Universal Compression of Piecewise iid Sources

Owen Cameron

October 2017

A thesis submitted in partial fulfilment of Honours in Mathematics of the Australian National University
Declaration

The work in this thesis is my own except where otherwise stated.

Owen Cameron
Acknowledgements

To my main supervisor, Marcus, I don’t know how to thank you for what you’ve provided me in supervision this year. For giving me a curious subject to focus on, I am grateful. Throughout the year I learnt more from you than I could have expected. To Farhana, I owe my gratitude for helping me understand the problem. Badri, and Pierre, I actually don’t known what to say. Your role this year has been indispensable; without both of you I doubt I’d have made it this far.

To the guys in the Brian Anderson dungeon, it’s been a hoot. To Sina, for showing me the value in looking out for others. When all seemed afar, Boris, I found a reliable mate. To Tom, who inspires me to think deeply. To Darren, and Ari for bringing people together. My thanks to Suraj, John and Jarryd, for being truly inspiring. To Elliot and Amy, who taught me about commitment. And finally many thanks to Wei for being happy and fun - you’ve been a breath of fresh air through these last months.

To the maths kids who I pushed through these last few weeks with, I wish I’d seen more of you. Particularly An Ran and James, with the calming presence of Adwait. I’ve been so glad to find that others have found this year a struggle, and it’s been a privilege to grow with you these last few weeks.

To my housemateys - all four of you - thank you for your patience and accommodation, your provision of impromptu counselling, bed-time serenading and being happy people that I could look forward to coming home to (however occasionally) and seeing.

To my other friends over the years who I’ve been able to journey through uni with. To Brad, Jarryd, Kate and Ian, my first year besties; Robbie, Dan and Layan, my most consistent pals; Eloise and Joseph, my examples of perfect balance; Ben and Sim, for reminding me that it’s good to monkey about with your brothers; Brad, Hugo, Dave and John, for keeping me moving; to the Cowper crew, Alex, Em and Asto for being an emotional refuge. And to the many other
wonderful humans I have the blessing of calling my friends. Let’s meet soon.

And of course, to my family in Tasmania, my stable rock. You know what you mean to me. I made it.
Abstract

We consider the problem of compressing for discrete random processes taking values a finite set, whose statistical parameters are not known, but assumed to come from some class. For instance, the process might be iid or form a Markov chain.

We consider the problem of estimating the probability of random sequences whose statistical parameters $\theta$ are unknown but assumed to come from some parametric class $\{P_{\theta} \mid \theta \in \Theta\}$ (for example, iid sequences or markov chains). More precisely, suppose $X_1, \ldots, X_n$ are random variables taking values in a finite set $\mathcal{A}$, with joint distribution $P_{\theta}$ for some (unknown) $\theta \in \Theta$. We seek a universal joint distribution $Q : \mathcal{A}^n \rightarrow [0,1]$ which is ‘not much worse’ than any potential ‘true’ distribution $P_{\theta} \in \{P_{\theta} \mid \theta \in \Theta\}$, where performance is measured by coding redundancy $\log \frac{P_{\theta}(x)}{Q(x)}$, $x \in \mathcal{A}^n$. The study of coding redundancy is motivated by probabilistic data compression, as it represents the difference in compression using distribution $Q$ instead of $P_{\theta}$. The problem of compressing optimally (and efficiently) with respect to a given distribution is solved (Huffman coding, arithmetic coding [6, 26]), so to compress a source with unknown parameter $\theta$ the problem remains to choose $Q$ appropriately.

We consider classes of “piecewise stationary” sources in which the parameters of the source occasionally change with $i$. [Include motivation??] For example, the sequence of random variables $X_1, \ldots, X_n$ might be iid. with parameter $\theta^{(1)}$ for $X_1, \ldots, X_t$ and with parameter $\theta^{(2)}$ for $X_{t+1}, \ldots, X_n$. In the piecewise setting, we prove a bound on the redundancy of one prediction method $Q$ which is already known to perform optimally for finite-memory stationary sources. This bound guarantees that our method is also universal over piecewise iid. sources, and therefore extends the class for which $Q$ is universal.
# Contents

Acknowledgements v

Abstract vii

1 Introduction 1

2 Source Coding 5
  2.1 Instantaneous codes 6
  2.2 Kraft inequality 8
  2.3 Shannon Entropy 9

3 Markov Sources 13
  3.1 iid processes 13
  3.2 Markov processes 16

4 Universal Compression 19
  4.1 Compressing with respect to the wrong distribution 19
  4.2 Universal compression 20
  4.3 Universal prediction for iid processes 22
    4.3.1 Bound for KT distribution 24
  4.4 Universal prediction for Markov processes 25
    4.4.1 Bound for $k$-KT distribution 26

5 Coding with $k(\cdot)$-KT for Piecewise iid Processes 27
  5.1 Piecewise iid Processes 27
  5.2 Basic proof 28
  5.3 Performance of $k(\cdot)$-KT for general piecewise environments 31
    .1 Probability and Independence 39

Bibliography 40
Chapter 1

Introduction

In his famous paper, “A Theory of Mathematical Communication” [28], Claude Shannon laid the foundations of information theory. The theory formalizes the problem of lossless\(^1\) compression of a random data source (random process). The purpose of compressing data is to minimize the cost of transmitting it across some communication channel, or storing it on a hard-drive, for instance. In Shannon’s formulation, we think of data/message as not given, but as an instance of some underlying random process. The problem is to engineer a system of communication between a ‘transmitter’ and ‘receiver’ so that messages can be communicated efficiently. If something is known about which messages are most probable, there may be scope to reduce the expected volume of data transmitted (or stored). The key elements of this problem are a countable set \(\mathcal{X}\) of possible ‘messages’ or values the data may assume, and an associated discrete probability\(^2\) \(P : \mathcal{X} \rightarrow [0, 1]\) for each possible value. Assuming transmission (storage) is across a binary channel (binary hard-drive), we want to assign a unique binary string or code \(c(x) \in \{0, 1\}^*\) to each \(x \in \mathcal{X}\) to transmit. To minimize cost, we want to choose \(c\) so that the length of the string/code is short, on average.

Roughly, this thesis is motivated by the problem of compressing optimally when \(\mathcal{X}\) is a set of long strings and the probability \(P\) is unknown.

We start by introducing the relevant parts of Shannon’s theory for lossless compression. The most important result is Shannon’s source coding theorem, which establishes an achievable lower limit on the compressibility of a source (ie.

\(^1\)lossless compression is where the original data can be perfectly recovered from the compressed form. To contrast, in many multimedia applications compression can be achieved by disposing of superfluous detail in the data.

\(^2\)if there was no randomness or uncertainty associated with data, it would be useless to collect it or communicate it.
establishes the expected code-length of the optimal code $c$ for $P$). This limit is called the Shannon entropy, $H(P)$, and is closely related (and named after) entropy in thermodynamics. The entropy indicates that to optimally compress with respect to $P$ one should assign a binary code $c(x)$ whose length is roughly $-\log_2 P(x)$ bits (for each $x \in \mathcal{X}$). This can be done with, for instance, the algorithms given in [28, 14, 26]. Thus, the problem of compressing optimally with respect to a given distribution $P$ is solved.

An interesting class of problems arises when the ‘true’ distribution $P$ is not known in advance. If we instead code $X$ optimally with respect to another distribution $Q$ (which is intended to approximate $P$) the difference in their optimal code-lengths is roughly $\log \frac{P(x)}{Q(x)}$, known as the redundancy. One natural approach to this problem is to assume $P$ comes from some class $\{P_\theta \mid \theta \in \Theta\}$, and choose an arbitrary distribution $Q$ which is, in some sense, ‘close’ to all $P_\theta$. This begs the question of what is meant by ‘close’ and how one should choose $Q$. For instance, one might choose $Q$ in order to minimize the worst-case redundancy over all $x \in \mathcal{X}$ and all $\theta$ in $\Theta$,

$$\sup_{x \in \mathcal{X}, \theta \in \Theta} \log \frac{P_\theta(x)}{Q(x)}$$

or to minimize the worst-case expected redundancy with respect to the unknown parameter

$$\sup_{\theta \in \Theta} \mathbb{E}_{P_\theta} \log \frac{P_\theta(x)}{Q(x)} = \sup_{\theta \in \Theta} D(P_\theta \parallel Q)$$

We consider the special case where $\mathcal{X}$ is a set of sequences (over some finite alphabet $\mathcal{A}$). This case is practically natural to study, since data in the real world almost exclusively takes the form of a sequence (eg. written text, DNA, binary storage on a hard-drive). Moreover, these sequences tend to be very long, often in the order of several megabytes or gigabytes.

Compression of such long sequences calls for the study of distributions over them. The models we consider are time-homogeneous Markov chains, and as a special case, iid processes. A stochastic process $(X_i)_{i \in \mathbb{N}}$ with each $X_i$ taking values in $\mathcal{A}$ defines a joint distribution $P^n$ on the product space $\mathcal{A}^n$ for each $n \in \mathbb{N}$.

The above problem of choosing $Q$ for long sequences with respect to some class $\{P_\theta \mid \theta \in \Theta\}$ (of stochastic processes) is the domain of universal compression. Similarly, we want to choose a process $Q$ (which defines a distribution $Q_n$ over $\mathcal{A}^n$) such that for all $\theta \in \Theta$ the redundancy $\log \frac{P_{\theta,n}}{Q_n}$ is small. More precisely,
the criterion of universal prediction is to choose a process $Q$ (which defines a distribution $Q_n$ over $\mathcal{A}^n$) with vanishing per-symbol redundancy as $n \to \infty$

$$
\lim_{n \to \infty} \frac{1}{n} \log \frac{P_n(x_1, \ldots, x_n)}{Q_n(x_1, \ldots, x_n)} = 0 \tag{1.3}
$$
either in expectation, over all sequences $(x_i)_{i \in \mathbb{N}}$ (ie. pointwise), or in some other topology. For compression, if the entropy of $P_n$ grows linearly (which includes most interesting stationary ergodic processes) the coding redundancy effectively vanishes when compared to the total length of encoding (as $n \to \infty$).

The original contribution of this thesis is a result in universal compression for piecewise iid sources. Informally, a sequence of random variables $X_1, \ldots, X_n$ is piecewise iid if it is the concatenation of a fixed number of iid sequences. For example, $X_1, \ldots, X_n$ is piecewise iid with 2 pieces if it can be written

$$
(X_1, \ldots, X_n) = (Y_1^{(1)}, \ldots, Y_{n_1}^{(1)}, Y_1^{(2)}, \ldots, Y_{n_2}^{(2)}) \tag{1.4}
$$
where $(Y_i^{(1)})_{i=1}^{n_1}$ and $(Y_i^{(2)})_{i=1}^{n_2}$ are each iid.

We give a guarantee that one method of prediction, which is a universal predictor for finite-memory Markov sources, is almost surely universal for piecewise iid categorical sources. We conclude with a discussion of some of the related literature.
Chapter 2

Source Coding

This section introduces the source coding setup. Of particular importance is Shannon’s entropy which gives a theoretic limit for the compressibility of a known source.

The theory of source coding is concerned with the problem of designing protocols to parsimoniously encode messages from a random source. More concretely, suppose that every second some random process produces a message or some ‘data’ from a set of possible messages. For example, a message might be a letter from the English alphabet. This has a natural probability distribution given by the frequency of symbol occurrences (for instance, the letter e is more frequent than q). A ‘sender’ and ‘receiver’ want to agree on a system so the messages can be sent across some communication channel. The canonical example of a channel is a binary electric cable, through which a continuous sequence of 0 and 1’s can be communicated. Thus, they must ‘encode’ the messages as binary sequences to transmit.

Naturally, communicating over a channel is costly, so they would like their agreed protocol to minimize how much they use the channel. We assume that both sender and receiver have a-priori knowledge of the source distribution. The study of lossless compression addresses the question of how cheaply they can do this - of how few bits they can get away with (on average) to communicate each message. The rough idea is that more probable messages should be communicated with shorter codes, since they will be transmitted more frequently.

