_1^n|y_1^n)} \\ &\leq \sum_{y_1^n} P(y_1^n)|S(y_1^n)| = |S|_n, \end{aligned}$$

where the inequality is due to the fact that only those x_1^n with $P(x_1^n|y_1^n) > 0$ contribute to the inner sum.

We conclude that:

$$P(-\log P(X_1^n|\phi_n(X_1^n)) \geq \log|S|_n + c_n) \leq 2^{-c_n}.$$

Applying the Borel-Cantelli lemma with e.g. $c_n = 2\log n$, we get the desired result. □

Define $R_n^*(D) = \nu + \frac{1}{n}\log\frac{1}{|S|_n}$. Lemma 3.5.2 combined with equation (3.28) gives:

$$l_n(X_1^n) \geq -\log P(X_1^n) + n(R_n^*(D) - \nu) - O(\log n), \text{ ev. a.s.} \tag{3.32}$$

Lastly, we prove:

Lemma 3.5.3.

$$n|R_n^*(D) - R_l(D)| = O(\log n).$$

Proof. Since d is balanced, notice that $d_n(x_1^n; y_1^n)$ only depends on the 'type' (the type of a string is a vector of counts of the appearances of each symbol in the string) of $\Phi_{y_1^n}(x_1^n)$. (Recall that Φ is the permutation with the property $d(x;y) = d(\Phi_y(x), 0)$.) By well known arguments resulting from the combinatorics of types (see e.g. chapter 2 of (14)), we know

$$(n+1)^{-|\mathbb{K}|}2^{nH(X)} \leq |S|_n \leq (n+1)^{|\mathbb{K}|}2^{nH(X)},$$

where X has the distribution that maximizes $H(X)$ subject to $Ed(X,0) \leq D$. Taking logarithms, we get

$$|\log |S|_n - \max_{Ed(X,0)\leq D} nH(X)| = O(\log n).$$

The result follows by the definitions of $R_n^*(D)$ and $R_l(D)$. □

Combining lemma 3.5.3 with eq. 3.32 we conclude the proof of the theorem. □

3.5.1 Proof of theorem 3.3.6

Having proved theorem 3.5.1, it only remains to show that:

Lemma 3.5.4.

$$-\log P(X_1^n) \geq -\log P(X_1^n|X_{-\infty}^0) - O(\log n), \ \textit{ev. a.s.} \ .$$

Proof. We argue as in (1) that

$$E\left[\frac{P(X_1^n)}{P(X_1^n|X_{-\infty}^0)}\right] \leq 1, \tag{3.33}$$

and thus

$$P(-\log P(X_1^n|X_{-\infty}^0) \geq -\log P(X_1^n) + c_n) \tag{3.34}$$

$$= P(\frac{P(X_1^n)}{P(X_1^n|X_{-\infty}^0)}) \geq 2^{c_n}) \tag{3.35}$$

$$\leq 2^{-c_n} E\left[\frac{P(X_1^n)}{P(X_1^n|X_{-\infty}^0)}\right] \leq 2^{-c_n}. \tag{3.36}$$

Picking $c_n = 2\log n$ and invoking the Borel-Cantelli lemma completes the proof. □

3.6 Mixing sources

Define the function $\rho_n = -\log P(X_n|X_{-\infty}^{n-1})$. Theorem 3.3.6 can be re-written as

$$l_n(X_1^n) \geq \sum_{i=1}^{n}(\rho_i - \nu) + nR_l(D) - O(\log n) \quad \text{ev. a.s.} \ .$$

This allows us to bound the limiting behavior of the code length sequence by applying well known limit theorems to the stationary sequence ρ_n. For instance, when (X_n) are i.i.d., it follows that (ρ_n) is also an i.i.d. sequence. It can easily be shown that the variance of $\rho_n = -\log P(X_n|X_{-\infty}^{n-1})$ is bounded:

Lemma 3.6.1. $E[\rho_1^2] < \infty$.

Proof.

$$\begin{aligned} E[\rho_1^2] &= \lim_{N\to\infty} \sum_{x_{-N}^1} P(x_{-N}^0)P(x_1|x_{-N}^0)\log^2 P(x_1|x_{-N}^0) \\ &\leq \lim_{N\to\infty} 2\sum_{x_{-N}^1} P(x_{-N}^0) = 2K, \end{aligned}$$

since $P\log^2 P$ terms are bounded above by 2. □

Therefore (ρ_n) satisfies a central limit theorem with limiting variance $var[\rho_0]$. It follows that for memoryless, finite state sources (X_n):

Corollary 3.6.2. *There exists a sequence of random variables (z_n) s.t.*

$$\frac{l_n(X_1^n) - nR_l(D)}{\sqrt{n}} \geq z_n$$

with $z_n \xrightarrow{d} N(0, var[\rho_0])$.

