MASTER

Entropy of hidden Markov models

van Wijk, A.C.C.

Award date:
2007

Technische Universiteit Eindhoven
Department of Mathematics and Computer Science

MASTER'S THESIS

Entropy of Hidden Markov Models

by A.C.C. van Wijk

Supervisors:

Dr. E.A. Verbitskiy (Philips Research)
Dr. R. Rietman (Philips Research)
Prof. dr. R.W. van der Hofstad (TUE)
Prof. dr. ir. S.C. Borst (TUE)

Eindhoven, August 2007

Preface

This thesis is the result of my final project to obtain the degree of Master of Science in Industrial and Applied Mathematics at the Eindhoven University of Technology. I did the specialization in Statistics, Probability and Operations Research. This document describes the work done during my eight-month internship at Philips Research Eindhoven in the Digital Signal Processing group.

I would like to thank my supervisors: Evgeny Verbitskiy and Ronald Rietman from Philips Research, and Sem Borst and Remco van der Hofstad from the university. In addition I want to thank the other students at Philips Research, which are too numerous to enumerate, for the great time I had there.

Summary

In this thesis we investigate the entropy of hidden Markov models. A hidden Markov model is a stochastic process $\{Y_n\}_{n\geq 0}$, which can be seen as a noisy observation of a Markov chain. The entropy is a measure for the randomness of the process. It is known that the conditional probability $\mathbb{P}[Y_0 = y_0 \mid Y_1 = y_1, \ldots Y_n = y_n]$ converges at an exponential rate. The literature on this is reviewed and different upper bounds for the convergence rate are compared. Next we give series expansions for this conditional probability in the special case of the so-called binary symmetric model. We consider expansions in different variables. A remarkable result for these expansions is that the coefficients in the beginning of the expansion will not change anymore as n becomes larger. Finally we describe a method to obtain a series expansion for the entropy making use of a recurrence relation for the given conditional probability.

Contents

Chapter 1

Introduction

In this thesis we investigate hidden Markov models: stochastic processes with an infinite memory. These processes can be seen as a noisy observation of a Markov chain. Where as for ordinary Markov chains the transition probabilities of going from one state to another depend only on the previous state, for hidden Markov models they depend on the entire history of the process.

Because of the mathematical structure of hidden Markov models, they have a wide range of applications. Examples are machine recognition, like speech and optical character recognition, and bioinformatics. In the last one they can be used to model DNA and protein sequences. The power of these models is that they can be very efficiently implemented and simulated.

The main focus of this work will be the entropy of hidden Markov models. The entropy is a measure for the amount of information a stochastic process contains. For this we will particularly look at the binary symmetric hidden Markov model, the simplest but non-trivial instance of these models.

The structure of this thesis is as follows. First hidden Markov models as well as entropy are introduced more precisely, and we will state the problem of interest. We will give a convergence result for the conditional probabilities in hidden Markov models, and we will review some literature for this result. Then we turn our attention to the entropy of these processes, especially to series expansions for this. Subsequently, making use of recurrence relations for the conditional probabilities, we derive an efficient way to compute the coefficients in one of these expansions. Finally we will give the most important conclusions of our work and recommendations for further research.

Chapter 2

Mathematical background

In this chapter we start by introducing hidden Markov models. These stochastic processes can be seen as a noisy observation of an ordinary Markov process. Hidden Markov processes give rise to three problems, which have already been solved efficiently in the literature. They will be briefly addressed. We will also briefly discuss some applications of hidden Markov models. We then focus on the binary symmetric case, the simplest hidden Markov models, which will be the main focus of this thesis. Given some conditions, a convergence result is proved. Two other definitions of a hidden Markov model are given as well, and it is proved that they are equivalent.

Next we will introduce the notion of entropy, especially the entropy rate of a stochastic process. We review a remarkable result from the literature, concerning the power series expansion for the entropy rate of a hidden Markov model. We will state and prove a theorem about the convergence of the conditional probabilities in the process. Finally we give the problem description of this thesis.

2.1 Hidden Markov models

2.1.1 Definition

A hidden Markov model is a stochastic process, wherein the transition probabilities of going from one state to another depend on the entire history of the process. It can be seen as a noisy observation of an ordinary Markov model. For this let $\mathcal{S}$ be a discrete state space, and let P be the $|\mathcal{S}| \times |\mathcal{S}|$ stochastic matrix with transition probabilities. Let $X = \{X_n\}_{n\geq 0}$ with $X_n \in \mathcal{S}$ be a Markov chain with transition probability matrix $P = (p_{ij})$, i.e.,

$$\mathbb{P}[X_{n+1} = x_j \mid X_n = x_i] = p_{ij},$$

for all $x_i, x_j \in \mathcal{S}$, and some initial distribution π for X. Let $\mathcal{S}'$ also be a discrete state space, not necessarily with an equal number of states as $\mathcal{S}$. Let Π be the so-called emission probability matrix, a stochastic matrix with dimension $|\mathcal{S}| \times |\mathcal{S}'|$. The states of $\mathcal{S}'$ are called the observed states. Then $Y = \{Y_n\}_{n\geq 0}$ is a hidden Markov model, where

$$\mathbb{P}[Y_n = y_k \mid X_n = x_j] = \Pi_{jk},$$

for all $y_k \in \mathcal{S}'$ and all $x_j \in \mathcal{S}$. So each state of X has a probability distribution over the possible states of Y. Given the state of X the state of Y is selected according to this distribution. These distributions are, for each state of X, given in the matrix Π.

Only the process Y is observed and the X-process is not known, i.e. hidden, which explains the name of these processes. The X-process will be referred to as the hidden or underlying Markov chain. In this way, the Y-process can be interpreted as observing the X-process through a noisy channel.

For a Markov process the next state of the process depends only on the previous state, or sometimes a fixed number of previous states. For hidden Markov models the transition probabilities depend on the entire history of the process, assuming that the underlying Markov chain is unknown. States further back in the past have fewer influence on these probabilities, although they still have some. This loss of influence happens at an exponential rate, as will be shown further on. Note that, although all past states have influence, hidden Markov processes can be efficiently simulated without having to keep track of the entire past. This is possible because of the fact that the underlying process X is Markovian, i.e. for that process only the last state has to be known. In order to simulate realizations of the process Y, one only has to keep track of the current state of X. Given the state X_n the observed state Y_n can be drawn, as well as the next state X_{n+1}.

2.1.2 Fundamental problems

Hidden Markov models have given rise to three fundamental problems, see [14, 31]. All three have already efficiently been solved in the literature. These problems as well as their solutions are briefly discussed here.

The first one, known as the *Evaluation problem*, asks for the probability that a given sequence of observations occurs, given the model parameters π, P and Π. So it asks for

$$\mathbb{P}[\{Y_i, \ldots, Y_j\} \mid \pi, P, \Pi\,].$$

This probability can be calculated using the *Forward-Backward Algorithm*.

The second one is the *Decoding problem*, which aims to find the most likely sequence of states of the underlying Markov process such that the probability of observing a sequence $\{Y_i, \ldots, Y_j\}$ is maximal. So it searches for the sequence $\{X_i, \ldots, X_j\}$ that maximizes

$$\mathbb{P}[\{X_i, \ldots, X_j\}, \{Y_i, \ldots, Y_j\} \mid \pi, P, \Pi\,].$$

A naive approach would be to calculate the probabilities for all possible sequences $\{X_i, \ldots, X_j\}$. An efficient algorithm however is the *Viterbi Algorithm*, see [16, 37].

The third fundamental problem is the *Learning problem*. For this the model parameters π, P, Π are supposed to be unknown and these are tried to be optimized given a sequence of observations, a so-called training sequence. So, which π, P, Π maximize

$$\mathbb{P}[\{Y_i, \ldots, Y_j\} \mid \pi, P, \Pi\,].$$

This is the hardest problem of the three. Although there is no known analytical method to solve this [31], the *Baum-Welch Algorithm* [4, 9] gives an iterative procedure to find a local maximum. Another algorithm for this is the *Segmental K-means Algorithm* [24].

2.1.3 Applications

We will briefly point out a few applications of hidden Markov models. More elaborate discussions can be found in [7, 12, 28].

One of the first applications was in speech recognition [31] to convert spoken language into text. For this, training sequences are used, to adapt the parameters of the model in order to obtain the best possible recognition. Other examples of pattern recognition are the recognition of optical characters, such as text and handwriting, gestures, body motion, etcetera.

Another field where hidden Markov models are used, is bioinformatics, see [11, 26], for instance in the modeling of DNA and protein sequences.

2.1.4 Binary symmetric case

The simplest non-trivial hidden Markov model is the binary symmetric one. For this let the state space of X be given by $\mathcal{S} = \{-1, 1\}$, and let the transition probability matrix P be given by

$$P = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix},$$

for some $p \in [0, 1]$. Let the state space of Y be given by $\mathcal{S}' = \mathcal{S}$, and let the emission probability matrix Π be given by

$$\Pi = \begin{pmatrix} 1-\delta & \delta \\ \delta & 1-\delta \end{pmatrix},$$

for some $\delta \in [0, 1]$.