There are several important considerations, however. To avoid confusion, different messages must not be assigned the same code. More subtly, however, the receiver must also have some way of knowing when the code for each symbol ends upon receiving it (‘decoding’). This brings us to the study of instantaneous
codes and the Kraft inequality.

**Definition 2.1** (Some computer science parlance). An alphabet $\mathcal{A}$ is merely a finite set. Its elements $\sigma$ may be referred to as symbols. A string $s$ from $\mathcal{A}$ is a finite sequence of symbols, $(\sigma_1, \sigma_2, \ldots, \sigma_n) \equiv \sigma_1 \sigma_2 \cdots \sigma_n$ (we will use both notations interchangeably). The set of strings from $\mathcal{A}$ is denoted $\mathcal{A}^*$. From any two strings $s = \sigma_1 \cdots \sigma_m$ and $t = \tau_1 \cdots \tau_n$ we may form another by the binary operation of concatenation,

$$s \cdot t = \sigma_1 \cdots \sigma_m \cdot \tau_1 \cdots \tau_n := \sigma_1 \cdots \sigma_m \tau_1 \cdots \tau_n \quad (2.1)$$

$s$ is a prefix of $t$ if there exists a string $r$ such that $s \cdot r = t$. Conversely, $r$ is a suffix of $t$ if there exists $s \in \mathcal{A}^*$ with $t = s \cdot r$. Note that every string is a prefix and suffix of itself (ie. there exists an empty string of length 0). We reserve $l(s)$ to denote the length of $s$. Where the indexing is well defined, we may denote a contiguous subsequence/substring $\sigma_j \sigma_{j+1} \cdots \sigma_k$ of $s$ by $s_{j:k}$. (Alternatively, an arbitrary string $s$ of length $n$ may sometimes be introduced as $x_{1:n}$ to specify that $x$ is a string of length $n$ with indexing $x_{1:n}$.

**Definition 2.2.** Let $\mathcal{X}$ be a countable set. A binary code for $\mathcal{X}$ is a map

$$c : \mathcal{X} \rightarrow \{0, 1\}^* \quad (2.2)$$

c is called prefix-free if $x \neq y$ implies $c(x)$ is not a prefix of $c(y)$.

More general ‘$m$-ary’ codes $c : \mathcal{X} \rightarrow \{1, \ldots, m\}^*$ can also be studied, relevant to when the storage or transmission medium has more than just two states, such as the 26-state English alphabet, or 3-colour fibre-optic cable. The concepts defined below have an analogous $m$-ary version, but up to a normalisation factor $(\log m)$ they are essentially equivalent. Therefore, in what follows we restrict to the base 2 standard.

## 2.1 Instantaneous codes

Observe that if $c$ is prefix free then it must be injective. Therefore, the code that is transmitted, $c(x)$, will not be confused with any other code $c(y)$. From another angle, the injectivity guarantees the existence of a ‘decompression’ function $c^{-1} : c(\mathcal{X}) \rightarrow \mathcal{X}$ for the receiver.

Note that if $c$ is prefix-free then the number of symbols needed to distinguish $c(x)$ from all other codes is no more than $l(c(x))$. For this reason, prefix-free codes are sometimes called instantaneous codes.
Figure 2.1: Source coding for transmission
2.2 Kraft inequality

The Kraft inequality characterizes the possible code-lengths for instantaneous codes.

**Proposition 2.3** (Kraft inequality). If \( c : \mathcal{X} \to \{0,1\}^* \) is an instantaneous code then

\[
\sum_{x \in \mathcal{X}} 2^{-l(c(x))} \leq 1 \quad (2.3)
\]

Conversely, if \( \mathcal{X} \) is countable and \( L : \mathcal{X} \to \mathbb{N} \) satisfies \( \sum_{x \in \mathcal{X}} 2^{-L(x)} \leq 1 \) then there exists an instantaneous code \( c : \mathcal{X} \to \{0,1\}^* \) with \( L(x) = l(c(x)) \quad \forall x \in \mathcal{X} \).

**Proof.** For each \( s = s_1 \cdots s_l(s) \in \{0,1\}^* \), define

\[
I(s) = [i(s), i(s) + 2^{-l(s)}] \quad (2.4)
\]

where \( i(s) = \sum_{i=1}^{k} 2^{-i} \cdot s_i \). This is a bijection between \( \{0,1\}^* \) and the set of dyadic intervals, \( \{[\frac{i}{2^m}, \frac{i+1}{2^m}) \mid m \in \mathbb{N}, i \in \{0, \ldots, 2^m - 1\}\} \). Further, for each \( s \) the length of \( I(s) \) is \( 2^{-l(s)} \).

Note that \( I(s_1) \cap I(s_2) \) is non-empty exactly when either \( s_1 \) is a prefix of \( s_2 \), or vice versa. Thus, if \( c \) is prefix-free and \( x, y \in \mathcal{X} \) then \( I(c(x)) \) and \( I(c(y)) \) are disjoint. Then \( \{I(s) \mid s \in c(\mathcal{X})\} \) is a set of pairwise disjoint intervals, each contained within \([0,1]\). Thus, their total length satisfies

\[
\sum_{x \in \mathcal{X}} 2^{-l(c(x))} = \sum_{x \in \mathcal{X}} |I(c(x))| \leq 1 \quad (2.5)
\]

To prove the converse, assume for simplicity of notation that \( |\mathcal{X}| \) is finite. We can enumerate \( \mathcal{X} = \{x_1, x_2, \ldots, x_{|\mathcal{X}|}\} \) such that \( L(x_1) \leq \ldots \leq L(x_{|\mathcal{X}|}) \). Then we can assign a dyadic interval to each \( x_i \) of appropriate length as follows: Let \( S(i) = \sum_{j<i} 2^{-L(x_j)} \) and define the interval \( I_i = [S_i, S_i + 2^{-L(x_i)}] \). By the hypothesis, this is indeed an assignment of disjoint dyadic intervals to each \( x \in \mathcal{X} \). Then define a code for \( x_i \) according to the correspondence of (2.5) with \( I_i \). \( \square \)

The Kraft inequality therefore indicates exactly which codelengths are possible for instantaneous codes. We now turn to an important concept in information theory known as Shannon entropy (or just entropy). Entropy is an important quantity in the study of instantaneous codes. In particular, it represents the optimal expected length of an instantaneous code for a random source.
Remark 2.4. If \((\mathcal{X}, \mathcal{P}(\mathcal{X}), P)\) is a probability space and \(c : \mathcal{X} \rightarrow \{0, 1\}^*\) a code for \(\mathcal{X}\) then \(c\) can be viewed as a random variable \((\mathcal{X}, \mathcal{P}(\mathcal{X})) \rightarrow (\{0, 1\}^*, \mathcal{P}(\{0, 1\}^*))\). Furthermore, the length function \(l : \{0, 1\}^* \rightarrow \mathbb{R}\) is Borel measurable, and therefore \(l(c(\cdot))\) can be viewed as a random variable \((\mathcal{X}, \mathcal{P}(\mathcal{X})) \rightarrow (\mathbb{R}, \mathcal{B})\).

2.3 Shannon Entropy

Definition 2.5. Let \((\mathcal{X}, \mathcal{P}(\mathcal{X}), P)\) be a discrete probability space. Note that \(P\) can be viewed as a random variable \((\mathcal{X}, \mathcal{P}(\mathcal{X})) \rightarrow (\mathbb{R}, \mathcal{B})\). We define the entropy of \(P\) as

\[
H(P) = \mathbb{E}_P[-\log_2 P]
\]

\[
= \sum_{x \in \mathcal{X}} -P(x) \log_2 P(x)
\]

Also, if \(X\) is a discrete source, we informally write \(H(X)\) to denote the entropy of its distribution.

Remark 2.6. Take note that the definition of entropy depends only on the distribution \(P\), and nothing about the discrete space \((\mathcal{X}, \mathcal{P}(\mathcal{X}))\) on which it is defined. Hence, it would perhaps be more natural to define \(H\) directly on an appropriate space of (equivalence classes of) distributions over finite sets.

Indeed, let \((\mathcal{A}, \mathcal{P}(\mathcal{A}))\) denote a finite discrete space, and \(\Delta_\mathcal{A}\) denote the class of probability distributions on \(\mathcal{A}\). Then we have

\[
H : \Delta_\mathcal{A} \rightarrow \mathbb{R}
\]

It is easily verified that \(H\) is nonnegative, continuous and strictly concave. \(H\) attains its maximum exactly on the uniform distribution \(P_u(\sigma) = \frac{1}{|\mathcal{A}|} \forall \sigma \in \mathcal{A}\), at which point, \(H(P_u) = \log_2 |\mathcal{A}|\). \(H\) is minimized on the Dirac masses, \(P = \delta_\sigma\) for some \(\sigma \in \mathcal{A}\).

Henceforth all logarithms are assumed to be base 2.

The entropy of a source gives a lower bound on the expected length of an instantaneous code. Conversely, given a source \((X, P)\) there is a code whose expected code length is at most 1 bit more than the entropy. As a result entropy serves as the benchmark for optimal source coding for a known distribution. These results are captured in the following two lemmas respectively:
Lemma 2.7. If \((\mathcal{X}, P)\) is a discrete probability space and \(c : \mathcal{X} \rightarrow \{0, 1\}^*\) an instantaneous code then

\[
H(P) \leq \mathbb{E}[l(c)]
\]

(2.9)

Proof. 

\[
\mathbb{E}[l(c(x))] = \sum_{x \in \mathcal{X}} P(x) \cdot l(c(x))
\]

(2.10)

\[
= H(x) - H(x) - \sum_{x \in \mathcal{X}} P(x) \cdot \log_2 2^{-l(c(x))}
\]

(2.11)

\[
= H(x) - \sum_{x \in \mathcal{X}} P(x) \cdot \log \left( \frac{2^{-l(c(x))}}{P(x)} \right)
\]

(2.12)

\[
\geq H(x) - \log \left( \sum_{x \in \mathcal{X}} P(x) \cdot 2^{-l(c(x))} \right)
\]

(2.13)

\[
= H(x) - \log \left( \sum_{x \in \mathcal{X}} 2^{-l(c(x))} \right)
\]

(2.14)

\[
\geq H(x)
\]

(2.15)

where (2.13) follows from Jensen’s inequality \((- \log \text{ convex})\) and (2.15) follows from the Kraft inequality.

Lemma 2.8. Let \((\mathcal{X}, P)\) be a discrete probability space. There exists an instantaneous code \(c : \mathcal{X} \rightarrow \{0, 1\}^*\) with

\[
\mathbb{E}[l(c)] \leq H(P) + 1
\]

(2.16)

Proof. Define

\[
L : \mathcal{X} \rightarrow \mathbb{N}
\]

(2.17)

\[
x \mapsto \left\lceil \log \frac{1}{P(x)} \right\rceil
\]

(2.18)

\(L\) satisfies the Kraft inequality -

\[
\sum_{x \in \mathcal{X}} 2^{-L(x)} \leq \sum_{x \in \mathcal{X}} 2^{-\left\lceil \log \frac{1}{P(x)} \right\rceil}
\]

(2.19)

\[
\leq \sum_{x \in \mathcal{X}} 2^{-\log \frac{1}{P(x)}}
\]

(2.20)

\[
= \sum_{x \in \mathcal{X}} P(x)
\]

(2.21)

\[
= 1
\]

(2.22)
2.3. **SHANNON ENTROPY**

Therefore there exists an instantaneous code \( c \) with \( L(x) = l(c(x)) \). The expected length of this code is

\[
\mathbb{E}[l(c)] = \mathbb{E}\left[\left\lceil \log \frac{1}{P} \right\rceil\right] \leq H(P) + 1
\]

(2.23)

\[
< \mathbb{E}\left[ \log \frac{1}{P} \right] + 1 \leq H(P) + 1
\]

(2.24)

\[
= H(P) + 1
\]

(2.25)

Thus, the entropy represents the optimal benchmark for instantaneous source coding: For a given source \((X, P)\), the expected code length of the optimal code \( c \) satisfies

\[
H(P) \leq \mathbb{E}(l(c)) \leq H(P) + 1
\]

(2.26)

If not for the restriction that code-lengths must be integers (but rather, they only have to satisfy the Kraft inequality) then the left inequality would be an equality. These may be thought of as integer code rounding errors. Indeed, quite often we will say a code \( c \) is optimal if its expected length satisfies (2.26).