When (X_n) are not i.i.d, but sufficiently fast mixing, one would expect that the same holds for the sequence (ρ_n). In general, suppose that the sequence

$$\sigma^2 = var(\rho_0) + 2\sum_{i=1}^{\infty} cov(\rho_0, \rho_i) \tag{3.37}$$

converges. Sufficient conditions for this to hold have been studied in (33). The convergence holds, for instance, when (X_n) is a finite state, finite order Markov source, or more generally when (X_n) has the following mixing properties (50):

$$\alpha(k) = O(k^{-336}) \text{ and } \gamma(k) = O(k^{-48})$$

with

$$\alpha(k) := \sup\{|P(B \cap A) - P(B)P(A)|; \\ A \in \mathcal{F}(X_{-\infty}^0), B \in \mathcal{F}(X_k^\infty)\},$$

$$\gamma(k) := \max_{x \in \mathbb{K}} E|\log P(X_1 = x|X_{-\infty}^0) - \log P(X_1 = x|X_{-k}^0)|.$$

An easy corollary to theorem 3.3.6 for the above cases is the following one sided central limit theorem.

Corollary 3.6.3. *There exists a sequence of random variables* (z_n) *s.t.*

$$\frac{l_n(X_1^n) - nR_l(D)}{\sqrt{n}} \geq z_n$$

with $z_n \xrightarrow{d} N(0, \sigma^2)$.

We refer to $\liminf \frac{1}{n} E[(l_n(X_1^n) - nR_l(D))^2]$ as the coding variance. Then σ^2 is a lower bound on the minimum coding variance. In the memoryless case, this can easily be calculated as $var(-\log P(X_0))$. In general, for sources that meet the Shannon lower bound, and for which the sum in (3.37) is absolutely summable, we conclude that the minimum coding variance is strictly positive unless ρ_n is equal to a deterministic constant. This confirms the conjecture raised in (35) in a more general setting.

What is perhaps more interesting is that minimum coding variance for lossy coding that meets the Shannon lower bound admits a lower bound that is independent of the distortion level D and is equal to the minimum lossless coding variance. [1] This is surprising, because it implies that the minimum coding variance can be discontinuous at distortion level $D_{max} := \inf_d\{R(d) = 0\}$. Consider an information source for which the Shannon lower bound holds with equality for the entire range of distortions $0 \leq D \leq D_{max}$. The i.i.d. $X_n \sim$ Bernoulli(p) process with Hamming distortion measure is one such source. It is easy to show that:

[1] For the lossless case see (45),(46).

Lemma 3.6.4. *For $D = D_{max} + \epsilon$, there exists an achievable coding scheme with*

$$l_n(X_1^n) \leq 1 \text{ eventually a.s. .}$$

Proof. Without loss of generality, let $p \leq \frac{1}{2}$. Note that $D_{max} = p$. We code as follows. If $\sum_{i=1}^{n} X_i < n(p+\epsilon)$, we map to all zeros. This is within distortion $D_{max} + \epsilon$. Otherwise, we transmit the exact string X_1^n. We use a 1 bit flag to indicate which event happens. Since we know

$$P(\sum_{i=1}^{n} X_i \geq n(p+\epsilon)) \leq e^{-O(n)},$$

the error event stops happening eventually almost surely by Borel-Cantelli, thus proving the lemma. □

This shows that the minimum coding variance is 0 when $D = D_{max} + \epsilon$ for any ϵ, while it is strictly non-zero when $D = D_{max}$.

3.7 Long range dependent sources

The results in the previous section imply that for sufficiently fast mixing information sources, the optimal pointwise redundancy in the code length process is bounded below by an order $\sqrt{n}$ random process. In this section, we investigate the case when the memory in the source decays much more slowly.

Assume that the entropy density (ρ_n) is LRD with Hurst index $\frac{1}{2} \leq H \leq 1$. From theorem 3.3.6, we conclude that the process $l_n(X_1^n) - R_l(D)$ is lower bounded by the partial sums of a zero-mean LRD process with Hurst index H and therefore the pointwise redundancy in code length is lower bounded by a process that is at least of order n^H. The result is true for any coding algorithm with fixed distortion that has average code length equal to the Shannon lower bound. In other words, long

range dependence is an information-theoretic quantity that persists under any coding scheme. This result was first suggested in (45) in the context of lossless coding of an LRD renewal process. The extension to the lossy case is important, because in practice, long range dependence is observed in the context of coding with distortion (e.g. at the output of a variable bit-rate video coder (5; 22; 52; 21)).

Therefore efforts to mitigate long range dependence using clever coding might be futile, at least in the constant distortion case. To maintain a less bursty rate, one might try to use codes with variable quality, in which case we conjecture that the distortion function will likely end up being long range dependent.

This entire discussion hinges on the fact that there exists information sources for which the entropy density (ρ_n) is LRD, and for which the Shannon lower bound is tight. Below we construct an example process with these properties, demonstrating that the above discussion is not vacuous.

3.7.1 Example

The first example of a concrete information source which has (ρ_n) LRD was presented in (45). There it is proved that if (X_n) is a stationary discrete time LRD renewal process with Hurst index H, then $\rho_n = -\log P(X_n|X_{-\infty}^{n-1})$ is also LRD with identical Hurst index H.

Here we demonstrate an information source such that (ρ_n) is LRD with Hurst index H for which the Shannon lower bound is tight for some strictly non-zero distortion $D > 0$.