The process X is a Markov process consisting of 1's and -1's. With probability p the state of X is 'flipped' with respect to the previous state, and so with probability $1-p$ it stays the same. For each X_n it is decided by means of an (unfair) coin flip if Y_n is just the value of X_n, or that it becomes $-X_n$. For this, let $Z_n \in \{-1, 1\}$, $n = 0, 1, \ldots$ be independent and identically distributed Bernoulli random variables, independent of the process X, with common distribution $\{\delta, 1-\delta\}$, i.e.

$$\mathbb{P}[Z_n = 1] = 1 - \delta = 1 - \mathbb{P}[Z_n = -1].$$

Now define Y_n as

$$Y_n = X_n \cdot Z_n.$$

This gives the process $\{Y_n\}_{n\geq 0}$. In this way this process can be seen as observing the process X through a so-called binary symmetric channel [8] with error probability δ. So with probability δ an error occurs and the state 1 is observed as an -1 or vice versa. Note that the process obtained in this way is symmetric both in p and in δ.

The initial distribution π of the X-process is taken to be equal to the stationary distribution of it, which is by symmetry:

$$\mathbb{P}[X_i = -1] = \mathbb{P}[X_i = 1] = \frac{1}{2}.$$

Again by symmetry this is also the stationary probability distribution for the Y-process:

$$\mathbb{P}[Y_i = -1] = \mathbb{P}[Y_i = 1] = \frac{1}{2}.$$

When investigating the entropy of hidden Markov processes in the sequel of this thesis, we will almost entirely focus on this binary symmetric process.

2.1.5 Alternative definitions

Hidden Markov models were introduced in Section 2.1.1 by the use of two stochastic matrices P and Π. Now two alternative definitions are given. The first one is named *Markov source*, the second one *grouped Markov chain*. It is proved that these definitions are equivalent to the previous one.

A Markov source, also called a function of a Markov chain, is defined as in [15]. Let $\tilde{X} = \{\tilde{X}_n\}_{n\geq 0}$ be a Markov chain with values in a finite set of states $\tilde{\mathcal{S}}$, with transition probability matrix Δ. Let the process $\tilde{Y} = \{\tilde{Y}_n\}_{n\geq 0}$ with state space $\tilde{\mathcal{S}}'$ be defined by the coordinate-by-coordinate transformation $f : \tilde{\mathcal{S}} \mapsto \tilde{\mathcal{S}}'$ given by

$$\tilde{Y}_n = f(\tilde{X}_n).$$

Here the number of values that $\tilde{Y}_n$ can take is smaller than that of $\tilde{X}_n$, i.e. $|\tilde{\mathcal{S}}| \leq |\tilde{\mathcal{S}}'|$. The process $\tilde{Y}$ is a Markov source.

In [22] a grouped Markov chain is defined. Let $\hat{X} = \{\hat{X}_n\}_{n\geq 0}$ with $\hat{X}_n \in \hat{\mathcal{S}}$ be a Markov chain, with transition probability matrix $\hat{P}$. Let the states of the chain be divided into $K = |\hat{\mathcal{S}}|$ mutually exclusive and exhaustive nonempty subsets $\mathcal{B}_1, \ldots, \mathcal{B}_K$. Define the process $\hat{Y} = \{\hat{Y}_n\}_{n\geq 0}$ with $\hat{Y}_n \in \{1, 2, \ldots, K\}$ by

$$\hat{Y}_n = i \;\Leftrightarrow\; \hat{X}_n \in \mathcal{B}_i.$$

The process $\hat{Y}$ is called a grouped Markov chain.

It is straightforward that a grouped Markov chain and a Markov source are equivalent. The following proposition, which is stated in [15] as an exercise, gives equivalence of hidden Markov models and grouped Markov chains.

Proposition 2.1. *Every grouped Markov chain can be written as a hidden Markov model, and conversely every hidden Markov model can be written as a grouped Markov chain.*

For the proof we refer to Appendix A.1.

Defining the binary symmetric process as a Markov source, one has the Markov process $V = \{V_n\}_{n\geq 0}$ where

$$V_n = (X_n, Z_n).$$

The state space of this process is given by

$$\tilde{\mathcal{S}} = \mathcal{S} \times \mathcal{S}' = \{(-1,1), (-1,-1), (1,1), (1,-1)\}.$$

Now $f(V_n)$ where $f : \tilde{\mathcal{S}} \mapsto \mathcal{S}'$ defined by

$$f(V_n) = f(X_n, Z_n) = X_n \cdot Z_n,$$

gives the hidden Markov process Y. Let Δ be the transition probability matrix of the process V, which as in (A.1.1) is given by

$$\Delta = \begin{pmatrix} (1-p)(1-\delta) & (1-p)\delta & p\,(1-\delta) & p\,\delta \\ (1-p)(1-\delta) & (1-p)\delta & p\,(1-\delta) & p\,\delta \\ p\,(1-\delta) & p\,\delta & (1-p)(1-\delta) & (1-p)\delta \\ p\,(1-\delta) & p\,\delta & (1-p)(1-\delta) & (1-p)\delta \end{pmatrix}. \tag{2.1.1}$$

This matrix gives the correct conditional probabilities for the individual processes X and Y. One could easily check that

$$\begin{aligned} &\mathbb{P}[X_n = X_{n+1}] = 1-p, && \mathbb{P}[Y_n = X_n] = 1-\delta, \\ &\mathbb{P}[X_n \neq X_{n+1}] = p, && \mathbb{P}[Y_n \neq X_n] = \delta, \end{aligned}$$

equivalent to the processes given by the matrices P and Π.

2.2 Entropy

The entropy is a measure for the amount of information a random variable or a stochastic process contains. We will only focus on the so-called Shannon entropy [32, 33]. This notion comes from the field of information theory [8]. The entropy gives a bound for the maximal achievable compression for the data generated by the process, and it that way it gives whether the data can be reliably transmitted over a given channel.

2.2.1 Definition

The *entropy* $H(U)$ [8] of a discrete random variable U, taking values in a set $\mathcal{U}$, is defined by

$$H(U) = -\sum_{U} \mathbb{P}[U] \log \mathbb{P}[U],$$

with the assumption $0 \log 0 = 0$. Here and throughout the sequel of this thesis, we use the notation $\mathbb{P}[U] = \mathbb{P}[U = u]$, and the summation should be understood as to be over all $u \in \mathcal{U}$. Note that $H(U)$ itself is not a random variable. From the definition it follows that the more uncertainty there is in U, the larger $H(U)$ will be. Note that the entropy can also be written as

$$H(U) = -\mathbb{E}[\log \mathbb{P}[U]],$$

which also defines the entropy of a continuous random variable.

Suppose that

$$U = \begin{cases} 1 & \text{with probability } p, \\ -1 & \text{with probability } 1-p, \end{cases}$$

for some $p \in [0, 1]$. Then the entropy of U is a function of p and is given by

$$H(U) = -p \log p - (1-p) \log(1-p) =: h(p).$$

The *entropy rate* $h(Y)$ [8] of a stochastic process $Y = \{Y_n\}_{n \geq 0}$ is defined by

$$h(Y) = \lim_{n\to\infty} \frac{1}{n+1} H(Y_0, \ldots, Y_n), \tag{2.2.1}$$

where

$$\begin{aligned} H(Y_0, \ldots, Y_n) &= -\sum_{Y_0}\sum_{Y_1} \cdots \sum_{Y_n} \mathbb{P}[Y_0, \ldots, Y_n] \log \mathbb{P}[Y_0, \ldots, Y_n] \\ &= -\mathbb{E}\left[\log \mathbb{P}[Y_0, \ldots, Y_n]\right], \end{aligned}$$

using the notation

$$\mathbb{P}[Y_0, \ldots, Y_n] = \mathbb{P}[Y_0 = y_0, Y_1 = y_1, \ldots, Y_n = y_n],$$

is the entropy of the random variable $U = (Y_0, \ldots, Y_n)$.

If the process Y is stationary, then the limit in (2.2.1) exists and is finite. In the next section we will give a proof of this, making use of the so-called subadditivity lemma.

Let $H(Y_n \mid Y_{n-1}, \ldots, Y_0)$ denote the conditional entropy, defined by

$$H(Y_n \mid Y_{n-1}, \ldots, Y_0) = -\mathbb{E}[\log \mathbb{P}[Y_n \mid Y_{n-1}, \ldots, Y_0]].$$

In [8] the following result for this is given:

Lemma 2.2. *For a stationary stochastic process Y it holds that*

$$\lim_{n\to\infty} \frac{1}{n+1} H(Y_0, \dots, Y_n) = \lim_{n\to\infty} H(Y_n \mid Y_{n-1}, \dots, Y_0).$$

This lemma gives an alternative way to calculate the entropy. The proof of is given in Appendix A.2.