From a practical perspective, there are many known algorithms that compress near optimally. Perhaps the most theoretically noteworthy is Huffman coding [14], an algorithm proven to be truly optimal for finite distributions. However, several others are well known, and can have practical advantages over Huffman coding.

The existence of these (near) ‘optimal’ coding algorithms means the problem of designing a code for a given distribution is essentially a solved problem. If one is given a distribution to code with respect to, we can just use one of these algorithms. However, when the distribution is not given, interesting questions arise. Instead a class of distributions might be given, for instance, and the problem is to try and find a suitable distribution \( Q \) to code with which satisfies some optimality condition (see for instance Definition 4.3).
Chapter 3

Markov Sources

Often the types of message we want to compress are themselves strings from some finite alphabet \( \mathcal{A} \). A file of written text (e.g. English), for example. This section introduces Markov chains, to which corresponds a widely studied and highly useful class of distributions over sequences.

In the remainder of this essay, we assume that \( \mathcal{A} \) always denotes a finite set/alphabet with an implicit order, \( \mathcal{A} = \{\sigma_1, \ldots, \sigma_{|\mathcal{A}|}\} \). Where a sigma algebra for \( \mathcal{A} \) is contextually implied, the discrete sigma algebra \( \mathcal{P}(\mathcal{A}) \) will be assumed implicitly (likewise for \( \mathcal{A}^n \) or any finite space).

**Definition 3.1.** Let \((\Omega, \mathcal{F}, \mathbb{P})\) denote a probability space. A sequence of random variables \((X_i : \Omega \rightarrow \mathcal{A})_{i \in \mathbb{N}}\) taking values in the same space \((\mathcal{A}, \mathcal{P}(\mathcal{A}))\) is called a discrete time stochastic process on \( \mathcal{A} \) (henceforth stochastic process or just process).

### 3.1 iid processes

**Example 3.2.** Suppose \((\Omega, \mathcal{F}, \mathbb{P})\) admits a sequence \((X_i)_{i \in \mathbb{N}}\) of independent\(^1\) and identically distributed (iid.) random variables with values in \( \mathcal{A} \). Thus, if \((x_1, \ldots, x_n) \in \mathcal{A}^n\) we have (using standard probabilistic informalities)

\[
\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i) \tag{3.1}
\]

\(^1\)see appendix
Since all $X_i$ have identical distribution on $\mathcal{A}$, the distribution of the whole process (on $\mathcal{A}^\infty$ with the standard sigma algebra) is induced by a single distribution over the (discrete) space $\mathcal{A}$.

Suppose we want to encode the result of the joint variable $(X_1, \ldots, X_n) : \Omega \to \mathcal{A}^n$ with a prefix-free code. That is, to give a code $c : \mathcal{A}^n \to \{0, 1\}^*$. By Lemma (2.7), the expected length of the code cannot be less than $H(X_1, \ldots, X_n)$. To reiterate, $H(X_1, \ldots, X_n)$ means the entropy of the distribution that $(X_1, \ldots, X_n)$ induces on $\mathcal{A}^n$. Thus,

$$H(X_1, \ldots, X_n) = \sum_{(x_1, \ldots, x_n) \in \mathcal{A}^n} -P(X_1 = x_1, \ldots, X_n = x_n) \log P(X_1 = x_1, \ldots, X_n = x_n)$$

(3.2)

$$= -\sum_{x_1, n \in \mathcal{A}^n} \prod_{i=1}^{n} P(X_i = x_i) \log \prod_{i=1}^{n} P(X_i = x_i)$$

(3.3)

$$= -\sum_{i=1}^{n} \sum_{x_i \in \mathcal{A}} P(X_i = x_i) \log P(X_i = x_i)$$

(3.4)

$$= \sum_{i=1}^{n} H(X_i)$$

(3.5)

$$= n \cdot H(X_1)$$

(3.6)

The proof of Lemma (2.8) indicates that to code optimally for $(X_1, \ldots, X_n)$ according to the condition (2.26) we should choose $c(x_1 \cdots x_n)$ with code-length

$$l(c(x_1 \cdots x_n)) \approx -\log P(X_1 = x_1, \ldots, X_n = x_n)$$

(3.7)

for each $(x_1, \ldots, x_n) \in \mathcal{A}^n$, and this will achieve an expected length less than $n \cdot H(X_1) + 1$. If $X_i$ is not constant (a.s.) then according to Remark 2.6 we have $H(X_1) > 0$. Thus, the entropy of $H(X_1, \ldots, X_n)$ increases linearly in $n$.

For a specific example, take $\mathcal{A} = \{0, 1\}$. Then $(X_i)_{i \in \mathbb{N}}$ is a sequence of iid Bernoulli variables, whose distribution can be characterized by some $\theta \in [0, 1]$, such that for all $i$

$$P(X_i = 1) = \theta$$

(3.8)

Then by the definition of Shannon entropy,

$$H(X_i) = \theta \log \frac{1}{\theta} + (1 - \theta) \log \frac{1}{1 - \theta}$$

(3.9)

**Definition 3.3.** For any finite set $\mathcal{X}$, let $\Delta_\mathcal{X}$ denote the class of (discrete) probability measures $P : \mathcal{X} \to [0, 1]$ on $\mathcal{X}$.
Remark 3.4. Given an order on $X$, $X = \{\sigma_1, \ldots, \sigma_{|X|}\}$, we can view $\Delta_X$ as a compact subset of $[0, 1]^{|X|}$ with the identification

$$\Delta_X \rightarrow \mathbb{R}^{|X|}$$

$$P \mapsto (P(\sigma_1), \ldots, P(\sigma_{|X|}))$$

In particular, we will often consider the case when $X = A^n$ for some $n \in \mathbb{N}$.

To any $A$-valued stochastic process, $(X)$ corresponds a distribution on $A^n$, namely the one induced by the joint variable $(X_1, \ldots, X_n)$, such as the one in (3.1) for the iid process. Let this be denoted $P^*_n(X) \in \Delta_{An}$. Furthermore, the induced distributions are compatible in the following way: if $m < n$ then for any $(\sigma_1, \ldots, \sigma_m) \in A^m$ we have

$$P^*_m(X)(\sigma_1, \ldots, \sigma_m) = \sum_{(\sigma_{m+1}, \ldots, \sigma_n) \in A^{n-m}} P^*_n(X)(\sigma_1, \ldots, \sigma_n)$$

Conversely, if $(X)$ and $(Y)$ are $A$-valued stochastic processes with $P^*_m(X)(\cdot) = P^*_m(Y)$ then $(X)$ and $(Y)$ are identically distributed (on $A^\infty$ with the canonical product sigma algebra).

Thus, $P(\cdot)$ gives a projection of $A$-valued stochastic processes onto their the distribution-equivalence classes. These are the objects of interest in our setup.

TAKE CARE: In this finite setting it is tempting to say that two distributions $P_1, P_2 \in \Delta_A$ are equivalent if there is a permutation $\phi : A \rightarrow A$ such that $P_1 = P_2 \circ \phi$. It might be even more tempting to say two processes are equivalent in these terms, and indeed this is a well defined equivalence relation. However, this is not the definition of equivalence ‘up to distribution’. The order of elements in $A$ matters.

Notation 3.5. Where it is understood from the context (e.g. length of argument) we will drop the subscript $n$ in $P^*_n$. For brevity, we will sometimes write $X_{1:n}$ to denote $(X_1, \ldots, X_n)$. Recall the analagous definition of $x_{1:n}$ given previously in Notation ??.

Remark 3.6. In the particular case where $(X_i)_{i \in \mathbb{N}}$ is iid, its induced distributions $P^*_n(X)$ can be (uniquely) inferred from $P^1(X)$ (using the properties of identicalness and independence). Thus, under the identification given in (3.10) and canonical order on $A^n$ we have a compact $\mathbb{R}$ parametrisation of the class of $A$-valued iid processes, up to equivalence in distribution. Throughout we will denote this parametrisation by $\Theta_0(A)$. Thus

$$\{P_\theta \mid \theta \in \Theta_0(A)\}$$
denotes the class of iid distributions for random processes.

Next we look at a more general class of processes, namely Markov processes, and a corresponding parametrisation for their distributions.

### 3.2 Markov processes

**Definition 3.7.** Let \((\Omega, \mathcal{F}, \mathbb{P})\) denote a probability space. Suppose \((X_i)_{i \in \mathbb{N}}\) is an \(\mathcal{A}\)-valued stochastic process such that for all \(i \in \{2, 3, 4, \ldots\}\), \((X_1, \ldots, X_{i-1})\) and \(X_{i+1}\) are conditionally independent given \(X_i\). Then \((X_i)\) is called a first order Markov process.

A formal definition of conditional independence is given in an appendix. Note that the condition implies that \(X_{i+1}\) and any subsequence of \((X_1, \ldots, X_{i-1})\) are conditionally independent given \(X_i\). Thus, given \(X_i\), \(X_{i+1}\) depends on nothing that happened before time \(i\).

As a special case, note that the independence of a family of random variables implies the conditional independence of the family when conditioned on any random variable. Thus, independently distributed processes are a Markov process.

Since \(\mathcal{A}\) is finite, \(X_i : \Omega \to \mathcal{A}\), we have the following equivalence:

**Proposition 3.8.** Let \(\mathcal{A}\) be finite. An \(\mathcal{A}\)-valued stochastic process \(\{X_i\}_{i \in \mathbb{N}}\) is a Markov source if and only if \(\forall i \in \mathbb{N}, \forall (x_j)_{j \in \mathbb{N}},\)

\[
\mathbb{P}\left( X_{i+1} = x_{i+1} \mid (X_1, \ldots, X_i) = (x_1, \ldots, x_i) \right) = \mathbb{P}\left( X_{i+1} = x_{i+1} \mid X_i = x_i \right)
\]

(3.14)

The proof is elementary, and not the focus of this thesis so has been omitted.

The following generalizes Definition 3.7.

**Definition 3.9.** Let \(k \in \mathbb{N}\). A stochastic process \((X_i)_{i \in \mathbb{N}}\) is a \(k^{th}\) order Markov source if \(\forall i \in \mathbb{N}, \{X_{i+k+1}\}\) is conditionally independent of \(\{X_1, \ldots, X_i\}\) given \(\{X_{i+1}, \ldots, X_{i+k}\}\).

An analogous version of Proposition (3.8) holds for \(k^{th}\) order processes.