Let $X_1(n) \in \{0, 1\}$ be an LRD renewal process with Hurst index H. Let $X_2(n) \in \{0, 1\}$ be an i.i.d. Bernoulli(p) process. Let X_1 be independent of X_2. Define

$$X_n = (X_1(n), X_2(n)) \in \{0, 1\}^2,$$

with $d(x;y) = 1 - \delta(x = y)$ for $x, y \in \{0,1\}^2$.

Note that we are able to write

$$X_n = (X_1(n), X_2(n)) = (X_1(n), 0) \oplus (0, X_2(n))$$

for the appropriate addition operation defined on $\{0,1\}^2$. Since d is a difference distortion measure, and the source can be decomposed into a group sum of i.i.d. components, by theorem 3.4.3, we conclude that the Shannon lower bound holds for this source for a strictly non-zero distortion level D.

By construction, we also have:

$$\begin{aligned}
\rho_n &= -\log P(X_n | X_{-\infty}^{n-1}) \\
&= -\log P(X_1(n) | (X_1)_{-\infty}^{n-1}) P(X_2(n)) \\
&= -\log P(X_1(n) | (X_1)_{-\infty}^{n-1}) - \log P(X_2(n)) \\
&:= \rho_1(n) + \rho_2(n),
\end{aligned}$$

which is LRD with Hurst index H by virtue of (ρ_1) having this property.

3.8 Achievability for LRD sources

For achievability we will use the well known random code construction. Let Q_n be a distribution on $\hat{\mathbb{K}}^n$. An infinite random codebook $\mathbb{C}$ drawn i.i.d. from Q_n is known both to the transmitter and the receiver. Let $W_n(x_1^n)$ be the index of the first $y_1^n \in \mathbb{C}$ such that $d(x_1^n; y_1^n) \leq D$ ($W_n(x_1^n) = \infty$ if no such match is found). The index $W_n(x_1^n)$ is transmitted using Elias coding of the integers whenever $W_n(x_1^n) < \infty$. X_1^n is transmitted as it is using $n\lceil \log K \rceil$ bits otherwise. A 1 bit flag is used to indicate which has occurred. The performance of this scheme is known to obey the following ((35), theorem 8):

Theorem 3.8.1.

$$l_n(X_1^n) \leq -\log Q_n(B(X_1^n, D)) + 2\log\log \frac{2n^2}{Q_n(B(X_1^n, D))} + O(\log n) \ \textit{ev. a.s.}$$

where $B(X_1^n, D) = \{y_1^n \in \hat{\mathbb{K}}^n : d(x_1^n; y_1^n) \leq D\}$.

We can tweak this by sending the exact representation of (X_1^n) whenever $-\log Q_n(B(X_1^n, D)) > n\lceil \log K \rceil$, which ensures that

Theorem 3.8.2.

$$l_n(X_1^n) \leq -\log Q_n(B(X_1^n, D)) + O(\log n) \ \textit{ev. a.s.}$$

where $B(X_1^n, D) = \{y_1^n \in \hat{\mathbb{K}}^n : d(x_1^n; y_1^n) \leq D\}$.

We will prove that $-\log Q_n(B(X_1^n, D))$ is well approximated by $-\log P(X_1^n) - n(\nu - R_l(D))$ given the right choice of output distribution Q_n.

Theorem 3.8.3. *Let* $(X_n) \in \mathbb{K}$ *be a stationary sequence for which* $R_l(D') = R(D')$ *for some balanced distortion measure d, for a range of distortions* $D - \epsilon < D' \leq D$ *for some* $\epsilon > 0$. *Then there exists a sequence of codes* $\phi_n(X_1^n)$ *operating at fixed distortion level* $D_n \leq D$ *s.t.*

$$l_n(X_1^n) \leq -\log P(X_1^n) - n(\nu - R_l(D)) - O(\sqrt{n}\log n) \ \textit{ev. a.s.}$$

We remark that the theorem is true for any source that meets the Shannon lower bound at distortion level D, however, the error term $O(\sqrt{n}\log n)$ is too loose to be meaningful in the fast mixing case, where the fluctuations of $-\log P(X_1^n)$ are of order $O(\sqrt{n})$.

Also the condition that the SLB holds for all average distortions in a small range $D - \epsilon \leq Ed(x; y) \leq D$ is usually implied by the slightly weaker condition $R_l(D) = R(D)$. This claim is discussed in the appendix.

Proof. By theorem 3.4.4 the source admits the decomposition

$$X_n = \Phi_{Y_n}(Z_n)$$

where Z_n are i.i.d and independent from Y_n with $H(Z_n) = \max_{X:Ed(X,0)\leq D} H(X)$ and Φ_y, $y \in \hat{\mathbb{K}}$, is the permutation function with $d(x;y) = d(\Phi_y(x), 0), \quad \forall x \in \mathbb{K}$, for a balanced distortion d.