In the sequel we will let the time run backwards. So we will consider the conditional probabilities $\mathbb{P}[\,Y_0 \mid Y_1, \dots, Y_n]$ and we consider the entropy as

$$h(Y) = \lim_{n\to\infty} H(Y_0 \mid Y_1, \dots, Y_n).$$

In [8] the following bounds are given for the entropy:

$$H(Y_0 \mid Y_1, \dots, Y_n, X_n) \le h(Y) \le H(Y_0 \mid Y_1, \dots, Y_n),$$

with equality in the limit as n tends to infinity.

2.2.2 Subadditivity lemma

In this section we prove that the limit in (2.2.1) exists. We follow the approach in [34] to prove the next proposition.

Proposition 2.3. *For a stationary stochastic process Y it holds that $\lim_{n\to\infty} \frac{1}{n+1} H(Y_0, \dots, Y_n)$ exists and is finite.*

For the proof we need the following lemma.

Lemma 2.4 (Subadditivity Lemma)**.** *If a sequence of real numbers $\{x_n\}$ satisfies the subadditivity condition*

$$x_{m+n} \le x_m + x_n, \text{ for all } m, n \ge 1, \tag{2.2.2}$$

then

$$\lim_{n\to\infty} \frac{x_n}{n} = \inf_{m\ge 1} \frac{x_m}{m}.$$

The proof is given in Appendix A.3. Making use of this lemma, the proof of Proposition 2.3 follows.

Proof of Proposition 2.3. For the entropy rate of the process Y it holds that [34]:

$$H(Y_0, \dots, Y_{m+n-1}) \le H(Y_0, \dots, Y_{m-1}) + H(Y_m, \dots, Y_{m+n-1}),$$

and so, by stationarity,

$$H(Y_0, \dots, Y_{m+n-1}) \le H(Y_0, \dots, Y_{m-1}) + H(Y_0, \dots, Y_{n-1}).$$

Let $h_n := H(Y_0, \dots, Y_{n-1})$, then this last line becomes

$$h_{m+n} \le h_m + h_n, \text{ for all } m, n \ge 1,$$

so $\{h_n\}$ satisfies the subadditivity condition (2.2.2). By Lemma 2.4 it then holds that

$$\lim_{n\to\infty} \frac{h_n}{n} = \inf_{m\ge 1} \frac{h_m}{m}.$$

As $h_n \ge 0$ we have $\frac{h_n}{n} \ge 0$ for all n, and the statement follows. □

2.3 Problem description

2.3.1 Series expansion entropy

We now return to the setting of hidden Markov models. The entropy rate of the binary symmetric hidden Markov model depends only on p and δ:

$$h(Y) = \lim_{n\to\infty} \frac{1}{n+1} H(Y_0, \ldots, Y_n) =: h_Y(p, \delta).$$

No closed-form expression for this seems to be known [39].

Han and Marcus [20] show that, under the assumption $p \in (0,1)$, $h_Y(p,\delta)$ is a real analytic function of p and δ. This will be discussed in Section 4.2. It implies that $h_Y(p,\delta)$ can be expressed as a convergent power series:

$$p \in (0,1): \; h_Y(p,\delta) = \sum_{k=0}^{\infty} C_k \delta^k,$$

where the C_k are functions of p.

Recall that $h(Y) = \lim_{n\to\infty} H(Y_0|Y_1, \ldots, Y_n)$. Zuk et al. [40] give a remarkable result for this, which holds for general hidden Markov processes:

$$H(Y_0|Y_1, \ldots, Y_n) = \sum_{k=0}^{\infty} C_k^{(n)} \delta^k,$$

where $C_k^{(n)} = C_k$ for $n \geq \lceil \frac{k+1}{2} \rceil$. So the coefficients $C_k^{(n)}$ 'stabilize' for n large enough. A proof of this is given in [40], but no intuition for this 'stabilization' of the $C_k^{(n)}$'s is given. The outline of this proof will be sketched in Section 4.1.3.

2.3.2 Convergence conditional probabilities

Given strict positivity of the matrices P and Π, the conditional probabilities in a general hidden Markov model can be shown to be positive and continuous. Recall the notation

$$\mathbb{P}[Y_0 \mid Y_1, \ldots, Y_n] = \mathbb{P}[Y_0 = y_0 \mid Y_1 = y_1, \ldots, Y_n = y_n].$$

Proposition 2.5. *Given $P, \Pi > 0$ it holds that*

$$\begin{gathered}\exists\, a, b \in (0,1) \quad \forall\, n \quad \forall\, Y_0, \ldots, Y_n \in \{-1, 1\}: \\ 0 < a \leq \mathbb{P}[Y_0|Y_1, \ldots, Y_n] \leq b < 1.\end{gathered}$$

This property is known as *finite energy.* It means that, regardless how much is known about the past, there is no absolute certainty about the next symbol of Y. This proposition will be proved in Section 3.1.

Let g be the limiting conditional probability as n tends to infinity, i.e.

$$\mathbb{P}[Y_0|Y_1, \ldots, Y_n] \overset{n\to\infty}{\longrightarrow} g(Y_0, Y_1, \ldots, Y_n, \ldots).$$

Proposition 2.6. *Given P, $\Pi > 0$ it holds that $\mathbb{P}[Y_0|Y_1, \ldots, Y_n]$ converges uniformly as $n \to \infty$. More precisely, let*

$$g_n(Y) = \mathbb{P}[Y_0|Y_1, \ldots, Y_n],$$

then there exist $\alpha > 0$ and C such that for all Y:

$$|g_n(Y) - g_m(Y)| \leq Ce^{-\alpha n}, \quad \forall\, n, m : n < m.$$

The proof of this proposition will be given in Section 3.1 as well. As a consequence of Proposition 2.5 and Proposition 2.6, we have that g is positive and continuous, hence $\mathbb{P}$ is a g-measure [25]. In the sequel of this thesis we will assume P, $\Pi > 0$.

For $h(Y)$ we have:

$$\begin{aligned} h(Y) &= \lim_{n \to \infty} H(Y_0|Y_1, \ldots, Y_n) \\ &= - \lim_{n \to \infty} \mathbb{E}[\, \log \mathbb{P}[\, Y_0|Y_1, \ldots, Y_n]\,] \\ &= -\mathbb{E}[\, \log g(Y_0, Y_1, \ldots, Y_n, \ldots)], \end{aligned}$$

where interchanging limit and expectation is allowed because of Proposition 2.6.

2.3.3 Result settlement coefficients

We now state one of the main results of this thesis, concerning the so-called ‘settlement’ of the coefficients in the power series expansion of the conditional probabilities.

Theorem 2.7. *Given transition probabilities $P > 0$ and emission probabilities $\Pi > 0$, there exist $F_k, \tilde{F}_k : \mathbb{R}^{k+3} \mapsto \mathbb{R}$ such that*

$$g(Y_0, Y_1, \ldots, Y_n, \ldots) = \sum_{k=0}^{\infty} F_k(p\,; Y_0, \ldots, Y_{k+1})\delta^k,$$

and even

$$g(Y_0, Y_1, \ldots, Y_n, \ldots) = \sum_{k=0}^{\infty} \tilde{F}_k(p\,; Y_0, \ldots, Y_{k+1})(\delta(1-\delta))^k.$$

Here $F_k^{(n)} = F_k$, for $n \geq k+1$. This is called the ‘settlement’ or ‘stabilization’ of the coefficients. It implies that the F_k are computable by a finite computation:

$$\mathbb{P}[Y_0|Y_1, \ldots, Y_n] = \sum_{k=0}^{\infty} F_k^{(n)} \delta^k,$$

The coefficients could be derived either analytically or by numerical computations.

This theorem will be proved in Section 4.1.2. It is a similar result to that found by Zuk et al. [39], which give this statement for the series expansion for the entropy, see Theorem 4.1.

This result is important as it reduces the computational complexity of the problem significantly. Instead of having to compute $F_k^{(n)}$ for all n to be able to compute $\lim_{n\to\infty} F_k^{(n)}$, we now only have to compute $F_k^{(n)}$ for one value of n large enough. Section 4.3 and Section 4.4, where we try to find an general expression for the coefficients F_k for the binary symmetric model, are based on this result.

Chapter 3

Literature review convergence

As stated in Section 2.3.2, it holds that $\mathbb{P}[Y_0|Y_1,\ldots,Y_n]$ converges uniformly as n tends to infinity. In this chapter we will review the literature on this result. We start by the work of Baum and Petrie (1966), who prove it along the lines of the two propositions given in Section 2.3.2. Then we focus on the work of Harris (1955), who gives a proof based on couplings. These were introduced by Doeblin (1937) and also studied by Vasershtein (1969). Next we look at the work of Birch (1962) and at the more recent work of Fernández, Ferrari and Galves (2002). The results hold for general hidden Markov models. The different upper bounds found for the rate of convergence, will be compared for the binary symmetric hidden Markov model.

3.1 Baum and Petrie

We follow the approach of Baum and Petrie [3] to prove the uniform convergence of $\mathbb{P}[Y_0|Y_1,\ldots,Y_n]$ as $n\to\infty$. The proof follows by proving Proposition 2.5 and Proposition 2.6, which we shall do in this section. We recall that the first proposition gave uniform bounds strictly between 0 and 1 for the conditional probability, where the second stated the convergence of it at an exponential rate.