As they are currently defined, the family of distributions \(\left\{P_n^{(X)}\right\}_{n \in \mathbb{N}}\) could be infinitely complex, since the condition of conditional independence imposes no time-consistent regularity on the distribution. Thus, the following can be taken as the appropriate analogue for Markov processes of being ‘identically distributed’ in the iid case.
3.2. MARKOV PROCESSES

Definition 3.10. Let \( \{X_i : \Omega \rightarrow \mathcal{A}\}_{i \in \mathbb{N}} \) be a first order Markov process, \( \mathcal{A} \) finite. The process is called *time homogeneous* if the conditional distribution is invariant under time shifts. That is, for all \( i, j \in \mathbb{N} \) and all \( x, y \in \mathcal{A} \)

\[
P(X_{i+1} = y \mid X_i = x) = P(X_{j+1} = y \mid X_j = x)
\]  

(3.15)

Likewise a \( k^{th} \) order Markov process is *time homogeneous* if for all \( i, j \in \mathbb{N}, y \in \mathcal{A} \) and \( x_{1:k} \in \mathcal{A}^k \)

\[
P(X_{i+k+1} = y \mid X_{i+1:i+k} = x_{1:k}) = P(X_{j+k+1} = y \mid X_{j+1:j+k} = x_{1:k})
\]  

(3.16)

Remark 3.11. Suppose \((X)\) is a time homogeneous \(k\)th order Markov processes. Then its distribution \(P_X\) can be deduced from the ‘starting’ distribution \(P_{\mathcal{X}}^k \in \Delta_{\mathcal{A}^k}\) and the conditional ‘update’ formulas \(P_X(\cdot \mid x) \in \Delta_{\mathcal{A}}\) for each \( x \in \mathcal{A}^k \), where

\[
P_X(y \mid x) := \frac{P_{\mathcal{X}}^{k+1}(xy)}{P_{\mathcal{X}}^k(x)}
\]  

(3.17)

\[
= P(X_{k+1} = y \mid X_{1:k} = x)
\]  

(3.18)

with \(xy\) denoting the concatenation defined in Notation ?? (and the usual assumptions for zero probabilities when defining conditional distributions). The deduction of \(P^n\) is be made using the chain rule for conditional probability, applying the \(k\)th order analogue of (3.8) from conditional independence, and then the time homogeneity.

The appropriate uniqueness condition also holds.

Thus, we have a parametrisation of time homogeneous \(k\)th order Markov processes \(\Delta_{\mathcal{A}^k} \times \Delta_{\mathcal{A}^k}^{\mid \mathcal{A}^k}\) (up to equivalence in distribution).

Notation 3.12. We denote the above parametrisation of distributions for \(k\)-order Markov processes by \(\Theta_k\) (consistent with Notation (??)). Thus

\[
\{P_\theta \mid \theta \in \Theta_k\}
\]  

(3.19)

denotes the class of distributions for time homogenous \(k\)-order Markov processes.

Lemma 3.13. Suppose \((X)\) is a \(k\)th order time homogeneous Markov process with parameter in the interior of \(\Theta_k\). Then the entropy \(H(X_1, \ldots, X_n)\) grows linearly with \(n\).
As a result of this lemma, Shannon’s source coding theory implies that the expected number of bits from an optimal encoding of \((X_{1:n})\) increases linearly.

The purpose of this chapter was to introduce an interesting and popular class of probability models for sequences/strings, with a view towards optimally compressing random strings from these models with known model parameters. The following section approaches the problem of compressing when the model parameters are not known a-priori.
Chapter 4

Universal Compression

This section introduces the problem of universal compression/prediction, and introduces ‘universal’ distributions for the model classes presented in the previous chapter. These universal distributions are the topic of next chapter.

4.1 Compressing with respect to the wrong distribution

From Chapter 2 it may be inferred that to optimally compress a random source, it suffices to know the distribution of the source, \( P \). Then one may use their favourite coding mechanism (Huffman, Arithmetic) and optimally encode a message \( x \) as a binary string \( c(x) \) whose length is roughly

\[
l(c(x)) = - \log P(x)
\]  

If we instead encoded \( x \) optimally with respect to another distribution \( Q \) by assigning a code of length \(- \log Q(X)\) to \( x \), then the difference in code-length would be

\[
- \log Q(x) + \log P(x) = \log \frac{P(x)}{Q(x)}
\]

**Definition 4.1** (Redundancy, Relative entropy/KL divergence). Suppose \( \mathcal{X} \) is a discrete space and \( P, Q \in \Delta_{\mathcal{X}} \) are two distributions. Then we define the (point-wise) redundancy of \( P \) and \( Q \) as

\[
\rho_{P,Q} : \mathcal{X} \to \mathbb{R} \cup \{\pm\infty\} \cup \{\text{undefined}\}
\]

\[
x \mapsto \log \frac{P(x)}{Q(x)}
\]
where we take $\log \frac{p}{0} = \infty$, etc.

$P$ is absolutely continuous with respect to $Q$ if for any $x$ such that $Q(x) = 0$, $P(x)$ is also 0. If we view $P(\cdot)$ and $Q(\cdot)$ as random variables $X \to \mathbb{R}$ and take the expectation with respect to the measure $P$,

$$D(\cdot \| Q) : \Delta^2_X \to \mathbb{R}$$

$$D(P \parallel Q) = E_P[\log \frac{P(\cdot)}{Q(\cdot)}]$$

$$= E_P \rho_{P,Q}$$

$$= \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)}$$

this is called the relative entropy/KL divergence from $P$ to $Q$. Note that if $P$ is not absolutely continuous w.r.t. $Q$ then their KL divergence is $\infty$.

**Lemma 4.2.** KL divergence is nonnegative, concave and vanishes exactly on the diagonal, $\{(P, P) \mid P \in \Delta_X\}$.

The lemma follows from Jensen’s inequality, since $- \log$ is convex.

From (2.26), the redundancy $\rho_{P,Q}(\cdot)$ measures the difference in the optimal code-length with respect to $Q$ and $P$ (ignoring integer-code rounding errors). As yet, no indication has been given that coding $x$ with $- \log P(x)$ is necessary for optimal encoding, only sufficient. However, this follows the fact that the divergence vanishes exactly on the diagonal.

The above gives well defined and justified notions of the cost of coding with respect to the wrong distribution. We now turn to the problem of trying to code well even in the absence of knowledge of the true distribution. If it is assumed that the true distribution comes from some class $\{P_\theta \mid \theta \in \Theta\}$, one might seek a distribution $Q$ for which the above difference is ‘guaranteed’ (in some sense) to be small over this class. The next section introduces a range of notions for ‘guarantees’ which are relevant to the classes of distributions studied in Chapter 3.

### 4.2 Universal compression

Let $\{P_\theta \mid \theta \in \Theta\}$ be a class of distributions for stochastic processes over $\mathcal{A}$. Suppose the distribution of $(X_i)_{i \in \mathbb{N}}$ is distributed in $\{P_\theta \mid \theta \in \Theta\}$, but its parameter

\footnote{From an optimisation point of view, we seek to find a distribution $Q$ (to code with respect to) which minimizes some loss function related to the redundancy (yet to be made precise).}
is not known. Universal compression concerned with the limits of compressing \((X_1, \ldots, X_n)\) as \(n \to \infty\). Lemma 3.13 indicates that most Markov processes have linearly increasing entropy \(H(X_1, \ldots, X_n)\) in \(n\). Therefore, the expected length of the optimal code increases linearly.

**Definition 4.3.** Suppose \(\{P_\theta \mid \theta \in \Theta(A)\}\) is a class of distributions for \(A\)-valued processes (and \((\Omega, F, P)\) admits a process for each distribution). For each \(\theta\) let \(X^{(\theta)}\) denote a process distributed according to \(\theta\). A coding distribution \(Q\) is **universal in expectation (or mean)** over the class \(\{P_\theta \mid \theta \in \Theta(A)\}\) if for all \(\theta \in \Theta(A)\)

\[
\lim_{n \to \infty} \frac{1}{n} \mathbb{E} \rho_{P_\theta, Q}(X^{(\theta)}_1, \ldots, X^{(\theta)}_n) = 0
\]  

(4.9)

The following conditions are in increasing order of strength:

**Q is universal in probability over \(\{P_\theta \mid \theta \in \Theta(A)\}\)** if for every \(\theta \in \Theta(A)\) and every \(\epsilon > 0\)

\[
\limsup_{n \to \infty} P \left( \frac{1}{n} \rho_{P_\theta, Q}(X^{(\theta)}_1, \ldots, X^{(\theta)}_n) < \epsilon \right) = 1
\]  

(4.10)

**Q is universal almost surely if for every \(\theta \in \Theta(A)\)**

\[
P \left( \lim_{n \to \infty} \frac{1}{n} \rho_{P_\theta, Q}(X^{(\theta)}_1, \ldots, X^{(\theta)}_n) = 0 \right) = 1
\]  

(4.11)

**Q is pointwise universal if for every \(\theta \in \Theta(A)\) and any sequence \((x_i)_{i \in \mathbb{N}} \in A^\mathbb{N}\)**

\[
\lim_{n \to \infty} \frac{1}{n} \rho_{P_\theta, Q}(x_1, \ldots, x_n) = 0
\]  

(4.12)

**Q is uniformly pointwise universal if for every \(\theta \in \Theta(A)\)**

\[
\lim_{n \to \infty} \sup_{(x_i)_{i \in \mathbb{N}} \in A^\mathbb{N}} \frac{1}{n} \rho_{P_\theta, Q}(x_1, \ldots, x_n) = 0
\]  

(4.13)

**Q is extremely uniformly pointwise universal if**

\[
\lim_{n \to \infty} \sup_{\theta \in \Theta(A)} \sup_{(x_i)_{i \in \mathbb{N}} \in A^\mathbb{N}} \frac{1}{n} \rho_{P_\theta, Q}(x_1, \ldots, x_n) = 0
\]  

(4.14)

If the entropy \(H(X_1, \ldots, X_n) = H(P^n_{(X)})\) grows linearly with \(n\), and \(Q\) is universal over \(\{P_\theta \mid \theta \in \Theta\}\) then the per-symbol redundancy vanishes as \(n \to \infty\) (in the appropriate topology). This essentially corresponds to \(Q\) being optimal for \(P^n_\theta\) as \(n \to \infty\).

Note that the definitions above only depend on the distribution of \((X^{\theta})\), so they are well defined.
4.3 Universal prediction for iid processes

In this section we present a distribution $Q$ which is extremely uniformly point-wise universal over the class $\Theta_0$ of distributions for iid processes. It is called the KT distribution, named after the authors of its origin [17]. It is the basic element of an extension which is universal over the class $\Theta_k$ of $k$-order time homogeneous Markov processes presented in the next section. Further, and an adaptation of this extension is studied next chapter for an even more general class. Before introducing the KT distribution, we make a definition which will be useful throughout.

**Notation 4.4.** Let $A = \{\sigma_1, \ldots, \sigma_d\}$ and assume $k \leq n$. Recall that for $x \in A^*$ we may denote a contiguous subsequence $(x_j, x_{j+1}, \ldots, x_k)$ by $x_{j:k}$. For each $s \in A^k$ and $x \in A^n$, denote the number of times the sequence $s$ occurs as a (contiguous) subsequence of $x_{1:n}$ by the counting function

$$c_s(x_{1:n}) = \sum_{i=0}^{n-k} 1_s(x_{i+1:i+k})$$

(4.15)

We may refer to $c_s(x_{1:n})$ as the ‘count of $s$ in $x_{1:n}$’.

Later the following will also be useful. Encode the number of times $s$ occurs in $x_{1:n}$ is followed by each symbol $\sigma \in A$ in the vector

$$\tau_s(x_{1:n}) = (c_{s\sigma_1}(x_{1:n}), \ldots, c_{s\sigma_{|A|}}(x_{1:n}))$$

(4.16)

We may refer to $\tau_s(x_{1:n})$ as the ‘counts following $s$ in $x_{1:n}$’.

**Definition 4.5 (Maximum likelihood estimate/estimator).** Let $x_{1:n} \in A^n$ and $\{P_\theta \mid \theta \in \Theta(A)\}$ be a (compact) class of distributions for stochastic processes over $A$ (such that for any $x_{1:n} \in A^n$, $P_\theta(x_{1:n})$ is continuous in $\theta$). The maximum likelihood estimate for $x_{1:n}$ is the parameter $\theta_{ML} \in \Theta$ which maximises the likelihood of $x_{1:n}$. i.e.

$$\theta_{ML}(x_{1:n}) := \arg\max_{\theta \in \Theta(A)} P_\theta(x_{1:n})$$

(4.17)

The maximum likelihood probability is

$$P_{ML}(x_{1:n}) := P_{\theta_{ML}(x_{1:n})}(x_{1:n}) = \max \{P_\theta(x_{1:n}) \mid \theta \in \Theta(A)\}$$

(4.18)

The conditions that $\Theta$ is compact and $P_\theta(x_{1:n})$ is continuous ensure the maximum is attained. Since the definition only depends on the distribution, the
definition is independent of parametrisation. Thus, the definition makes sense. Note that these conditions hold for Markov processes (and therefore iid).