In this construction, we pick Q_n to be equal to the distribution of Y_1^n, corresponding to an expected distortion of $D_- := D - \frac{\log n}{\sqrt{n}}$. We assume n sufficiently large such that $\frac{\log n}{\sqrt{n}} < \epsilon$, and so that such a decomposition exists. The distribution maximizing $H(X)$ subject to $Ed(X,0) \leq D_-$ has the form

$$Z_n \sim \frac{e^{sd(z,0)}}{K} \tag{3.38}$$

with $K = \sum_{x\in\mathbb{K}} e^{sd(x,0)}$, where $s < 0$ is chosen such that $E[d(Z_n, 0)] = D_-$. This can be seen by introducing the Lagrange multiplier s for the condition $Ed(X,0) \leq D_-$ and then solving the unconstrained optimization. Let $\delta_n := D - D_-$. We note that

$$K^n P(x_1^n) = \sum_{\hat{y}_1^n} e^{sd(x_1^n;\hat{y}_1^n)} Q_n(\hat{y}_1^n) \tag{3.39}$$

and

$$P(y_1^n|x_1^n) = \frac{e^{sd(x_1^n;y_1^n)} Q_n(y_1^n)}{\sum_{\hat{y}_1^n} e^{sd(x_1^n;\hat{y}_1^n)} Q_n(\hat{y}_1^n)}. \tag{3.40}$$

$$Q_n(B(x_1^n, D)) = \sum_{B(x_1^n, D)} Q_n(y_1^n) \tag{3.41}$$

$$= \sum_{B(x_1^n, D)} e^{-sd(x_1^n; y_1^n)} e^{sd(x_1^n; y_1^n)} Q_n(y_1^n) \tag{3.42}$$

$$\geq \sum_{B(x_1^n, D) - B(x_1^n, D-2\delta_n)} \tag{3.43}$$

$$\sum_{\hat{y}_1^n} e^{sd(x_1^n; \hat{y}_1^n)} Q_n(\hat{y}_1^n) \tag{3.44}$$

$$e^{-sd(x_1^n; y_1^n)} \frac{e^{sd(x_1^n; y_1^n)} Q_n(y_1^n)}{\sum_{\hat{y}_1^n} e^{sd(x_1^n; \hat{y}_1^n)} Q_n(\hat{y}_1^n)} \tag{3.45}$$

$$\geq K^n P(x_1^n) e^{-nsD + 2sn\delta_n} \tag{3.46}$$

$$\sum_{B(x_1^n, D) - B(x_1^n, D-2\delta_n)} \frac{e^{sd(x_1^n; y_1^n)} Q_n(y_1^n)}{\sum_{\hat{y}_1^n} e^{sd(x_1^n; \hat{y}_1^n)} Q_n(\hat{y}_1^n)} \tag{3.47}$$

$$= K^n P(x_1^n) e^{-nsD + 2sn\delta_n} \sum_{B(x_1^n, D) - B(x_1^n, D-2\delta_n)} P(y_1^n | x_1^n). \tag{3.48}$$

We can re-write the last term as:

$$\sum_{y_1^n \in B(x_1^n, D) - B(x_1^n, D-2\delta)} P(y_1^n | x_1^n) = \tag{3.49}$$

$$\sum_{z_1^n : D_- - \delta_n \leq \frac{1}{n} d(z_1^n, 0) \leq D_- + \delta_n} P(Z_1^n = z_1^n | x_1^n) \tag{3.50}$$

$$:= P(B_\delta | x_1^n) \tag{3.51}$$

where $B_\delta := \{z_1^n : D_- - \delta_n \leq \frac{1}{n} d(z_1^n, 0) \leq D_- + \delta_n\}$.

We show

Lemma 3.8.4.

$$P(Z_1^n \in B_\delta | x_1^n) \geq \frac{1}{2} \ \textit{ev. a.s.}$$

Proof.

$$P(P(Z_1^n \notin B_\delta | x_1^n) \geq \frac{1}{2}) = \tag{3.52}$$

$$P(P(Z_1^n \notin B_\delta | x_1^n) P(x_1^n) \geq \frac{1}{2} P(x_1^n)) \tag{3.53}$$

$$= \sum_{x_1^n : P(x_1^n) \leq 2P(Z_1^n \notin B_\delta, x_1^n)} P(x_1^n) \tag{3.54}$$

$$\leq \sum_{x_1^n} 2P(Z_1^n \notin B_\delta, x_1^n) \leq 2P(Z_1^n \notin B_\delta) \tag{3.55}$$

Since $d(Z_1^n, 0)$ is a sum of i.i.d. bounded variables with mean D_-, we can bound $P(Z_1^n \notin B_\delta)$ by a moderate deviations argument. From (16) (1.2) we have for S_n, the sum of i.i.d. variables,

$$\limsup_n \frac{n}{b_n^2} \log P(|\frac{S_n}{b_n}| > 1) \leq -C.$$

Picking $S_n = d(Z_1^n, 0) - nD$, $b_n = \sqrt{n} \log n$ gives

$$P(Z_1^n \notin B_\delta) \leq e^{-C \log^2 n}$$

for some constant $C > 0$. Since the sequence on the RHS is summable, we deduce by Borell-Cantelli lemma that

$$1 - P(Z_1^n \notin B_\delta | x_1^n) = P(Z_1^n \in B_\delta | x_1^n) \geq \frac{1}{2} \text{ ev. a.s. .} \tag{3.56}$$