3.1.1 Bounds conditional probability

In order to prove Proposition 2.5, we need the following lemma.

Lemma 3.1. *Suppose $x_i>0, y_i>0$ for all i, then*

$$\min_{i=1,\ldots,n}\frac{x_i}{y_i} \;\le\; \frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n y_i} \;\le\; \max_{i=1,\ldots,n}\frac{x_i}{y_i}.$$

Proof. As

$$\min_j \frac{x_j}{y_j} \le \frac{x_i}{y_i} \le \max_j \frac{x_j}{y_j}, \quad \forall i,$$

we have

$$\min_j \frac{x_j}{y_j} = \frac{\sum_{i=1}^{n} y_i \min_j \frac{x_j}{y_j}}{\sum_{i=1}^{n} y_i} \leq \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} y_i} = \frac{\sum_{i=1}^{n} y_i \frac{x_i}{y_i}}{\sum_{i=1}^{n} y_i} \leq \frac{\sum_{i=1}^{n} y_i \max_j \frac{x_j}{y_j}}{\sum_{i=1}^{n} y_i} = \max_j \frac{x_j}{y_j}. \qquad \square$$

We now prove Proposition 2.5, which holds for general hidden Markov models:

Proof of Proposition 2.5. The proof is based on writing out $\mathbb{P}[Y_0|Y_1,\dots,Y_n]$. For this the following four ideas will be used.

(1) By Bayes' Theorem we have

$$\mathbb{P}[Y_0|Y_1,\dots,Y_n] = \frac{\mathbb{P}[Y_0,Y_1,\dots,Y_n]}{\mathbb{P}[Y_1,\dots,Y_n]}.$$

(2) By conditioning on the states of X, we can write

$$\mathbb{P}[Y_0,\dots,Y_n] = \sum_{X_0,\dots,X_n} \mathbb{P}[Y_0,\dots,Y_n|X_0,\dots,X_n]\mathbb{P}[X_0,\dots,X_n].$$

(3) As X is a Markov Chain, we have

$$\begin{aligned}\mathbb{P}[X_0,\dots,X_n] &= \mathbb{P}[X_0]\mathbb{P}[X_1|X_0]\dots\mathbb{P}[X_n|X_{n-1}]\\ &= \mathbb{P}[X_0]\prod_{i=0}^{n-1}\mathbb{P}[X_{i+1}|X_i].\end{aligned}$$

(4) As Y_i only depends on X_i, for $i = 0,\dots,n$, we have

$$\begin{aligned}\mathbb{P}[Y_0,\dots,Y_n|X_0,\dots,X_n] &= \mathbb{P}[Y_0|X_0]\mathbb{P}[Y_1|X_1]\dots\mathbb{P}[Y_n|X_n]\\ &= \prod_{i=0}^{n}\mathbb{P}[Y_i|X_i].\end{aligned}$$

This gives:

$$\begin{aligned}&\mathbb{P}[Y_0|Y_1,\dots,Y_n]\\ &\overset{(1)}{=} \frac{\mathbb{P}[Y_0,Y_1,\dots,Y_n]}{\mathbb{P}[Y_1,\dots,Y_n]}\\ &\overset{(2)}{=} \frac{\sum_{X_0,X_1,\dots,X_n} \mathbb{P}[Y_0,\dots,Y_n|X_0,\dots,X_n]\mathbb{P}[X_0,\dots,X_n]}{\sum_{X_1,\dots,X_n} \mathbb{P}[Y_1,\dots,Y_n|X_1,\dots,X_n]\left(\sum_{X_0}\mathbb{P}[X_1,\dots,X_n|X_0]\mathbb{P}[X_0]\right)}\\ &\overset{(3,4)}{=} \frac{\sum_{X_0,X_1,\dots,X_n} \mathbb{P}[Y_0|X_0]\mathbb{P}[X_0]\prod_{i=0}^{n-1}\mathbb{P}[X_{i+1}|X_i]\prod_{i=1}^{n}\mathbb{P}[Y_i|X_i]}{\sum_{X_0,X_1,\dots,X_n} \mathbb{P}[X_0]\prod_{i=0}^{n-1}\mathbb{P}[X_{i+1}|X_i]\prod_{i=1}^{n}\mathbb{P}[Y_i|X_i]}.\end{aligned}$$

Note that the nominator and denominator are equal up to the term $\mathbb{P}[Y_0|X_0]$. Lemma 3.1 now gives

$$\min_{X_0} \mathbb{P}[Y_0|X_0] \le \mathbb{P}[Y_0|Y_1,\dots,Y_n] \le \max_{X_0} \mathbb{P}[Y_0|X_0].$$

As $\Pi > 0$ it follows that one could take

$$a = \min_{X_0} \mathbb{P}[Y_0|X_0] > 0, \qquad b = \max_{X_0} \mathbb{P}[Y_0|X_0] < 1,$$

which proves the statement of the proposition. □

For the binary symmetric case, we find, assuming $0 < \delta \le \frac{1}{2}$:

$$0 < \delta = \min_{X_0} \mathbb{P}[Y_0|X_0] \le \mathbb{P}[Y_0|Y_1,\dots,Y_n] \le \max_{X_0} \mathbb{P}[Y_0|X_0] = 1-\delta < 1.$$

In Appendix A.4 we will give two alternative proofs of this proposition.

3.1.2 Uniform convergence conditional probability

In this section we prove Proposition 2.6. The proof is rather long, but the result of this proposition is important, as it established the converges of $\mathbb{P}[Y_0|Y_1,\dots,Y_n]$ as $n \to \infty$.

Proof of Proposition 2.6. Let

$$\begin{aligned} g_n(Y) &= \mathbb{P}[Y_0|Y_1,\dots,Y_n], \\ g_n(Y,i) &= \mathbb{P}[Y_0|Y_1,\dots,Y_n, X_{n+1}=i], \\ \overline{g}_n(Y) &= \max_i g_n(Y,i), \\ \underline{g}_n(Y) &= \min_i g_n(Y,i). \end{aligned}$$

First we prove that

$$\underline{g}_n(Y) \le g_n(Y) \le \overline{g}_n(Y).$$

We have that, by conditioning on X_{n+1},

$$\begin{aligned} g_n(Y) &= \sum_i g_n(Y,i)\, \mathbb{P}[X_{n+1}=i \mid Y_1,\dots,Y_n] \\ &\le \left(\max_j g_n(Y,j)\right) \sum_i \mathbb{P}[X_{n+1}=i \mid Y_1,\dots,Y_n] \\ &= \overline{g}_n(Y), \end{aligned}$$

and

$$g_n(Y) \ge \min_j g_n(Y,j) = \underline{g}_n(Y).$$

Now we prove that, for some $\kappa \in (0,1)$:

$$\begin{aligned} \overline{g}_{n+1}(Y) &\le \kappa \underline{g}_n(Y) + (1-\kappa)\overline{g}_n(Y), \\ \underline{g}_{n+1}(Y) &\ge \kappa \overline{g}_n(Y) + (1-\kappa)\underline{g}_n(Y). \end{aligned}$$

For $g_{n+1}(Y,i)$ we have:

$$
\begin{aligned}
&g_{n+1}(Y,i)\\
&= \mathbb{P}[Y_0|Y_1,\dots,Y_{n+1},X_{n+2}=i]\\
&= \frac{\sum_j \mathbb{P}[Y_0,Y_1,\dots,Y_{n+1},X_{n+1}=j,X_{n+2}=i]}{\sum_j \mathbb{P}[Y_1,\dots,Y_{n+1},X_{n+1}=j,X_{n+2}=i]}\\
&= \frac{\sum_j \mathbb{P}[Y_0,X_{n+2}=i|Y_1,\dots,Y_n,X_{n+1}=j]\mathbb{P}[Y_1,\dots,Y_n,X_{n+1}=j]\mathbb{P}[Y_{n+1}|X_{n+1}=j]}{\sum_j \mathbb{P}[Y_1,\dots,Y_{n+1},X_{n+1}=j,X_{n+2}=i]}\\
&= \frac{\sum_j g_n(Y,j)\mathbb{P}[X_{n+2}=i|X_{n+1}=j]\mathbb{P}[Y_1,\dots,Y_n,X_{n+1}=j]\mathbb{P}[Y_{n+1}|X_{n+1}=j]}{\sum_j \qquad \mathbb{P}[X_{n+2}=i|X_{n+1}=j]\mathbb{P}[Y_1,\dots,Y_n,X_{n+1}=j]\mathbb{P}[Y_{n+1}|X_{n+1}=j]}\\
&= \sum_j g_n(Y,j)q(j,i),
\end{aligned}
$$

where

$$
q(j,i) = \frac{\mathbb{P}[X_{n+2}=i|X_{n+1}=j]\mathbb{P}[Y_1,\dots,Y_n,X_{n+1}=j]\mathbb{P}[Y_{n+1}|X_{n+1}=j]}{\sum_{j'} \mathbb{P}[X_{n+2}=i|X_{n+1}=j']\mathbb{P}[Y_1,\dots,Y_n,X_{n+1}=j']\mathbb{P}[Y_{n+1}|X_{n+1}=j']}.
$$

This gives that $g_{n+1}(Y,i)$ is a weighted sum of the $g_n(Y,j)$'s.