Take special note that \( P_{ML}^n(\cdot) \) is not a probability distribution over \( \mathcal{A}^n \). However, it is a useful function to have as it upper bounds the probability of all distributions in the relevant class.

For iid processes, \( \Theta = \Delta_A \), it is easily verified using elementary calculus that the maximum likelihood estimate for \( x_{1:n} \) is the parameter of symbol frequencies in \( x_{1:n} \). That is,

\[
\theta_{ML}(x_{1:n}) = \left( \frac{c_1(x_{1:n})}{n}, \ldots, \frac{c_d(x_{1:n})}{n} \right) = \left( \frac{c_1}{n}, \ldots, \frac{c_d}{n} \right) \tag{4.19}
\]

Then the ML probability for \( x_{1:n} \) is

\[
P_{ML}^n(x_{1:n}) = \prod_{\sigma \in \mathcal{A}} \frac{c_{\sigma}}{n^n} \tag{4.20}
\]

**Proposition 4.6** (Properties of ML probability). Symmetry under permuting \( x_{1:n} \) and under permuting symbols \( \sigma_1, \ldots, \sigma_n \). Convex as function of \( c_i \)s on the \( n \) simplex \( \Delta_n \).

**Definition 4.7.** Let \( \mathcal{A} \) be a finite set. The KT distribution for \( \mathcal{A} \) is the \( \mathcal{A} \)-valued stochastic process distribution \( \{P_{KT}^n\}_{n \in \mathbb{N}} \) (namely satisfying (3.12)) where for each \( n \in \mathbb{N} \)

\[
P_{KT}^n : \mathcal{A}^n \to [0, 1] \tag{4.21}
\]

\[
P_{KT}^n(x) := \frac{\prod_{\sigma \in \mathcal{A}} \Gamma(c_{\sigma}(x) + \frac{1}{2})}{\Gamma(\frac{1}{2}|\mathcal{A}|) \cdot \Gamma(n + \frac{|\mathcal{A}|}{2})} \tag{4.22}
\]

where \( \Gamma \) is the Gamma function. Notationally, since \( n \) is implied by the length of its argument, it will often be omitted.

**Lemma 4.8.** As defined above,

\[
P_{KT}(X_{n+1} = \sigma | X_{1:n} = x_{1:n}) = \frac{c_{\sigma}(x_{1:n}) + \frac{1}{2}}{n + \frac{|\text{mathcal{A}}|}{2}} \tag{4.23}
\]

This follows from the factorial-like properties of the Gamma function.

**Proposition 4.9** (Properties of KT estimator). For each \( n \), \( P_{KT}^n \) is minimized on the sequences \( x_{1:n} \) such that \( \forall \sigma_1, \sigma_2 \in \mathcal{A}, |c_{\sigma_1}(x) - c_{\sigma_2}(x)| \leq 1 \) (that is, the ‘flat’ sequences). \( P_{KT}^n \) is symmetric under permuting \( x_{1:n} \) and reordering of \( \mathcal{A} \).
CHAPTER 4. UNIVERSAL COMPRESSION

Notation 4.10. Given the symmetry of both the KT distribution and the ML
distribution for \( \mathcal{A} \), it is sometimes convenient to write probabilities with respect
to the number of counts. That is, if \((c_1, \ldots, c_d) = (c_1(x_1: n), \ldots, c_d(x_1: n)) \) for some
\( x_1: n \in \mathcal{A}^n \) define

\[
P_{KT}(c_1, \ldots, c_d) := P_{KT}(x_1: n) \tag{4.24}
\]

\[
P_{ML}(c_1, \ldots, c_d) := P_{ML}(x_1: n) \tag{4.25}
\]

4.3.1 Bound for KT distribution

Theorem 4.11 (KT bound for iid processes). Consider the class of distributions,
\( \{ P_{\theta} \mid \theta \in \Theta_0(\mathcal{A}) \} \). Denote the redundancy coding with respect to the maximum
likelihood probability (the iid ML) and the KT estimator by

\[
\rho_{ML,KT}(x_1: n) := \rho_{P_{\theta ML}(x_1: n), P_{KT}(x_1: n)} \tag{4.26}
\]

Let \( d = |\mathcal{A}| \). Then for any \((x_i)_{i \in \{1, \ldots, n\}}\)

\[
\rho_{ML,KT}(x_1: n) \leq d - \frac{1}{2} \cdot \log n + \log d \leq d \log n + \log d \tag{4.27}
\]

Corollary 4.12. The KT estimator \( P_{KT} \) is extremely uniformly pointwise uni-
versal over the class of iid process distributions.

Proof. For any \( \theta \in \Theta_0(\mathcal{A}) \), any \((x_i)_{i \in \mathbb{N}} \) and \( n \in \mathbb{N} \) we have \( P_{\theta}(x_1: n) \leq P_{ML}(x_1: n) \).
Also

\[
\lim_{n \to \infty} \frac{1}{n} \cdot \left( d - \frac{1}{2} \cdot \log n + \log d \right) = 0 \tag{4.28}
\]

Proof of theorem (4.11). Assume \( \mathcal{A} = \{1, \ldots, d\} \). For \( i \in \{1, \ldots, d\} \), let \( c_i = c_i(x_1, \ldots, x_n) \). Exponentiating both sides of (4.27) it is equivalent to show

\[
\frac{P_{ML}(c_1, \ldots, c_d)}{P_{KT}(c_1, \ldots, c_d)} \leq n^{\frac{d-1}{2}} \cdot d \tag{4.29}
\]

The bound holds for \( n = 1 \) since for any \( \sigma \in \mathcal{A} \) we have \( P_{ML}(\sigma) = 1 \), \( P_{KT}(\sigma) = \frac{1}{d} \),
and therefore \( \rho = \log d \). Now assume the bound holds for counts \((c_1, \ldots, c_d)\)
(i.e. assume equation (4.29) holds). We will show that if count \( c_1 \) increases by
1, the bound still holds. Since both probabilities are symmetric under symbol
permutations, this will imply the bound holds if any of \( c_1, \ldots, c_d \) increases. From this, it follows by induction that the bound holds for any sequence \( x_{1,n} \).

\[
P_{KT}(c_1 + 1, c_2, \ldots, c_d) = \frac{c_1 + \frac{1}{2}}{n + \frac{d}{2}} \cdot P_{KT}(c_1, \ldots, c_d)
\]

\[
P_{ML}(c_1 + 1, c_2, \ldots, c_d) = \frac{n^n}{(n+1)^{n+1}} \cdot \frac{(c_1 + 1)^{c_1 + 1}}{c_1^{c_1}} \cdot P_{ML}(c_1, \ldots, c_d)
\]

where we assume \( 0^0 = 1 \). Thus, by substitution, the result holds for counts \((c_1 + 1, c_2, \ldots, c_d)\) if

\[
\frac{n^n}{(n+1)^{n+1}} \cdot \frac{(c_1 + 1)^{c_1 + 1}}{c_1^{c_1}} \cdot P_{KT}(c_1, \ldots, c_d) \leq (n+1)^{\frac{d-1}{2}} \cdot d
\]  

(4.30)

By the inductive assumption, this holds if

\[
\frac{n^n}{(n+1)^{n+1}} \cdot \frac{(c_1 + 1)^{c_1 + 1}}{c_1^{c_1}} \cdot \frac{c_1^{c_1}}{n + \frac{d}{2}} \cdot n^{\frac{d-1}{2}} \cdot d \leq (n+1)^{\frac{d-1}{2}} \cdot d
\]  

(4.31)

Simplifying,

\[
\frac{n^{n+\frac{d+1}{2}} (n + \frac{d}{2})}{(n+1)^{n+1+\frac{d+1}{2}}} \cdot \frac{(c_1 + 1)^{c_1 + 1}}{(c_1 + \frac{1}{2}) \cdot c_1^{c_1}} \leq 1
\]  

(4.32)

It can be shown analytically (using elementary calculus) that the left hand fraction is bounded above by \( \frac{1}{e} \) and the right hand fraction is bounded above by \( e \). For details, see ([2]). By induction, we conclude the result holds for all sequences.

\[ \square \]

### 4.4 Universal prediction for Markov processes

We have seen that for iid sources the KT estimator is an extremely uniformly universal distribution over iid sequences. We now present an extremely uniformly universal distribution for the class of \( k \)-order Markov processes, the \( k \)-KT distribution. It uses the KT distribution for each conditional distribution \( P(\cdot | x) \), \( x \in \mathcal{A}^k \) in the parametrisation given in Remark 3.11.

**Definition 4.13.** \( k \in \mathbb{N} \). The \( k \)-KT distribution for \( \mathcal{A} \) is the stochastic process distribution \( \{ P^n_{k-KT} \}_{n \in \mathbb{N}} \) defined by

\[
P_{k-KT}(x_{1:n}) := \frac{1}{|\mathcal{A}|^k} \prod_{s \in \mathcal{A}^k} P_{KT}(c_s(x_{1:n}))
\]  

(4.33)
It is clear that this defines a distribution over $\mathcal{A}^n$ for each $n \in \mathbb{N}$.

### 4.4.1 Bound for $k$-KT distribution

The following theorem is a generalized version of Theorem (4.11) for $k$-order Markov processes.

**Theorem 4.14.** Let $k \in \mathbb{N}$ and $\Theta_k(\mathcal{A})$ parametrise the $k$-order Markov distributions for $\mathcal{A}$ valued processes, as in Remark 3.11. Let $d = |\mathcal{A}|$. Then the $k$-KT distribution achieves the following redundancy bound: For any $\theta \in \Theta(\mathcal{A})$ and $(x_i)_{i \in \{1, \ldots, n\}} \in \mathcal{A}^n$

$$
\rho_{\theta,k KT}(x_{1:n}) \leq d^{k+1}(\log n + \log d) \quad (4.34)
$$

Note that slightly stronger bounds can be derived, but this bound clearly emphasizes the important qualitative aspect, namely:

**Corollary 4.15.** The $k$-KT distribution is extremely uniformly pointwise universal over the class of $k$-order Markov processes.

In the next chapter a new result is sketched that a slight adaptation of $k$-KT performs universally with respect to a different parametric class of distributions. The proof of Theorem 4.14 is omitted here but the essential ideas form the skeleton of the proof in the following chapter.
Chapter 5

Coding with $k(.)$-KT for Piecewise iid Processes

This section sketches\footnote{unfortunately, due to time constraints, typing the proof was not able to be completed.} a new theorem related to the performance of the $k$-KT distribution. In particular, we consider a variation of the $k$-KT distribution, which allows $k$ to depend on $n$. The variation is shown to be is almost surely universal over the class of piecewise iid sources (defined below). Roughly, these are sequences which are for the most part iid, but whose parameters abruptly change at a few points.

It seems far fetched to suggest that the $k(.)$-KT distribution, whose brother $k$-KT was shown last chapter to be optimal for $k$ order Markov sources, might perform well for such a different class.

It is also conjectured (with reasons given) that an analogous result holds for more general piecewise Markov sources.

Roughly, the variation of $k$-KT allows a varying Markov order $k(.)$ which grows slowly with $n$.

5.1 Piecewise iid Processes

**Definition 5.1.** An $m$ piece iid process is defined by an $m$-tuple of iid processes, $((X_i^{(1)})_{i \in \mathbb{N}}, \ldots, (X_i^{(m)})_{i \in \mathbb{N}})$ such that the whole collection $\{X_i^{(j)} \mid j \in \{1, \ldots, m\}, i \in \mathbb{N}\}$ forms an independent collection random variables.