□

Combining 3.48 with this lemma and noting that $\log K - sD = \nu - R_l(D)$ (3.20) gives

$$Q_n(B(x_1^n, D)) \geq \frac{1}{2} K^n P(x_1^n) e^{-nsD + 2ns\delta} \text{ ev. a.s. ,} \tag{3.57}$$

$$-\log Q_n(B(x_1^n, D)) \leq -\log P(x_1^n) - n(\nu - R_l(D)) \tag{3.58}$$

$$+ O(\sqrt{n} \log n) \text{ ev. a.s. .} \tag{3.59}$$

□

Appendix

Let D_c equal the supremum of distortion values such that the Shannon lower bound is tight. i.e. $D_c := \sup\{d : R_l(d) = R(d)\}$. Then, assuming balanced distortion measures, according to 3.4.4, the source admits the decomposition

$$X_n = \Phi_{Y_n}(Z_n) \tag{3.60}$$

where Z_n are i.i.d and independent from Y_n with $H(Z_n) = \max_{X:Ed(X,0)\leq D_c} H(X)$. One would expect that a similar decomposition exists for all $0 \leq D \leq D_c$. Indeed, this is shown to be true for finite state Markov sources under balanced distortion in (26),(27). Here we will argue that this behavior is quite generally true.

Let $\mathbf{x_n}$ be the vector in $\mathbb{R}^{\mathbb{K}^n}$ consisting of the probabilities $P(x_1^n)$ for each $x_1^n \in \mathbb{K}^n$. Assume $\hat{\mathbb{K}} = \mathbb{K}$ and similarly define $\mathbf{y_n} \in \mathbb{K}^n$. Take a balanced distortion function $d(\cdot;\cdot)$ with $d(x;x) = 0$ and $d(x;y) > 0$ whenever $x \neq y$. Define matrix $\mathbf{Z_n} \in \mathbb{R}^{\mathbb{K}^n \mathrm{x} \mathbb{K}^n}$ as having entries $\{e^{sd(x_1^n;y_1^n)}\}$. Now (3.60) can be written as

$$\mathbf{x_n} = \frac{1}{K^n}\mathbf{Z_n}\mathbf{y_n}.$$

In other words, the Shannon lower bound is tight whenever the vector $\mathbf{Z}_n^{-1}\mathbf{x}_n$ has non-negative entries.

Note that $\mathbf{Z_n} = \mathbf{Z}^{\otimes n}$ where $\mathbf{Z} \in \mathbb{R}^{\mathbb{K}}$ consists of the entries $\{e^{sd(x;y)}\}$. Thus $\mathbf{Z}_n^{-1} = (\mathbf{Z}^{-1})^{\otimes n}$. Further note that for any distortion value $0 \leq \tilde{D} \leq D_c$, the transition matrix $\tilde{\mathbf{Z}}_n$ will have entries $\{e^{\tilde{s}d(x_1^n;y_1^n)}\}$ where $\tilde{s} \leq s < 0$. i.e. we can write $\tilde{\mathbf{Z}}_n = \mathbf{Z}_n^{\circ r}$ where $r \geq 1$ and $\circ$ denotes element-wise exponentiation.

Theorem 3.8.5. *Let D_s correspond to the value of distortion parametrized by s. If $R_l(D_s) = R(D_s)$ and $(\mathbf{Z}^{\circ \mathbf{r}})^{-1}\mathbf{Z} \geq 0$, then $R_l(D_{sr}) = R(D_{sr})$*

Proof. Let $\mathbf{y_n} = \mathbf{Z}_n^{-1}\mathbf{x}_n \geq 0$ by the assumption $R_l(D_s) = R(D_s)$. Then $(\mathbf{Z}_n^{\circ r})^{-1}\mathbf{x}_n = (\mathbf{Z}_n^{\circ r})^{-1}\mathbf{Z_n}\mathbf{y_n} \geq 0$ since we assume $(\mathbf{Z}^{\circ \mathbf{r}})^{-1}\mathbf{Z} \geq 0$ and $\mathbf{y_n} \geq 0$. □

The condition $(\mathbf{Z}^{\circ \mathbf{r}})^{-1}\mathbf{Z} \geq 0$ holds for all $r \geq 1$ quite generally. For example

Corollary 3.8.6. *(Binary alphabet) Let $|\mathbb{K}| = 2$. Then the Shannon lower bound is tight for all $0 \leq D \leq D_c$.*

Proof. Let

$$\mathbb{Z} = \begin{pmatrix} 1 & a \\ b & 1 \end{pmatrix}.$$

Then

$$\mathbf{Z}^{\circ \mathbf{r}-1} = \frac{1}{\Delta} \begin{pmatrix} 1 & -a^r \\ -b^r & 1 \end{pmatrix}.$$

Since $a, b < 1$ and $r \geq 1$, one can easily verify that $\mathbf{Z}^{\circ \mathbf{r}-1}\mathbf{Z} \geq 0$. □

Corollary 3.8.7. *(Probability of error distortion) Let $d(x; y) = \delta(x \neq y)$. Then the Shannon lower bound is tight for all $0 \leq D \leq D_c$.*