Note that

$$
\begin{aligned}
\frac{\mathbb{P}[Y_1,\dots,Y_n,X_{n+1}=j]}{\sum_{j'} \mathbb{P}[Y_1,\dots,Y_n,X_{n+1}=j']} &= \frac{\mathbb{P}[Y_1,\dots,Y_n,X_{n+1}=j]}{\mathbb{P}[Y_1,\dots,Y_n]}\\
&= \mathbb{P}[X_{n+1}=j|Y_1,\dots,Y_n],
\end{aligned}
$$

and $0 < \mathbb{P}[X_{n+1}=j|Y_1,\dots,Y_n] < 1$, by the same reasoning as in Proposition 2.5.

Let

$$
\kappa := \min_{j,i} q(j,i)
$$

then

$$
\begin{aligned}
\kappa \geq & \frac{\min\ \mathbb{P}[X_{n+2}=i|X_{n+1}=j]\min\ \mathbb{P}[Y_{n+1}|X_{n+1}=j]}{\max \mathbb{P}[X_{n+2}=i|X_{n+1}=j]\max \mathbb{P}[Y_{n+1}|X_{n+1}=j]}\\
&\cdot \min \mathbb{P}[X_{n+1}=j|Y_1,\dots,Y_n] =: \kappa'.
\end{aligned}
\tag{3.1.1}
$$

Assuming $P>0, \Pi>0$ we have $\kappa'>0$. As $\kappa \in (0,1)$, either κ or $1-\kappa$ is in $(0,1/2]$. We can assume w.l.o.g. that $\kappa \in (0,1/2]$ (otherwise take $\kappa = 1-\min_{j,i} q(j,i)$).

It follows that:

$$
\begin{aligned}
\overline{g}_{n+1}(Y) &= \max_i g_{n+1}(Y,i)\\
&\leq \kappa \underline{g}_n(Y) + (1-\kappa)\overline{g}_n(Y),
\end{aligned}
$$

and

$$
\begin{aligned}
\underline{g}_{n+1}(Y) &= \min_i g_{n+1}(Y,i)\\
&\geq \kappa \overline{g}_n(Y) + (1-\kappa)\underline{g}_n(Y).
\end{aligned}
$$

Now

$$\begin{aligned}\overline{g}_{n+1}(Y) - \underline{g}_{n+1}(Y) &\leq \Big((1-\kappa)\overline{g}_n(Y) + \kappa\underline{g}_n(Y)\Big) - \Big((1-\kappa)\underline{g}_n(Y) + \kappa\overline{g}_n(Y)\Big) \\ &= (1-2\kappa)(\overline{g}_n(Y) - \underline{g}_n(Y)).\end{aligned}$$

Taking

$$\tilde{\kappa} = 1 - 2\kappa \tag{3.1.2}$$

this gives that for some $0 \leq \tilde{\kappa} < 1$:

$$0 \leq \overline{g}_{n+1}(Y) - \underline{g}_{n+1}(Y) \leq \tilde{\kappa}\left(\overline{g}_n(Y) - \underline{g}_n(Y)\right).$$

Note that as $\kappa \in (0, 1/2]$ we have that $\tilde{\kappa} = 1 - 2\kappa \in [0, 1)$ as desired.

Iterating gives:

$$\begin{aligned}\overline{g}_{n+1}(Y) - \underline{g}_{n+1}(Y) &\leq \tilde{\kappa}\left(\overline{g}_n(Y) - \underline{g}_n(Y)\right) \\ &\leq \tilde{\kappa}^2\left(\overline{g}_{n-1}(Y) - \underline{g}_{n-1}(Y)\right) \\ &\;\;\vdots \\ &\leq \tilde{\kappa}^{n+1}\left(\overline{g}_0(Y) - \underline{g}_0(Y)\right) \\ &\leq \tilde{\kappa}^{n+1},\end{aligned}$$

where the last inequality holds as $0 \leq \overline{g}_0(Y) - \underline{g}_0(Y) \leq 1$.

We have

$$\begin{aligned}g_{n+1}(Y) - g_n(Y) &\leq \overline{g}_{n+1}(Y) - \underline{g}_n(Y) \\ &\leq (1-\kappa)\overline{g}_n(Y) + \kappa\underline{g}_n(Y) - \underline{g}_n(Y) \\ &= (1-\kappa)\left(\overline{g}_n(Y) - \underline{g}_n(Y)\right),\end{aligned}$$

and

$$\begin{aligned}g_{n+1}(Y) - g_n(Y) &\geq \underline{g}_{n+1}(Y) - \overline{g}_n(Y) \\ &\geq (1-\kappa)\underline{g}_n(Y) + \kappa\overline{g}_n(Y) - \overline{g}_n(Y) \\ &= (1-\kappa)\left(\underline{g}_n(Y) - \overline{g}_n(Y)\right).\end{aligned}$$

From this it follows that

$$\begin{aligned}|g_n(Y) - g_{n+1}(Y)| &\leq (1-\kappa)(\overline{g}_n(Y) - \underline{g}_n(Y)) \\ &\leq \tilde{\kappa}^n.\end{aligned}$$

Using telescoping sums and the triangle inequality we have that for all m, n such that $m \geq n$:

$$\begin{aligned}
|g_n(Y) - g_m(Y)| &= \sum_{l=n}^{m-1} |g_l(Y) - g_{l+1}(Y)| \\
&\leq \sum_{l=n}^{m-1} \tilde{\kappa}^l \\
&= \frac{\tilde{\kappa}^n - \tilde{\kappa}^m}{1 - \tilde{\kappa}} \\
&\leq \frac{\tilde{\kappa}^n}{1 - \tilde{\kappa}}.
\end{aligned}$$

This proves the exponential convergence, with $C = \frac{1}{1-\tilde{\kappa}}$ and $\alpha = -\log \tilde{\kappa} > 0$. $\square$

3.2 Harris

Another proof of the convergence of $\mathbb{P}[Y_0 \mid Y_1, \ldots, Y_n]$ can be found in Harris [22]. This proof is based on a technique called coupling, which dates back to Doeblin [10]. This technique will be introduced first, especially couplings for Markov chains. We give two examples of such a coupling, one by Doeblin and one by Vasershtein [36]. Next the notion of a successful coupling is explained, and we will show how this can be used to show the convergence result for hidden Markov models. This leads to three convergence rates, based on the results of Doeblin, Vasershtein and Harris. Finally we will prove that the second one will always give the best result of these three.

3.2.1 Coupling

A *coupling* [27, 35] of two or more random variables X^i, $i = 1, \ldots, n$ is a n-dimensional random variable $\tilde{X} = (\tilde{X}^1, \ldots, \tilde{X}^n)$ such that

$$\tilde{X}^i \stackrel{d}{=} X^i, \quad \forall\, i,$$

where $\stackrel{d}{=}$ denotes equality in distribution. So the X^i are the marginal distributions of $\tilde{X}$, while the joint distribution of $(X^1, \ldots, X^n)$ is in general not the same as the distribution of $\tilde{X}$.

Couplings for Markov chains

We consider a coupling for Markov chains consisting of two running Markov chains, constructed in such a way that from the first moment on they meet, they will coincide. More precisely, let $\{X_n\}_{n\geq 0}$ be a Markov chain with state space $\mathcal{S}$ and with transition probability matrix $P = (p_{ij})$. Let $\{\tilde{X}_n\}_{n\geq 0}$ be a stochastic process, where $\tilde{X}_n = (X_n^1, X_n^2)$. Denote the state of $\tilde{X}_n$ by $\tilde{x}_n = (x_n^1, x_n^2) \in \mathcal{S} \times \mathcal{S}$. So $\tilde{X}$ consists of two copies of the Markov chain X. Let X^1 start in g and X^2 start in h, for some $g, h \in \mathcal{S}$. Now construct the transition probabilities for $\tilde{X}$ in such a way, that from the first time on X^1 and X^2 take on the same value, both will keep taking on the same value. Denote these transition probabilities for $\tilde{X}_n$ by $\tilde{P} = (\tilde{p}_{ij})$. The time when both chains first meet is called the *coupling time* [18], denoted by T:

$$T = \min_{j\geq 0}\{j \mid X_j^1 = X_j^2\}.$$

By definition $X_n^1 = X_n^2$ for $n \geq T$.