Concatenation of partial sequences: Given a tuple of piecelengths, $(n^{(1)}, \ldots, n^{(m)})$
we are interested in compressing the joint random variable
\[ (X_1^{(1)}, \ldots, X_n^{(1)}, X_1^{(2)}, \ldots, X_n^{(2)}, \ldots, X_1^{(m)}, \ldots, X_n^{(m)}) \tag{5.1} \]
with respect to the joint distribution, particularly as the total length \( \sum n_i \) diverges to infinity. For simplicity, the simpler case of two pieces with equal length \( n \) will be analysed:
\[ X_1^{(1)}, \ldots, X_n^{(1)}, X_1^{(2)}, \ldots, X_n^{(2)} \tag{5.2} \]

**Definition 5.2** \((k_{(\cdot)}\text{-KT estimator})\). As usual, let \( \mathcal{A} \) denote a finite space, and for each \( k \in \mathbb{N} \) let \( P_{k,-KT}(x_{1:n}) \) denote the \( k \)-KT distribution for \( \mathcal{A} \). Then for a function
\[ k_{(\cdot)} : \mathbb{N} \rightarrow \mathbb{N} \tag{5.3} \]
define the \( k_{(\cdot)} \)-KT distribution by
\[ P^n_{k_{(\cdot)}} := P^n_{k_{(n)}-KT} \tag{5.4} \]

**Notation 5.3.** Parametrise the class of distributions for \( \mathcal{A} \)-valued \( m \)-piece iid processes \(((X_i^{(i)})_{i \in \mathbb{N}}, \ldots, (X_i^{(m)})_{i \in \mathbb{N}})\) by \( \Theta_0^m(\mathcal{A}) := \Theta_0(\mathcal{A})^m \), and use the canonical parametrisation of \( \theta = (\theta^1, \ldots, \theta^m) \) where \( \theta^i \) is the parameter of the distribution of \( (X^{(i)}) \) for each \( i \in \{1, \ldots, m\} \). Also, denote
\[ P_{\theta}(x_1, \ldots, x_{m:n}) := \prod_{i=1}^m P_{\theta^i}(x_{(i-1)n+1}, \ldots, x_{in}) \tag{5.5} \]
where \( P_{\theta^i}(x) \) is defined [somewhere above] for any \( x \in \mathcal{A}^*; \theta^i \in \Theta_0(\mathcal{A}) \). Then since the whole collection is independent, we have
\[ P_{\theta}(x_1, \ldots, x_{m:n}) = \mathbb{P}(X_1^{(1)} = x_1, \ldots, X_n^{(m)} = x_{m:n}) \tag{5.6} \]

**Remark 5.4.** With this notation, the definition of redundancy (Definition (4.1)) and notions of universality (Definition (4.3)) are well/clearly defined for \( m \)-piece iid sources.

### 5.2 Basic proof

This section presents the simplest guarantee of ‘universality’ of the \( k_{(\cdot)} \)-KT distribution for 2 piece binary iid processes. More precisely, it is shown that for an appropriate choice of function \( k : \mathbb{N} \rightarrow \mathbb{N} \) the \( k_{(\cdot)} \)-KT distribution is universal almost surely for any binary valued two piece iid process \(((X^{(1)}, X^{(2)}))\), provided its parameter \( \theta \) in the interior of \( \Theta_0^m(\mathcal{A}) \).
Theorem 5.5 (Basic binary interior version). Let $\mathcal{A} = \{0, 1\}$. There exists a function (increasing) $k : \mathbb{N} \to \mathbb{N}$ such that the $k(\cdot)$-KT distribution for $\mathcal{A}$ is almost surely universal over the interior of $\Theta_0^{(2)}(\mathcal{A})$.

Initial outline of proof. For fixed $n$ and undetermined $k : \mathbb{N} \to \mathbb{N}$ the (random) counts $c_s(X_n^{(1)}, \ldots, X_n^{(2)})$ following each context $s \in \{0, 1\}^k$ are shown to be concentrated near the mean. This is done using a novel decomposition of $c_s\sigma$ into sums of (non-overlapping) independent random variables and applying Hoeffding’s inequality. It is shown that these ‘typical’ sequences (those with counts concentrated near the mean) have small redundancy using the key bound of the KT distribution (Theorem 4.11).

Theorem 5.6 (Hoeffding’s Inequality). Let $Z_1, \ldots, Z_L$ be iid Bernoulli, $Z_i \in \{0, 1\}$ a.s. Let $Y = \sum_{i=1}^L Z_i$. Then $\forall \epsilon > 0$

$$
P \left( \left| Y - \mathbb{E}Y \right| > \epsilon \right) \leq 2e^{-\frac{\epsilon^2}{2}}
$$

(5.7)

Theorem 5.7 (Hoeffding’s Inequality (advanced)). Let $Z_1, \ldots, Z_L$ be iid Bernoulli distributed with $Z_i \in \{0, 1\}$ a.s. and $p = \mathbb{P}(Z_1 = 1)$. If $Y = \sum_{i=1}^L Z_i$ then $\forall \epsilon > 0$

$$
P( Y - \mathbb{E}Y > \epsilon ) \leq e^{-\frac{\epsilon^2 g(p)}{2L}}
$$

(5.8)

where

$$
g(p) = \begin{cases} 
\frac{1}{1-2p} \ln \frac{1-p}{p} & \text{if } 0 < p < \frac{1}{2} \\
\frac{1}{2p(1-p)} & \text{if } \frac{1}{2} \leq p < 1
\end{cases}
$$

(5.9)

Hoeffding’s inequality is well known. The above is a (partial) direct quotation from Hoeffding’s original paper [13] where a proof can be found.

Remark 5.8. In what follows, Hoeffding’s inequality will be applied to sums of indicator random variables, $1_s(X_{1:k})$ for $s \in \mathcal{A}^k$. As $k$ grows, the probability $\mathbb{P}(X_{1:k} = s)$ decreases exponentially, and uniformly over $s$. We take advantage of this property using this version of Hoeffding’s inequality as follows:

Let $g$ be as defined in (5.9). Then for any $p \in (0, \frac{1}{2})$ we have $g(p) > \ln \frac{1-p}{p}$. Note also that Theorem 5.7 yields a bound for $\mathbb{P}(\mathbb{E}Y - Y > \epsilon)$ by symmetry. Then applying the union bound and noting that $\frac{1}{2p(1-p)} > \ln \frac{1-p}{p}$ we get the following bound

$$
P \left( \left| Y - \mathbb{E}Y \right| > \epsilon \right) \leq 2e^{-\frac{\epsilon^2 \ln \frac{1-p}{p}}{2}}
$$

(5.10)

$$
= 2 \left( \frac{p}{1-p} \right) \frac{\epsilon^2}{2}
$$

(5.11)
Further, \( \frac{p}{1-p} < 2p \) which yields

\[
P(|Y - \mathbb{E}Y| > \epsilon) \leq 2 \cdot (2p)^{\epsilon^2/L} \tag{5.12}
\]

Since \( p < \frac{1}{2} \), this bound is already exponentially decreasing in \( \epsilon^2 \). However, if \( p = \mathbb{P}(X_{1:k} = s) \) then \( p \) also decreases exponentially with \( k \) (unless the distribution of \( (X_i) \) is a Dirac mass, in which case there is no need for such a bound). Thus, (5.12) gives an exponentially decreasing bound in both \( k \) and \( \epsilon^2 \).

**Lemma 5.9.** Suppose \( (X_i)_{i \in \mathbb{N}} \) is iid according to \( P \in \{ P \theta \mid \theta \in \Theta_0 \} \). The expected count for a given string \( s \) in \( X_{1:n} \) is

\[
\mathbb{E}[c_s(X_1, \ldots, X_n)] = (n - l(s) + 1) \cdot P(s) \tag{5.13}
\]

The following corollary is a novel extension of Hoeffding’s inequality to the count of a string \( s \) in an iid sequence (to compare, Theorem 5.7 can be viewed as a result about the count of strings whose length is one). The result uses the form (5.12) and is an essential ingredient \(^2\) of the proof of Theorem 5.5/5.11.

**Corollary 5.10.** Suppose \( X_1, \ldots, X_n \) are iid, \( k \in \mathbb{N} \) and \( s \in A^k \). Also, assume \( \mathbb{P}(X_{1:k} = s) = p_s < \frac{1}{2} \). Then the following tail bound holds for the expected count of \( s \) in \( (X_1, \ldots, X_n) \)

\[
P\left(\left| c_s(X_{1:n}) - \mathbb{E}[c_s(X_{1:n})]\right| > k \epsilon\right) \leq 2k(2p)^{\epsilon^2/k/n} \tag{5.14}
\]

**Proof.** By Notation (4.4), \( c_s(X_1, \ldots, X_n) \) is a sum of indicator random variables. If \( l(s) \geq 2 \), contiguous indicator variables are correlated and so Hoeffding’s inequality can’t be applied directly. We can, however, decompose \( c_s \) as a sum of sums of disjoint (and therefore independent) indicator variables as follows:

\[
c_s(X_{1:n}) = \sum_{t=0}^{n-l} \mathbb{I}_s(X_{t+1:t+k})
\]

\[
= \sum_{j=1}^k \sum_{r=0}^{L_j-1} \mathbb{I}_s(X_{r+k+j : j+(r+1)k+j})
\]

\[
= \sum_{j=1}^k Y_j
\]

\(^2\)there is no claim that the particular form (5.12) is necessary for the later proof, but being the strongest known bound, it it yields the fastest known convergence rate.
5.3. **Performance of \( k(\cdot) \)-KT for General Piecewise Environments**

Theorem 5.11 (Main theorem). Let \( m \in \mathbb{N} \), \( \mathcal{A} \) a finite space. There exists a function \( k : \mathbb{N} \to \mathbb{N} \) such that the \( k(\cdot) \)-KT distribution for \( \mathcal{A} \) is almost surely universal over the class of \( m \)-piece iid processes, \( \{ P_\theta \mid \theta \in \Theta_0^m \} \).

**Proof of theorem 5.11.** Let \( X = (X_j)_{j \in \mathbb{N}}, \ldots, (X_j)_{j \in \mathbb{N}} \) be an \( m \)-piece iid process with distribution parametrised by \( \theta = (\theta^0, \ldots, \theta^m) \). Since \( X \) is arbitrary, it suffices to show the existence of a function \( k : \mathbb{N} \to \mathbb{N} \), which does not depend on...
\[ \theta, \text{ for which} \]
\[ \frac{1}{n} \rho_{\theta, k(n)KT}(X_1^{(i)}, \ldots, X_m^{(m)}) \xrightarrow{n \to \infty} 0 \]  
(5.21)

almost surely.

Define
\[ A_n(\alpha, k) = \left\{ x \in A^{m \cdot n} : \forall s \in A^k, \forall \sigma \in A, \forall i \in \{1, 2\}, \left| c_{ss}(x^{(i)}) - \mathbb{E}[c_{ss}(X^{(i)})] \right| \leq \alpha_{ss}^{(i)} \right\} \]  
(5.22)

where
\[ \alpha_{ss}^{(i)} := \begin{cases} \alpha & \text{if } P_{\theta}(s\sigma) > 0 \\ 0 & \text{if } P_{\theta}(s\sigma) = 0 \end{cases} \]  
(5.23)

\(A_n(\alpha, k)\) may be referred to as the ‘typical sequences’, as it contains only sequences for which all counts following each \(k\)-length context \(s\) are \(\alpha\)-close to the mean. We will show that the redundancy on \(A_n\) is uniformly bounded and grows sublinearly in \(n\) for a suitable choice of \(k\).

Let \(x = x_{1:mn} \in A_n(k(\cdot), \alpha_n)\), where \(k\) and \(\alpha\) are functions \(\mathbb{N} \to \mathbb{N}\) yet to be determined. For brevity, let \(k = k(\cdot), \alpha = \alpha_n\), and let \(x^{(i)} := x_{(i-1)n+1:i:n}\. \) Thus, by concatenating, we have \(x = x^{(1)} \cdots x^{(m)}\). That is, \(x^{(i)}\) denotes the part of \(x\) corresponding to \(X^{(i)}\) in the redundancy.