Proof. We can write $\mathbf{Z}^{\circ \mathbf{r}}$ as $(1 - e^{sr})\mathbf{I} + e^{sr}\mathbf{1}\mathbf{1}^{\mathbf{T}}$. This is a Bose-Mesner Matrix, the inverse of which can be calculated as

$$(\mathbf{Z}^{\circ \mathbf{r}})^{-1} = \frac{(1 + (|\mathbb{K}| - 1)e^{sr}\mathbf{I} - e^{sr}\mathbf{1}\mathbf{1}^{\mathbf{T}}}{1 + e^{sr}(|\mathbb{K}| - 2) - e^{2sr}(|\mathbb{K}| - 1)}.$$

We get

$$(1 + e^{sr}(|\mathbb{K}| - 2) - e^{2sr}(|\mathbb{K}| - 1))(\mathbf{Z}^{\circ \mathbf{r}})^{-1}\mathbf{Z} = \tag{3.61}$$

$$(1 - e^s)(1 + (|\mathbb{K}| - 1)e^{sr})\mathbf{I} \tag{3.62}$$

$$+ \mathbf{1}\mathbf{1}^{\mathbf{T}}(e^s(1 + (|\mathbb{K}| - 1)e^{sr}) \tag{3.63}$$

$$- e^{s(r+1)} - e^{sr}(1 - e^s)) \tag{3.64}$$

It can be verified that $1 + e^{sr}(|\mathbb{K}| - 2) - e^{2sr}(|\mathbb{K}| - 1) \geq 0$ and $e^s(1 + (|\mathbb{K}| - 1)e^{sr}) - e^{s(r+1)} - e^{sr}(1 - e^s) \geq 0$ for all $r \geq 1$. □

In general we have

Corollary 3.8.8. *Let $D_c > \delta > 0$ exist. Then there exists $\epsilon > 0$ such that $R_l(D) = R(D)$ for all $0 \leq D \leq \epsilon$.*

Proof. Since $d(x;x) = 0$ and $d(x;y) > 0$ whenever $x \neq y$, $\lim_{r\to\infty}(\mathbf{Z}^{\circ r})^{-1} = \mathbf{I}$. Since $\mathbf{Z} > 0$, there exists $s_c > -\infty$ such that $\mathbf{Z}^{\circ \mathbf{r}-1}\mathbf{Z} \geq 0$ for all $s \geq s_c$. Since D_s is monotonically decreasing in s, we have the result. □

Chapter 4

Concluding remarks

Long range dependence shows up surprisingly often in real world data. Here, we tried to understand how long range dependence arises, in what way long range dependent processes are transformed as they pass through natural and engineered systems, and whether it is possible to suppress this property in systems when we don't want it.

We adopted a versatile model for long range dependence based on countable state Markov chains. We have provided conditions under which the growth rate of the variance of a function of a Markov chain is identical to that of the chain itself. The theorem implies that many instantaneous functions of such chains share the same Hurst index. Our results are widely applicable, however there is considerable art in using them.

In finance, our results imply that while market forces mold prices into roughly a martingale process, long range dependence still persists in the higher order statistics of price returns. In queuing networks, we saw that if long range dependent traffic enters a system, then no choice of routing/scheduling algorithms will alleviate this

problem. In fact, long range dependence might spread to other nodes in the network through coupling at shared service points if enough care is not taken.

We made similar observations for variable-length source codes that operate on information sources which exhibit long memory. As expected, the fluctuations of the rate function of a coder is lower bounded by the fluctuations of the information source. The results generalize to lossy source coding, in the case where the codec is forced to operate at fixed distortion under a balanced distortion measure. The results collectively suggest that long range dependence is fundamental in an information-theoretic sense, persisting under all feasible coding algorithms. Therefore efforts to mitigate long range dependence using clever coding might be futile, at least in the constant distortion case. To maintain a less variable bit-rate, one might try to use codes with variable quality, in which case we conjecture that the distortion function will likely end up being long range dependent.

An overall conclusion from the results of this thesis is that long range dependence is difficult both to create and to destroy. It acts largely as an invariant under fairly general classes of transformations. The source of long range dependence should probably be sought in human actions (as in the case of finance), or complicated natural systems (such as weather patterns) rather than the simpler systems which process and manipulate raw data created by such sources.

Bibliography

[1] P. Algoet and T. Cover, "A sandwich proof of the Shannon-McMillan-Breiman theorem," *The Annals of Probability*, pp. 899–909, 1988.

[2] V. Anantharam, "Scheduling strategies and long-range dependence," *Queueing systems*, vol. 33, no. 1, pp. 73–89, 1999.

[3] A. Barron, "Logically smooth density estimation." Ph.D. dissertation, Stanford University, Department of Electrical Engineering, 1985.

[4] J. Beran, *Statistics for long-memory processes*. Chapman & Hall/CRC, 1994, vol. 61.

[5] J. Beran, R. Sherman, M. Taqqu, and W. Willinger, "Long-range dependence in variable-bit-rate video traffic," *IEEE Transactions on Communications*, vol. 43, no. 234, pp. 1566–1579, 1995.