As in [18], we introduce two examples of such a coupling $\tilde{X}_n$. The *Doeblin coupling* [10] is given by:

$$\begin{array}{lll} \tilde{X}_n & \tilde{X}_{n+1} & \tilde{p}_{..} \\ (i,i) & (k,k) & p_{ik}, \\ (i,j) & (k,l) & p_{ik}p_{jl}, \end{array} \tag{3.2.1}$$

for $i \neq j$. This was the first known coupling, and it is often referred to as Doeblin's coupling or the classical coupling. For the *Vasershtein coupling* [36] for Markov chains, the transition probabilities are given by:

$$\begin{array}{lll} \tilde{X}_n & \tilde{X}_{n+1} & \tilde{p}_{..} \\ (i,i) & (k,k) & p_{ik}, \\ (i,j) & (k,k) & \min\{p_{ik}, p_{jk}\}, \\ (i,j) & (k,l) & \frac{(p_{ik}-p_{jk})^+(p_{jl}-p_{il})^+}{1-\sum_k \min\{p_{ik},p_{jk}\}}, \end{array} \tag{3.2.2}$$

for $i \neq j, k \neq l$, where $a^+ = \max\{a, 0\}$.

Successful coupling

Denote by $\tilde{P}_{g,h}$ the probability distribution of $\tilde{X}$ with $\tilde{X}_0 = (g, h)$. A coupling is called *successful* if with probability 1 both chains meet in finite time, so if

$$\tilde{P}_{g,h}[T < \infty] = 1, \quad \forall\, (g, h) \in \mathcal{S} \times \mathcal{S}.$$

Proposition 3.2. *If $p_{ij} > 0$ for all i, j, then both Doeblins and Vasershteins coupling as given in* (3.2.1) *respectively* (3.2.2) *are successful.*

Proof. Let $\mathcal{D} = \{(k, k) \mid k \in \mathcal{S}\}$. By definition

$$\min_{j\geq 0}\{\tilde{X}_j \in \mathcal{D}\} = T,$$

and $\tilde{X}_n \in \mathcal{D}$ for $n \geq T$. We have:

$$\begin{aligned} \mathbb{P}[T > m] &= \mathbb{P}[\tilde{X}_m \notin \mathcal{D} \mid \tilde{X}_{m-1} \notin \mathcal{D}] \dots \mathbb{P}[\tilde{X}_1 \notin \mathcal{D} \mid \tilde{X}_0 \notin \mathcal{D}]\mathbb{P}[\tilde{X}_0 \notin \mathcal{D}] \\ &\leq \mathbb{P}[\tilde{X}_m \notin \mathcal{D} \mid \tilde{X}_{m-1} \notin \mathcal{D}] \dots \mathbb{P}[\tilde{X}_1 \notin \mathcal{D} \mid \tilde{X}_0 \notin \mathcal{D}] \\ &= \prod_{n=1}^{m} \mathbb{P}[\tilde{X}_n \notin \mathcal{D} \mid \tilde{X}_{n-1} \notin \mathcal{D}]. \end{aligned}$$

We derive an upper bound for $\mathbb{P}[\tilde{X}_n \notin \mathcal{D} \mid \tilde{X}_{n-1} \notin \mathcal{D}]$. For Doeblins coupling, with $i \neq j$, we have

$$\begin{aligned} \mathbb{P}[\tilde{X}_n = (k,k) \in \mathcal{D} \mid \tilde{X}_{n-1} = (i,j) \notin \mathcal{D}] &= p_{ik}p_{jk}, \\ \mathbb{P}[\tilde{X}_n \in \mathcal{D} \mid \tilde{X}_{n-1} = (i,j) \notin \mathcal{D}] &= \sum_k p_{ik}p_{jk}, \\ \mathbb{P}[\tilde{X}_n \in \mathcal{D} \mid \tilde{X}_{n-1} \notin \mathcal{D}] &\geq \min_{i,j} \sum_k p_{ik}p_{jk}, \\ \mathbb{P}[\tilde{X}_n \notin \mathcal{D} \mid \tilde{X}_{n-1} \notin \mathcal{D}] &\leq 1 - \min_{i,j} \sum_k p_{ik}p_{jk}. \end{aligned} \tag{3.2.3}$$

Writing

$$\lambda_D := \min_{i,j} \sum_k p_{ik} p_{jk}, \tag{3.2.4}$$

this is

$$\mathbb{P}[\tilde{X}_n \notin \mathcal{D} \mid \tilde{X}_{n-1} \notin \mathcal{D}] \leq 1 - \lambda_D.$$

So continuing we have

$$\begin{aligned} \mathbb{P}[T > m] &\leq \prod_{n=1}^{m} \mathbb{P}[\tilde{X}_n \notin \mathcal{D} \mid \tilde{X}_{n-1} \notin \mathcal{D}] \\ &= (1 - \lambda_D)^m. \end{aligned}$$

As all $p_{ij} > 0$ we have $0 \leq (1 - \lambda_D) < 1$ and it follows that $T < \infty$ almost surely. So both chains meet in finite time with probability 1, and hence the coupling is successful.

Along the same lines it can be proved that the Vasershtein coupling is successful. For this coupling (3.2.3) becomes

$$\mathbb{P}[\tilde{X}_n = (k,k) \in \mathcal{D} \mid \tilde{X}_{n-1} = (i,j) \notin \mathcal{D}] = \min\{p_{ik}, p_{jk}\},$$

from which the same result follows, with parameter

$$\lambda_V := \min_{i,j} \sum_k \min\{p_{ik}, p_{jk}\} \tag{3.2.5}$$

instead of λ_D. □

Weak ergodicity

The Markov chain X is called *weakly ergodic* if for all $g, h \in \mathcal{S}$ it holds that

$$\lim_{n\to\infty} \sum_k |p_{gk}^{(n)} - p_{hk}^{(n)}| = 0,$$

where $p_{ij}^{(n)}$ is the n-step transition probability from i to j. This property implies that there is an asymptotic 'loss of memory' for the initial state of the Markov chain.

Proposition 3.3. *If the coupling $\tilde{X}$ is successful, then X is weakly ergodic.*

The proof is given in Appendix A.5.

3.2.2 Harris' result

Harris [22] states that for any Markov chain X with $X_n \in \mathcal{S}$ and transition probabilities $p_{ij} > 0$ for all i, j:

$$\begin{aligned} &|\mathbb{P}[X_{n+1} \in \mathcal{A}_{n+1} \mid X_0 = g, X_1 \in \mathcal{A}_1, \ldots, X_n \in \mathcal{A}_n] \\ &\qquad - \mathbb{P}[X_{n+1} \in \mathcal{A}_{n+1} \mid X_0 = h, X_1 \in \mathcal{A}_1, \ldots, X_n \in \mathcal{A}_n]| \leq (1 - \lambda_H)^n, \end{aligned}$$

where $\lambda_H \in (0,1)$, $\mathcal{A}_i$, $i = 1,\dots,n$ non-empty subsets of $\mathcal{S}$, and $g,h \in \mathcal{S}$ two arbitrary states. Without any proof or explanation, Harris gives the following expression for λ_H:

$$\lambda_H = \min_{i,j,k,l} \frac{p_{kj}p_{il}}{K^2 p_{ij}p_{kl}}, \tag{3.2.6}$$

where $K = |\mathcal{S}|$.

From this statement the convergence of $\mathbb{P}[Y_0 \mid Y_1,\dots,Y_n]$ follows. For this, consider the Markov chain $V = \{V_n\}_{n\geq 0}$ where $V_n = (X_n, Y_n)$, with state space $\mathcal{V} = \mathcal{S}\times\mathcal{S}'$. Here X is the underlying Markov chain, and Y is the hidden Markov process. Now $K = |\mathcal{V}|$. Let $\mathcal{A}_1,\dots,\mathcal{A}_{|\mathcal{S}'|}$ be a partition of $\tilde{\mathcal{S}}$, such that

$$V_n = (X_n, Y_n) \in \mathcal{A}_i \;\Leftrightarrow\; Y_n = i.$$

This gives:

$$\Big|\mathbb{P}[Y_0 = y_0 \mid Y_1 = y_1,\dots,Y_m = y_m, Y_{m+1} = y_{m+1},\dots] \\ - \mathbb{P}[Y_0 = y_0 \mid Y_1 = y_1,\dots,Y_m = y_m, Y_{m+1} = y'_{m+1},\dots]\Big| \leq (1-\lambda_H)^{m-1}. \tag{3.2.7}$$

3.2.3 Proof $\lambda_V \geq \lambda_D \geq \lambda_H$

For the three convergence rates found, it holds that

$$\lambda_V \geq \lambda_D \geq \lambda_H.$$

This means that the Vasershtein coupling gives a faster convergence than the Doeblin coupling, which in turn gives faster convergence than Harris. In other words, $(1-\lambda_V)^n$ goes faster to zero than $(1-\lambda_D)^n$, which goes faster to zero than $(1-\lambda_H)^n$.