We decompose the redundancy of \(x\) according to the counts following each context \(s\):
\[ \rho = \log \frac{\prod_{i \in \{1, \ldots, m\}} \left( P_{\theta}(x^{(i)}_{1:k}) \cdot \prod_{s \in A^k} P_{\theta}(c_s(x^{(i)})) \right)}{\frac{1}{d^k} \cdot \prod_{s \in A^k} P_{KT}(c_s(x^{(i)} \cdots x^{(m)}))} \]  
(5.24)

\[ = \sum_{s \in A^k} \log \frac{\prod_i P_{\theta}(c_s(x^{(i)}))}{P_{KT}c_s(x^{(i)} \cdots x^{(m)})} + \log \left( d^k \cdot \prod_i P_{\theta}(x^{(i)}_{1:k}) \right) \]  
(5.25)

where the last term is from the first \(k\) bits of each segment, for which the \(k\)-KT would ordinarily have no context. This is bounded above by \(\log d^k\) whose growth is of the order \(k(n)\). We thus assert the following restriction on \(k(n)\):
\[ k(n) \in o(n) \]  
(5.26)

where \(f(\cdot) \in o(g(\cdot))\) if and only if \(\lim_{n \to \infty} f(n)/g(n) = 0\).
5.3. PERFORMANCE OF $K(\cdot)$-KT FOR GENERAL PIECEWISE ENVIRONMENTS

Since the KT distribution approximates the ML distribution well (Theorem 4.11) the following decomposition seems reasonable:

$$\sum_{s \in A^k} \log \frac{\prod_i P_{\theta_i}(\bar{c}_s(x^{(i)}))}{P_{KT}(\bar{c}_s(x^{(1)} \ldots x^{(m)}))} = \sum_{s \in A^k} \log \frac{\prod_i P_{\theta_i}(\bar{c}_s(x^{(i)}))}{\prod_i P_{\lambda_s}(\bar{c}_s(x^{(i)}))} + \sum_{s} \log \frac{\prod_i P_{\lambda_s}(\bar{c}_s(x^{(i)}))}{P_{KT}(\bar{c}_s(x^{(1)} \ldots x^{(m)}))}$$

(5.27)

We will first bound the second term of (5.27). By independence and recalling Notation 4.10 the following holds for any $\lambda \in \Theta_0$:

$$\prod_{i=1}^{m} P_{\lambda}(\bar{c}_s(x^{(i)})) = P_{\lambda_s}(\sum_{i=1}^{m} c_s(x^{(i)}))$$

(5.28)

Also, the counts following a given context $s$ in $x$ can be decomposed as follows:

$$c_s(x^{(1)} \ldots x^{(m)}) = \sum_{i=1}^{m} c_s(x^{(i)}) + \sum_{i=2}^{m} c_s(x^{(i-1)}_{n-k+1:n} \cdot x^{(i)}_{1:k})$$

(5.29)

Combining (5.29) and (5.28) yields

$$\prod_{i} P_{\lambda_s}(\bar{c}_s(x^{(i)})) = \frac{P_{\lambda_s}(\bar{c}_s(x^{(1)} \ldots x^{(m)}))}{P_{\lambda_s}(\sum_{i=2}^{m} c_s(x^{(i-1)}_{n-k+1:n} \cdot x^{(i)}_{1:k}))}$$

(5.30)

Therefore the rightmost term of (5.27) can be decomposed as follows:

$$\sum_{s} \log \frac{\prod_{i} P_{\lambda_s}(\bar{c}_s(x^{(i)}))}{P_{KT}(\bar{c}_s(x^{(1)} \ldots x^{(m)}))}$$

$$= \sum_{s} \log \frac{P_{\lambda_s}(\bar{c}_s(x^{(1)} \ldots x^{(m)}))}{P_{\lambda_s}(\sum_{i=2}^{m} c_s(x^{(i-1)}_{n-k+1:n} \cdot x^{(i)}_{1:k}))} - \sum_{s} \log P_{\lambda_s}(\sum_{i=2}^{m} c_s(x^{(i-1)}_{n-k+1:n} \cdot x^{(i)}_{1:k}))$$

(5.32)

$$\leq d^k \left( \frac{d-1}{2} \log (m \cdot n) + \log d \right) - \sum_{s} \log P_{\lambda_s}(\sum_{i=2}^{m} c_s(x^{(i-1)}_{n-k+1:n} \cdot x^{(i)}_{1:k}))$$

(5.33)

where the inequality (5.33) follows from Theorem 4.11. We will thus require

$$\left( n \mapsto d^k \left( \frac{d-1}{2} \log (m \cdot n) + \log d \right) \right) \in o(n)$$

(5.34)

The rightmost term of (5.33) corresponds to the probability component of the KT estimator in the first $k$ bits of each piece, other than the first. Provided the choice of $\lambda_s$ is reasonable, this redundancy incurred will be negligible.

Indeed, let us assert that $\lambda_s$ has a ‘uniform buffer’ - that is, $\lambda_s$ is of the form

$$\lambda_s = \gamma(n) \cdot u + (1 - \gamma(n)) \bar{\lambda}_s$$

(5.35)
where \( u \) is the parameter for the uniform distribution. Thus, the rightmost term of (5.33) can be written

\[
- \sum_s \log P_\lambda \left( \sum_{i=2}^m \frac{c_s(x_{n-k+1:n}' \cdot x_{1:k}^0)}{\bar{c}_s(x_{n-k+1:n}' \cdot x_{1:k}^0)} \right)
\]

\[
= - \sum_s \log \left( \gamma(n) P_u \left( \sum_{i=2}^m \frac{c_s(x_{n-k+1:n}' \cdot x_{1:k}^0)}{\bar{c}_s(x_{n-k+1:n}' \cdot x_{1:k}^0)} \right) - (1 - \gamma(n)) P_\lambda \sum_{i=2}^m \frac{c_s(x_{n-k+1:n}' \cdot x_{1:k}^0)}{\bar{c}_s(x_{n-k+1:n}' \cdot x_{1:k}^0)} \right)
\]

\[
\leq - \sum_s \log \left( \gamma(n) P_u (\cdots) \right)
\]

\[
\leq - \sum_s \log \left( \gamma(n) \cdot \left( \frac{1}{d} \right)^k \right)
\]

\[
= - d^k \log \gamma + k \cdot d^k \log d
\]  

We will thus require that

\[
( n \mapsto - d^k (n) \log \gamma(n) + k(n) \cdot d^k(n) \log d ) \in o(n)
\]  

We now bound the middle sum in (5.27). Since \( x \in A_n(k, \alpha) \), we have

\[
| c_{s\sigma}(x^0_i) - \mathbb{E}[c_{s\sigma}(x^0_i)] | \leq \alpha^{(i)}_{s\sigma}
\]  

for all \( \sigma \in \mathcal{A} \). Using the \( d \)-dimensional mean \( \mathbb{E}[c_s(x^0_i)] = (n - k) P_{\theta_0}(s) \cdot \theta^0 \) (this follows from Lemma 5.9 with strings of length \( k + 1 \)) and denoting \( \bar{c}_s^\sigma = (\alpha^{(i)}_{s\sigma_1}, \ldots, \alpha^{(i)}_{s\sigma_d}) \). Thus,

\[
\sum_{s \in \mathcal{A}^k} \log \frac{\prod_i P_{\theta_0}(c_s(x^0_i))}{\prod_i P_{\lambda_s}(c_s(x^0_i))} = \sum_{s \in \mathcal{A}^k} \sum_{i=1}^m \log \frac{P_{\theta_0}(c_s(x^0_i))}{P_{\lambda_s}(c_s(x^0_i))}
\]

\[
\leq \sum_{s \in \mathcal{A}^k} \sum_{i=1}^m \log \frac{P_{\theta_0}(\mathbb{E}[c_s(x^0_i)] - \bar{c}_s^\sigma)}{P_{\lambda_s}(\mathbb{E}[c_s(x^0_i)] + \bar{c}_s^\sigma)}
\]
5.3. PERFORMANCE OF $K(\cdot)$-KT FOR GENERAL PIECEWISE ENVIRONMENTS

This follows since the [CONTINUOUS EXTENSION] defined is decreasing in all variables.

\[
\begin{align*}
&= \sum_{s \in A_k} \sum_{i=1}^{m} \log \frac{P_{\theta^0}(s) \cdot (n-k) \cdot P_{\theta^0}(s) \cdot \theta^0 - \bar{\alpha}_s^0}{P_{\lambda^0}(s) \cdot (n-k) \cdot P_{\theta^0}(s) \cdot \theta^0 + \bar{\alpha}_s^0} \\
&\leq (n-k) \cdot \sum_{s \in A_k} \sum_{i=1}^{m} P_{\theta^0}(s) \cdot \log \frac{P_{\theta^0}(\theta^0)}{P_{\lambda^0}(\theta^0)} \\
&\quad + \sum_s \sum_i \log \frac{P_{\theta^0}(\bar{\alpha}_s^0)}{P_{\lambda^0}(-\bar{\alpha}_s^0)} \\
&\leq n \cdot \sum_{s \in A_k} \sum_{i=1}^{m} P_{\theta^0}(s) \cdot D_{KL}(P_{\theta^0} \| P_{\lambda^0}) \\
&\quad + \sum_s \sum_i \log P_{\theta^0}(\bar{\alpha}_s^0) \cdot P_{\lambda^0}(\bar{\alpha}_s^0) \\
\end{align*}
\]

Noting that divergence is decreasing in its second argument. The damping factor $\gamma$ gives a bound on $P_{\lambda^0}(\bar{\alpha}_s^0)$ namely

\[
P_{\lambda^0}(\bar{\alpha}_s^0) \geq \gamma P_{\lambda}(\bar{\alpha})
\]

If we define $\eta = \min \{ P_{\theta^0}(\sigma) \mid \sigma \in A, i \in \{1, \ldots, m\} \}$ we get the lower bound

\[
P_{\theta^0}(\bar{\alpha}_s^0) \leq \eta^{ad}
\]

Thus, the last term of (5.47) is bounded above by

\[
\sum_s \sum_i \log(\gamma d^{-ad} \eta^{ad})
\]

\[
= d^{k+1}(\log \gamma + ad \log(\eta/d))
\]

Thus we assert

\[
d^{k+1}(\log \gamma + ad \log(\eta/d)) \in o(n)
\]

The final, and most interesting part of the bound has been left to last. To bound the first term in (5.47) we now set

\[
\lambda_s = \text{argmax}_{\theta_i} P_{\theta^0}(s)
\]

Then for $\theta^0 = \lambda_s$ we get

\[
D_{KL}(P_{\theta^0}, P(\lambda_s)) \leq P_{\theta^0}(1 - \gamma)
\]
CHAPTER 5. CODING WITH $K_\gamma$-KT FOR PIECEWISE IID PROCESSES

!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!! INCOMPLETE SECTION OF PROOF:

All that remains is to show final divergence term is bounded by $mn \log \frac{1}{1-\gamma} + mn \log \frac{1}{\gamma} \cdot \kappa^k$ for some $\kappa \in (0, 1)$. Then all the boxed bounds are collected and you show the existence of a triple functions $\gamma, k$ and $\alpha$ with $\alpha > 1/2$ which satisfy all of the boxed conditions. !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!
We now turn to bound the probability of the event

\[ \tilde{A}_n(k, \alpha) = \left\{ (X_1^{(i)}, \ldots, X_m^{(i)}) \in A_n(k, \alpha) \right\} \]  

(5.56)

in terms of our undetermined functions \( k, \alpha \).