[6] T. Berger, *Rate-Distortion Theory*. Wiley Online Library, 1971.

[7] N. Bingham, C. Goldie, and J. Teugels, *Regular Variation*. Cambridge University Press, 1989.

[8] T. Bollerslev, "A conditionally heteroskedastic time series model for speculative prices and rates of return," *The review of economics and statistics*, pp. 542–547, 1987.

[9] K. Carpio and D. Daley, "Long-range dependence of Markov chains in discrete time on countable state space," *Journal of Applied Probability*, vol. 44, no. 4, pp. 1047–1055, 2007.

[10] K. Chung, *Markov chains with Stationary Transition Probabilities*. Berlin: Springer-Verlag, 1967.

[11] R. Cont, "Empirical properties of asset returns: stylized facts and statistical issues," *Quantitative Finance*, vol. 1, no. 2, pp. 223–236, 2001.

[12] T. Cover, J. Thomas, J. Wiley, *et al.*, *Elements of information theory*. Wiley Online Library, 1991.

[13] M. Crovella and A. Bestavros, "Self-similarity in world wide web traffic: evidence and possible causes," *Networking, IEEE/ACM Transactions on*, vol. 5, no. 6, pp. 835–846, 1997.

[14] I. Csiszár and J. Körner, *Information theory: Coding theorems for discrete memoryless systems*. Academic Press (New York and Budapest), 1981.

[15] D. Daley, "The Hurst index of long-range dependent renewal processes," *Annals of Probability*, pp. 2035–2041, 1999.

[16] A. De Acosta, "Moderate deviations and associated laplace approximations for sums of independent random vectors," *Trans. Amer. Math. Soc*, vol. 329, no. 1, pp. 357–375, 1992.

[17] P. Doukhan, G. Oppenheim, and M. Taqqu, *Theory and applications of long-range dependence*. Birkhauser, 2003.

[18] R. Durrett, *Probability: theory and examples*. Brooks/Cole Advanced Books & Software, 2005.

[19] P. Elias, "Universal codeword sets and representations of the integers," *Information Theory, IEEE Transactions on*, vol. 21, no. 2, pp. 194–203, 1975.

[20] A. Es-Saghouani and M. Mandjes, "On the correlation structure of a lévy-driven queue," *Journal of Applied Probability*, vol. 45, no. 4, pp. 940–952, 2008.

[21] F. Fitzek and M. Reisslein, "Mpeg4 and h. 263 video traces for network performance," *IEEE Network*, vol. 15, no. 6, pp. 40–54, 2001.

[22] M. Garrett and W. Willinger, "Analysis, modeling and generation of self-similar VBR video traffic," *ACM SIGCOMM Computer Communication Review*, vol. 24, no. 4, pp. 269–280, 1994.

[23] G. Giller. (2010, May) Giller investments, llc. [Online]. Available: http://blog.gillerinvestments.com

[24] L. Giraitis, P. Robinson, and D. Surgailis, "A model for long memory conditional heteroscedasticity," *Annals of Applied Probability*, vol. 10, no. 3, pp. 1002–1024, 2000.

[25] C. Granger, "The typical spectral shape of an economic variable," *Econometrica: Journal of the Econometric Society*, pp. 150–161, 1966.

[26] R. Gray, "Information rates of autoregressive processes," *Information Theory, IEEE Transactions on*, vol. 16, no. 4, pp. 412–421, 1970.

[27] ——, "Rate distortion functions for finite-state finite-alphabet markov sources," *Information Theory, IEEE Transactions on*, vol. 17, no. 2, pp. 127–134, 1971.

[28] D. Heath, S. Resnick, and G. Samorodnitsky, "Patterns of buffer overflow in a class of queues with long memory in the input stream," *The Annals of Applied Probability*, vol. 7, no. 4, pp. 1021–1057, 1997.

[29] C. Heyde and Y. Yang, "On defining long-range dependence," *Journal of Applied Probability*, pp. 939–944, 1997.

[30] H. Hurst, "Methods of using long-term storage in reservoirs." in *ICE Proceedings*, vol. 5, no. 5. Ice Virtual Library, 1956, pp. 519–543.

[31] S. Jalali and T. Weissman, "New bounds on the rate-distortion function of a binary markov source," in *Information Theory, 2007. ISIT 2007. IEEE International Symposium on.* IEEE, 2007, pp. 571–575.

[32] J. Kieffer, "Sample converses in source coding theory," *Information Theory, IEEE Transactions on*, vol. 37, no. 2, pp. 263–268, 1991.

[33] I. Kontoyiannis, "Second-order noiseless source coding theorems," *Information Theory, IEEE Transactions on*, vol. 43, no. 4, pp. 1339–1341, 1997.

[34] ——, "Asymptotic recurrence and waiting times for stationary processes," *Journal of Theoretical Probability*, vol. 11, no. 3, pp. 795–811, 1998.

[35] ——, "Pointwise redundancy in lossy data compression and universal lossy data compression," *Information Theory, IEEE Transactions on*, vol. 46, no. 1, pp. 136–152, 2000.