The first inequality is straightforward. As $0 < p_{ij} < 1$ we have

$$\min\{p_{ik}, p_{jk}\} \geq p_{ik}p_{jk},$$

and it directly follows that $\lambda_V \geq \lambda_D$:

$$\lambda_V = \min_{i,j}\sum_k \min\{p_{ik}, p_{jk}\} \geq \min_{i,j}\sum_k p_{ik}p_{jk} = \lambda_D.$$

The next proposition states that $\lambda_D \geq \lambda_H$, from which it follows that

$$\lambda_D = \min_{i,j}\sum_k p_{ik}p_{jk} \geq \min_{i,j,m,n}\frac{p_{im}p_{jn}}{K^2 p_{jm}p_{in}} = \lambda_H.$$

Proposition 3.4. *For any $K\times K$ stochastic matrix $P = (p_{ij}) > 0$, we have for all i,j:*

$$\sum_k p_{ik}p_{jk} \geq \frac{1}{K^2}\min_{m,n}\frac{p_{im}p_{jn}}{p_{jm}p_{in}}.$$

The proof of this is given in Appendix A.6.

3.3 Birch

Birch [5, 6] gives a similar result similar to that of Harris, see (3.2.7), for the convergence of $\mathbb{P}[Y_0 \mid Y_1, \ldots, Y_n]$. Instead of Harris' value for λ_H of (3.2.6), Birch gives the value, say λ_B. So Birch states that for a general hidden Markov process Y, when the transition probability matrix of the process (X_n, Y_n) is strictly positive, it holds that

$$\Big| \mathbb{P}[Y_0 = y_0 \mid Y_1 = y_1, \ldots, Y_m = y_m, Y_{m+1} = y_{m+1}, \ldots] \\ - \mathbb{P}[Y_0 = y_0 \mid Y_1 = y_1, \ldots, Y_m = y_m, Y_{m+1} = y'_{m+1}, \ldots] \Big| \leq (1 - \lambda_B)^{m-1}.$$

For this again the definition for Y as a grouped Markov chain is used, as was the case for Harris' result. We consider the Markov chain $V = \{V_n\}_{n \geq 0}$ where $V_n = (X_n, Y_n)$. Let $\mathcal{A}_1, \ldots, \mathcal{A}_{|\mathcal{S}'|}$ be a partition of the state space $\mathcal{V}$ of V, such that

$$V_n = (X_n, Y_n) \in \mathcal{A}_i \;\Leftrightarrow\; Y_n = i.$$

Define

$$K_{min} := \min_i |\mathcal{A}_i|, \qquad K_{max} := \max_i |\mathcal{A}_i|.$$

We have $1 \leq K_{min} \leq K_{max} \leq K = |\mathcal{V}|$.

The expression for λ_B is given by:

$$\lambda_B = \min_{i,j,k,l,m} \frac{K_{min}}{K_{max}^2} \left(\frac{p_{il} p_{lk}}{p_{ij} p_{jm}} \right)^2. \tag{3.3.1}$$

3.4 Fernández, Ferrari, Galves

In this section we state the claim of [15] for the upper bound on the convergence rate. It is based on regeneration times of the underlying Markov chain. First we introduce Countable Mixtures of Markov Chains. Next we introduce regeneration times and show that the regeneration times for a Markov chain have a geometric distribution. Then we give the main result of this section: the upper bound for the convergence rate based on regeneration times.

3.4.1 CMMC

A *Countable Mixtures of Markov Chain* (CMMC) [15] is a process whose transition probabilities are a countable convex combination of Markov transitions of increasing order. Denote by x_j^i the vector $(x_i, \ldots, x_j)$, then the general form of a CMMC is given by

$$\mathbb{P}[a \mid \underline{x}] = \lambda_0 \mathbb{P}^{(0)}[a] + \sum_{k=1}^{\infty} \lambda_k \mathbb{P}^{(k)}[a \mid x_{-1}^{-k}],$$

where $\lambda_k \geq 0$, $\sum_{k=0}^{\infty} \lambda_k = 1$, each $\mathbb{P}^{(k)}[a \mid x_{-1}^{-k}]$ is a Markov transition of order k for $k \geq 1$ and $\mathbb{P}^{(0)}$ is a probability measure. By a Markov transition of order k we mean that the transition probabilities for the next state depend only on the last k states. A hidden Markov model can be written in this form, as can a Markov chain. For this last one, we have $\lambda_k = 0$ for $k \geq 2$, as a state depends only on the previous state. There is not necessarily a unique representation in this form for a Markov chain. We will exploit this in Section 3.4.3.

The *regeneration time* [15] for the window $(X_l, \dots, X_m)$ is given by

$$\tau[l, m] := \max\left\{t \leq l \mid t \leq n - L_n, \text{ for all } n \in [t, m]\right\},$$

with the convention $\tau[l, m] = -\infty$ if the set in the right-hand side is empty. Here $L_n, n \in \mathbb{Z}$, called *random orders*, give on how many states back the transition probabilities for the state X_n depend. Write $\tau[l] := \tau[l, l]$. We have that, when $\tau[l]$ is a regeneration time, then $\{X_{n+\tau[l]}\}_{n\geq 0}$ and $\{X_{\tau[l]-n}\}_{n\geq 0}$ are independent.

3.4.2 Geometric distribution

In the next lemma we consider a CMMC whose transition probabilities either depend only on the previous state or do not depend on the past at all.

Lemma 3.5. *For a CMMC defined by*

$$\mathbb{P}[a \mid \underline{x}] = \lambda_0 \mathbb{P}^{(0)}[a] + \lambda_1 \mathbb{P}^{(1)}[a \mid x_{-1}], \tag{3.4.1}$$

with $\lambda_0 + \lambda_1 = 1$, it holds that, for any $l \in \mathbb{Z}$, $\tau[l]$ has a geometric distribution with parameter λ_0.

Proof. Noting that in this case $L_n \in \{0, 1\}$, we have

$$\tau[l] = \max\{n \leq l \mid L_n = 0\}. \tag{3.4.2}$$

We have

$$\begin{aligned} L_n = 0 &\Leftrightarrow \{0 \leq U_n \leq \lambda_0\}, \\ L_n = 1 &\Leftrightarrow \{\lambda_0 \leq U_n \leq 1\}, \end{aligned}$$

with (U_n) a sequence of i.i.d. random variables uniformly distributed on $[0, 1]$. If $U_n < \lambda_0$ then the next state depends not on the past at all, otherwise it depends on the previous state. It follows that L_n is a Bernoulli distributed random variable: With probability λ_0 it is equal to 0 and it is equal to 1 otherwise. We can interpret l as the number of times $L_n = 1$ occurs before the first time $L_n = 0$ occurs. This directly gives that $\tau[l]$ has a geometric distribution, with parameter $1 - \lambda_1 = \lambda_0$. □

In general (3.4.2) does not hold for a CMMC, as for this $L_n \in \{0, 1, \dots\}$.

3.4.3 Result Fernández, Ferrari, Galves

We now give the claim made by [15].

Claim 3.6. *Denote by τ_X the regeneration time for X_0. Then it holds that*

$$\sup_{Y, \tilde{Y}} \left| \mathbb{P}[Y_0 \mid Y_{-\infty}^{-1}] - \mathbb{P}[Y_0 \mid Y_s^{-1} \tilde{Y}_{-\infty}^{s-1}] \right| \leq \mathbb{P}[\tau_X < s],$$

for every Y_0 and $s \leq 0$.

As the underlying Markov chain can be written as in (3.4.1), it follows that τ_X has a geometric distribution, so

$$\mathbb{P}[\tau_X < s] = (1-\lambda_0)^{-s}.$$

3.5 Comparison convergence rates

In the previous sections we found a number of different upper bounds for the convergence rate

$$\gamma(n) := \sup_{Y,Y'} \left| \mathbb{P}[Y_0 \mid Y_1, \dots, Y_n, Y_{n+1} \dots] - \mathbb{P}[Y_0 \mid Y_1, \dots, Y_n, Y'_{n+1} \dots \right|.$$

Now will compare these for the binary symmetric hidden Markov model, see Section 2.1.4 and Section 2.1.5. Throughout this section we assume $0 < \delta \leq 0.5$ and $0 < p \leq 0.5$. Note that the result of Baum and Petrie is based on the definition of hidden Markov model as given 2.1.4, where all others make use of the definition as a Markov source, see Section 2.1.5. For this we consider the so-called extended Markov chain V with $V_n = (X_n, Z_n)$. The hidden Markov model Y follows from this by the function f which gives

$$V_n = f(X_n, Z_n) = X_n \cdot Z_n.$$

3.5.1 Convergence rates

Baum and Petrie

For Baum and Petrie's approach we consider the hidden Markov model which was defined as in Section 2.1.4. So $X_n \in \{-1, 1\}$ with transition probability matrix P, and $Y_n \in \{-1, 1\}$ with emission probability matrix Π. From Proposition 2.6 it follows that

$$\gamma(n) \leq (1-\tilde{\kappa})^n e^{n \log \tilde{\kappa}},$$

where is in (3.1.2) $\tilde{\kappa} = 1 - 2\kappa$. From (3.1.1) we have that κ' is a lower bound for κ, given by:

$$\begin{aligned}\kappa' &= \frac{\min\, \mathbb{P}[X_{n+2} = i | X_{n+1} = j] \min\, \mathbb{P}[Y_{n+1} | X_{n+1} = j]}{\max \mathbb{P}[X_{n+2} = i | X_{n+1} = j] \max \mathbb{P}[Y_{n+1} | X_{n+1} = j]} \\ &\qquad \cdot \min \mathbb{P}[X_{n+1} = j | Y_1, \dots, Y_n] \\ &= \frac{p^2\, \delta}{(1-p)(1-\delta)}.\end{aligned}$$