This part relies heavily on Corollary 5.10. Let \( s \in \mathcal{A}^k \) and \( \sigma \in \mathcal{A} \). Then applying Corollary 5.10 to the string \( s\sigma \) yields

\[ P \left( \left| c_{s\sigma}(X_1^{(i)}) - \mathbb{E}[c_{s\sigma}(X_1^{(i)})] \right| > (k + 1)\epsilon \right) \leq 2(k + 1)(2P_{\theta^{(i)}}(s))^{\epsilon^2(k+1)/n} \]

(5.57)

for each \( i \in \{1, \ldots, m\} \). Note the condition in the Corollary that the probability parameter \( P(X_1^{(i)} = s) = p_s \) must be less than \( \frac{1}{2} \). (The bound also holds trivially if the distribution of \( X_1^{(i)} \) is a Dirac mass). Now if we define

\[ \bar{\eta} := \max \left\{ P_{\theta^{(i)}}(\sigma) \mid i \in \{1, \ldots, m\}, \sigma \in \mathcal{A} \right\} \cap (0, 1) \]  

(5.58)

then for sufficiently large \( k \) we have for any \( i \in \{1, \ldots, m\} \) that for all \( s \in \mathcal{A} \), \( P_{\theta^{(i)}}(s) < \bar{\eta}^k \) (or \( P_{\theta^{(i)}} \) is a Dirac measure), and \( \bar{\eta}^k \xrightarrow{k \to \infty} 0 \). Thus for sufficiently large \( k \) the Corollary may be applied.

Now to each piece \( i \) we apply another union bound across \( s \in \mathcal{A}^k \) and \( \sigma \in \mathcal{A} \) to yield

\[ P \left( \exists s \in \mathcal{A}^k \exists \sigma \in \mathcal{A} \left| c_{s\sigma}(X_1^{(i)}) - \mathbb{E}[c_{s\sigma}(X_1^{(i)})] \right| > (k + 1)\epsilon \right) \]

\[ \leq \sum_{\sigma \in \mathcal{A}} \sum_{s \in \mathcal{A}^k} P \left( \left| c_{s\sigma}(X_1^{(i)}) - \mathbb{E}[c_{s\sigma}(X_1^{(i)})] \right| > (k + 1)\epsilon \right) \]

(5.59)

\[ \leq \sum_{\sigma \in \mathcal{A}} \sum_{s \in \mathcal{A}^k} 2(k + 1)(2P_{\theta^{(i)}}(s))^{\epsilon^2(k+1)/n} \]

(5.60)

\[ \leq 2^{1+\epsilon^2(k+1)/n}(k + 1) \sum_{\sigma \in \mathcal{A}} \sum_{s \in \mathcal{A}^k} (2P_{\theta^{(i)}}(s))^{\epsilon^2(k+1)/n} \]

(5.61)

\[ \leq 2^{1+\epsilon^2(k+1)/n}(k + 1) \cdot k \cdot \sum_{s \in \mathcal{A}^k} (P_{\theta^{(i)}}(s))^{\epsilon^2(k+1)/n} \]

(5.62)

\[ = 2^{1+\epsilon^2(k+1)/n}(k + 1) \cdot k \cdot \left( \sum_{\sigma \in \mathcal{A}} P_{\theta^{(i)}}(\sigma) \right)^{\epsilon^2(k+1)/n} \]

(5.63)

\[ \leq 2^{1+\epsilon^2(k+1)/n}(k + 1) \cdot k \cdot \left( d \cdot \bar{\eta}^{\epsilon^2(k+1)/n} \right)^k \]

(5.64)

\[ = 2^{1+\epsilon^2(k+1)/n}(k + 1) \cdot k \cdot d^k \cdot \bar{\eta}^{\epsilon^2(k+1)/n} \]

(5.65)

Finally, if we let \( \epsilon = \frac{\alpha}{k+1} \) and apply a final union bound over \( i \in \{1, \ldots, m\} \) we
get the following bound on the probability of \( \tilde{A}_n(k, \alpha) \):

\[
1 - P\left( \tilde{A}_n(k, \alpha) \right) = P\left( \{ \exists i \in \{1, \ldots, m\}, \exists s \in A^k \exists \sigma \in A \mid c_{s\sigma}(X_{1:n}^i) - E[c_{s\sigma}(X_{1:n}^i)] > \alpha \} \right) 
\]

(5.67)

\[
\leq \sum_{i=1}^{m} P\left( \{ \exists s \in A^k \exists \sigma \in A \mid c_{s\sigma}(X_{1:n}^i) - E[c_{s\sigma}(X_{1:n}^i)] > \alpha \} \right) 
\]

(5.68)

\[
\leq m \cdot 2^{1+ \frac{\alpha^2}{(k+1)n}} (k+1) \cdot k \cdot d^k \cdot \bar{\eta}^\alpha k/(k+1) 
\]

(5.70)

This is summable for \( \alpha = n^\delta \) for any \( \delta > 1/2 \). Thus we can apply the Borel Cantelli lemma.

\[\square\]

**Conjecture 5.12.** The \( k \cdot -KT \) distribution is also universal for the class of piecewise Markov processes \( \Theta_j^{(m)} \).

There is a corresponding version of Hoeffding’s inequality for Markov processes - [12]. It is believed that this may be used to replicate the above proof for the class \( \Theta_j^{(m)} \). The bound may give a suitable concentration bound for the typical sequences for Markov processes. This has not yet been investigated.

**Remark 5.13** (Remarks on the proof). Initially, it was not apparent that Hoeffding’s inequality could be applied to \( c_s = \sum \mathbb{1}_s(x_{i+1:i+k}) \) since the Bernoulli parameters were correlated. The more general inequality of Chebyshev was used in place of Hoeffding’s to bound the tail, and a ‘universal in probability’ result became apparent.

It is noted that the stronger version of Hoeffding’s inequality used in the proof (namely Theorem 5.7, Equation 5.12) makes no qualitative difference versus using the basic version in Theorem ?? (cf. Definition 4.3). However it technically gives you more ‘space to move’ with \( \alpha \) in (5.70) and could be useful for fine-tuning bounds on the convergence rate.

If the \( m \) parameters are all in the interior of \( \Theta_0^{(m)} \), there is no need to use the smoothing factor \( \gamma_n \). Instead, the discrepancy is bounded by the KL divergence \( D_{KL}(\cdot \parallel \cdot) \) between any two distributions. This leads to a simpler proof.

In line with Conjecture 5.12, there may be scope for applying the extended version of Hoeffding’s inequality for ergodic Markov chains directly for the iid case. It is conjectured that this could replace of the decomposition of \( c_s(X_{1:n}) \) (c.f. Conjecture 5.10) as a sum of independent random variables.
.1 Probability and Independence

Definition .14 (Probability space). Let $\Omega$ be a set, $\Sigma \subset \mathcal{P}(\Omega)$. Then $\Sigma$ is a sigma algebra ($\sigma$-algebra) for $\Omega$ if $\Sigma$ is closed under complements

$$A \in \Sigma \implies A^c \in \Sigma \quad (71)$$

and for all countable subsets $\{A_i \mid i \in I\} \subset \Sigma$

$$\bigcup_{i \in I} A_i \in \Sigma \quad (72)$$

If $\Sigma$ is a sigma algebra for $\Omega$ then $(\Omega, \Sigma)$ forms a measurable space. A probability space is a triple, $(\Omega, \Sigma, P)$ where $(\Omega, \Sigma)$ is a measurable space and $P : \Sigma \to [0, 1]$ satisfies

1. $P(\Omega) = 1$ and

2. whenever $\{A_i \mid i \in I\} \subset \Sigma$ is a subset of pairwise disjoint sets

$$P\left(\bigcup_{i \in I} A_i\right) = \sum_{i \in I} P(A_i) \quad (73)$$

$P$ is called a probability measure. $\Omega$ is called the sample space, elements $A$ of $\Sigma$ are called events and $P(A)$ is the probability of $A$.

Remark .15. We sometimes talk of discrete probability spaces, $(\Omega, \Sigma, P)$ where $\Omega$ is a countable set and $\Sigma = \mathcal{P}(\Omega)$. Where it is clear from context, mention of $\Sigma = \mathcal{P}(\Omega)$ is omitted. We may talk informally of the probability $P(\omega)$ of a point $\omega \in \Omega$, really meaning probability $P(\{\omega\})$ of the singleton event $\{\omega\} \in \Sigma$.

Definition .16 (Random variables). If $(\Omega, \Sigma)$ and $(N, \mathcal{B})$ are measurable spaces then a map $X : \Omega \to N$ is called measurable if for every $B \in \mathcal{B}$ we have $X^{-1}(B) \in \Sigma$. We write $X : (\Omega, \Sigma) \to (N, \mathcal{B})$. If $(\Omega, \Sigma, P)$ is a probability space and $X : (\Omega, \Sigma) \to (N, \mathcal{B})$, then $X$ is an $(N, \mathcal{B})$-valued random variable.

Note that if $(\Omega, \Sigma, P)$ is a probability space and $X : (\Omega, \Sigma) \to (N, \mathcal{B})$ then $X$ induces a probability measure $P'$ on $(N, \mathcal{B})$ defined for all $B \in \mathcal{B}$ by $P'(B) = P(X^{-1}(B))$. In this case, we write $X : (\Omega, \Sigma, P) \to (N, \mathcal{B}, P')$.

Definition .17. Sigma algebra generated by a random variable.
Definition 18 (Independence). Let $(\Omega, \Sigma, P)$ be a probability space. Events $A, B \in \Sigma$ are independent if $P(A \cap B) = P(A) \cdot P(B)$. A set $\{A_i\}_{i \in I} \subset \Sigma$ is independent if for every finite subset $J \subset I$,

$$P\left(\bigcap_{i \in J} A_i \right) = \prod_{i \in J} P(A_i) \quad (74)$$

A set of random variables $\{X_i : (\Omega, \Sigma, P) \rightarrow (c_i, B_i, P'_i)\}_{i \in I}$ is independent if for every finite subset $J \subset I$ and every map $f : J \rightarrow B_i$ the set $\{X^{-1}(f(i)) : i \in J\}$ is independent as a subset of $\Sigma$.

Definition 19 (Conditional probability). Let $(\Omega, \Sigma, P(\cdot))$ be a probability space and $B \in \Sigma$ with $P(B) \neq 0$. Then for $A \in \Sigma$ we define the conditional probability of $A$ given $B$ as

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (75)$$

Note that conditional probability here is not defined on sets of probability zero.

Lemma 20. The map

$$P(\cdot | B) : \Sigma \rightarrow \mathbb{R}$$

$$A \mapsto P(A|B) \quad (76)$$

is a probability measure on $(\Omega, \Sigma)$.

Thus, conditioning on $B$ induces a new probability space, $(\Omega, \Sigma, P(\cdot | B))$.

Definition 21 (Conditional independence). Let $(\Omega, \Sigma, P)$ be a probability space and $B \in \Sigma$ with $P(B) \neq 0$. A set $\{A_i\}_{i \in I} \subset \Sigma$ is conditionally independent given $B$ if $\{A_i\}_{i \in I}$ is an independent set of events in $(\Omega, \Sigma, P(\cdot | B))$. Likewise, a collection of random variables $C = \{X_i : (\Omega, \Sigma, P) \rightarrow (c_i, B_i, P'_i)\}_{i \in I}$ are conditionally independent given $B$ if they form an independent set in $(\Omega, \Sigma, P(\cdot | B))$.

Further, we define conditional independence given a random variable: Let $X : (\Omega, \Sigma, P) \rightarrow (N, B, P')$ be a random variable. A collection of random variables $C = \{X_i : (\Omega, \Sigma, P) \rightarrow (c_i, B_i, P'_i)\}_{i \in I}$ is conditionally independent given $X$ if for every $B \in B$ with $P'(B) \neq 0$, $C$ is conditionally independent of $X^{-1}(B)$ (which is an event in $\Sigma$). In particular, if $C = \{X_1, X_2\}$ we say $X_1$ is conditionally independent of $X_2$ given $B$. 

Bibliography