[36] W. Leland, M. Taqqu, W. Willinger, and D. Wilson, "On the self-similar nature of ethernet traffic (extended version)," *Networking, IEEE/ACM Transactions on*, vol. 2, no. 1, pp. 1–15, 1994.

[37] N. Likhanov and R. Mazumdar, "Cell loss asymptotics for buffers fed with a large number of independent stationary sources," *Journal of Applied Probability*, vol. 36, no. 1, pp. 86–96, 1999.

[38] N. Likhanov, B. Tsybakov, and N. Georganas, "Analysis of an atm buffer with self-similar (fractal) input traffic," in *INFOCOM'95. Fourteenth Annual Joint*

Conference of the IEEE Computer and Communications Societies. Bringing Information to People. Proceedings. IEEE. IEEE, 1995, pp. 985–992.

[39] T. Linder and R. Zamir, "On the asymptotic tightness of the shannon lower bound," *Information Theory, IEEE Transactions on*, vol. 40, no. 6, pp. 2026–2031, 1994.

[40] B. Mandelbrot, "Forecasts of future prices, unbiased markets, and martingale models," *Journal of Business*, vol. 39, no. 1, pp. 242–255, 1966.

[41] ——, "When can price be arbitraged efficiently? a limit to the validity of the random walk and martingale models," *The Review of Economics and Statistics*, vol. 53, no. 3, pp. 225–236, 1971.

[42] ——, *Fractals and scaling in finance: discontinuity, concentration, risk: selecta volume E.* Springer Verlag, 1997.

[43] M. Markakis, E. Modiano, and J. Tsitsiklis, "Scheduling policies for single-hop networks with heavy-tailed traffic," *Proceedings of the 47th annual Allerton conference on Communication, control, and computing*, pp. 112–120, 2009.

[44] M. Mureşan, *A concrete approach to classical analysis.* Springer Verlag, 2009, vol. 28.

[45] B. Oğuz and V. Anantharam, "Compressing a long range dependent renewal process," in *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on.* IEEE, 2010, pp. 1443–1447.

[46] ——, "Hurst index of functions of long range dependent Markov chains," *Journal of Applied Probability*, vol. 49, no. 2, 2012.

[47] ——, "Pointwise lossy source coding theorem for sources with memory," in *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*. IEEE, 2012.

[48] K. Park, G. Kim, and M. Crovella, "On the relationship between file sizes, transport protocols, and self-similar network traffic," in *Network Protocols, 1996. Proceedings., 1996 International Conference on*. IEEE, 1996, pp. 171–180.

[49] K. Park and W. Willinger, "Self-similar network traffic: An overview," *Self-Similar Network Traffic and Performance Evaluation*, pp. 1–38, 2000.

[50] W. Philipp and W. Stout, *Almost sure invariance principles for partial sums of weakly dependent random variables*. Amer. Mathematical Society, 1975.

[51] S. Resnick and G. Samorodnitsky, "Activity periods of an infinite server queue and performance of certain heavy tailed fluid queues," *Queueing Systems*, vol. 33, no. 1, pp. 43–71, 1999.

[52] O. Rose, "Statistical properties of MPEG video traffic and their impact on traffic modeling in ATM systems," *Conference on Local Computer Networks: Proceedings*, p. 397, 1995.

[53] Z. Sahinoglu and S. Tekinay, "On multimedia networks: self-similar traffic and network performance," *Communications Magazine, IEEE*, vol. 37, no. 1, pp. 48–52, 1999.

[54] G. Samorodnitsky and M. Taqqu, *Stable non-Gaussian processes: Stochastic models with infinite variance*. Chapman & Hall, 1994.

[55] R. Schassberger, "Insensitivity of steady-state distributions of generalized semi-markov processes. part i," *The Annals of Probability*, pp. 87–99, 1977.

[56] C. E. Shannon, "A mathematical theory of communication," *The Bell Systems Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.

[57] ——, "Coding theorems for a discrete source with a fidelity criterion," *IRE Nat. Conv. Rec*, vol. 4, no. 142-163, 1959.

[58] A. Wyner and J. Ziv, "Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression," *Information Theory, IEEE Transactions on*, vol. 35, no. 6, pp. 1250–1258, 1989.

[59] ——, "Fixed data base version of the lempel-ziv data compression algorithm," *Information Theory, IEEE Transactions on*, vol. 37, no. 3, pp. 878–880, 1991.

[60] E. Yang and Z. Zhang, "On the redundancy of lossy source coding with abstract alphabets," *Information Theory, IEEE Transactions on*, vol. 45, no. 4, pp. 1092–1110, 1999.

[61] Z. Zhang, E. Yang, and V. Wei, "The redundancy of source coding with a fidelity criterion. 1. known statistics," *Information Theory, IEEE Transactions on*, vol. 43, no. 1, pp. 71–91, 1997.

[62] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *Information Theory, IEEE Transactions on*, vol. 24, no. 5, pp. 530–536, 1978.

[63] B. Zwart, S. Borst, and M. Mandjes, "Exact queueing asymptotics for multiple heavy-tailed on-off flows," in *INFOCOM 2001. Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 1. IEEE, 2001, pp. 279–288.