Harris

For the Markov chain V we have from (3.2.6) and (3.2.7):

$$\gamma(n) \leq (1-\lambda_H)^n, \qquad \lambda_H = \min_{i,j,k,l} \frac{p_{kj} p_{il}}{K^2 p_{ij} p_{kl}}.$$

Here $K = 4$ and the transition probabilities are the elements of matrix Δ as given in (2.1.1), so

$$\lambda_H = \frac{p^2}{16(1-p)^2}.$$

Birch

For the Markov chain V we have from (3.3.1):

$$\gamma(n) \leq (1-\lambda_B)^n, \qquad \lambda_B = \min_{i,j,k,l,m} \frac{K_{min}}{K_{max}^2} \left(\frac{p_{il}p_{lk}}{p_{ij}p_{jm}}\right)^2.$$

Here $K_{min} = K_{max} = 2$ and the transition probabilities are elements from Δ, so we get

$$\lambda_B = \frac{p^4\delta^4}{2\,(1-p)^4(1-\delta)^4}.$$

Fernández, Ferrari, Galves

Consider the extended Markov chain V. Recall that the representation of a Markov chain as CMMC is not unique, see Section 3.4.1. This enables us to write V as in (3.4.1) in the following way:

$$\begin{aligned}\mathbb{P}[(X_{n+1}, Z_{n+1}) \mid (X_n, Z_n)] = 2p\,\mathbb{P}^{(0)}[X_{n+1}]\,\mathbb{P}^{(0)}[Z_{n+1}] \\ + (1-2p)\,\mathbb{P}^{(1)}[X_{n+1} \mid X_n]\mathbb{P}^{(1)}[Z_{n+1} \mid Z_n],\end{aligned}$$

where

$$\mathbb{P}^{(1)}[X_{n+1} \mid X_n] = \begin{cases} 1 & \text{if } X_{n+1} = X_n, \\ 0 & \text{if } X_{n+1} \neq X_n. \end{cases}$$

This can be easily checked to be correct by plugging in all possibilities of 1's and -1's. By symmetry we have $\mathbb{P}^{(0)}[X_{n+1}] = \frac{1}{2}$ for $X_{n+1} \in \{-1, 1\}$, and $\mathbb{P}^{(1)}[Z_{n+1} \mid Z_n] = \mathbb{P}^{(0)}[Z_{n+1}]$ by independence of the Z_i, and

$$\mathbb{P}^{(0)}[Z_{n+1}] = \begin{cases} 1-\delta & \text{if } Z_{n+1} = 1, \\ \delta & \text{if } Z_{n+1} = -1. \end{cases}$$

This representation gives that $\lambda_0 = 2p =: \lambda_F$, so the distribution of τ_V is given by

$$\mathbb{P}[\tau_V > n] = (1-2p)^n.$$

According to Claim 3.6, it follows that for the binary symmetric model

$$\gamma(n) \leq (1-2p)^n,$$

for all $n \geq 0$.

Doeblin's and Vasershtein's coupling

For comparison we calculate the value of λ_D for the extended Markov chain V with transition probabilities Δ. From (3.2.4) it follows that

$$\lambda_D = \min_{i,j} \sum_k p_{ik}p_{jk} = 2\,p\,(1-p)\,\left((1-\delta)^2 + \delta^2\right).$$

From (3.2.5) we have

$$\lambda_V = \min_{i,j} \sum_k \min\{p_{ik}, p_{jk}\} = 2p.$$

Recall that these results give a bound on the convergence rate of a coupling of two Markov chains:

$$|\mathbb{P}[X_0 \mid X_n = g] - \mathbb{P}[X_0 \mid X_n = h]\,| \leq (1 - \lambda_{D,H})^n.$$

The results of Baum and Petrie, Harris, Birch and Fernández, Ferrari and Galves hold for the convergence rate of a hidden Markov model.

3.5.2 Comparison

We compare the upper bounds for $\gamma(n)$ for the values $p = 0.4$ and $\delta = 0.1$. The upper bounds of Birch, Harris and Fernández, Ferrari and Galves (FFG) are given by $(1 - \lambda_*)^n$. For these values of p and δ the expressions for λ_* become:

Birch	λ_B	$\approx$	$1.505 \cdot 10^{-5}$,	Doeblin	λ_D	$=$	0.3936,
Harris	λ_H	$\approx$	$2.777 \cdot 10^{-2}$,	Vasershtein	λ_V	$=$	0.8,
FFG	λ_F	$=$	0.8.				

The values of λ_D and λ_V are added for comparison as they give a result for the convergence rate of a Markov chain, where the other three hold for hidden Markov models.

For Baum and Petrie's result $p = 0.4$ and $\delta = 0.1$ give $\kappa' = \frac{4}{135}$ and

$$\gamma(n) \leq \frac{(1 - 2\kappa')^n}{2\kappa'} \approx 16.9 \cdot (1 - 0.06)^n.$$

For small n this upper bound is large, but as n grows this will drop rather fast to zero. Note that it becomes smaller than 1 only for $n \geq 47$.

The plot of the convergence rates plotted against (continuous) n is given in Figure 3.1.

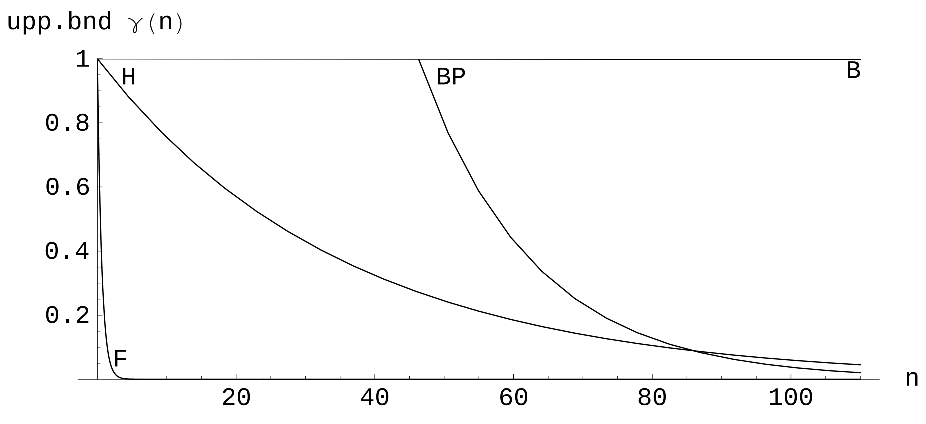


Figure 3.1: Upper bounds for convergence speed $\gamma(n)$ plotted against (continuous) n, for Birch (B), Harris (H), FFG (F) and Baum and Petrie (BP).

Chapter 4

Series expansions

In this chapter we investigate series expansions for the entropy. We start with the remarkable result of Zuk et al. [39], who give a so-called 'stabilization' for the coefficients in these expansions. We will show that the same idea is applicable for the conditional probabilities $\mathbb{P}[Y_0 \mid Y_1, \ldots, Y_n]$, and we will give a proof of this. The idea of the 'stabilization' will be illustrated by a simple example. Next we discuss the results of Han and Marcus [19, 20], concerning the analyticity of the series expansions for the entropy. These results all hold for general hidden Markov models. We then focus on the binary symmetric model. For this we derive several series expansions in various parameters. The aim is to find a general expression for the coefficients that appear in these, but this turns out to be quite challenging.

4.1 Settlement coefficients

4.1.1 Result Zuk et al.

In [38, 39, 40, 41, 42] Zuk et al. study series expansions for the entropy of general as well as binary symmetric hidden Markov models. In [39] they give the following result:

Theorem 4.1. *Let Y be a general hidden Markov model. Given the series expansions*

$$h(Y_0, Y_1, \ldots, Y_n) = \sum_{k=0}^{\infty} C_k^{(n)} \delta^k$$

and

$$h(Y_0, Y_1, Y_2, \ldots) = \sum_{k=0}^{\infty} C^k \delta^k,$$

where $C_k^{(n)}$ and C_k are functions of P. Then

$$n \geq \left\lceil \frac{k+1}{2} \right\rceil \;\Rightarrow\; C_k^{(n)} = C_k.$$

We say that the coefficients 'settle' or 'stabilize'. The term $\lceil \frac{k+1}{2} \rceil$ does not depend on the alphabet size. Note that we use a different indexing than in [39]. In that indexing the result holds for $n \geq \lceil \frac{k+3}{2} \rceil$. The statement is proved for an arbitrary hidden Markov model Y. We will give a short outline of the proof in Section 4.1.3.

For the binary symmetric model, the coefficients C_k are given up to eleventh order in [42]:

$$\begin{aligned}
C_0 &= -p \log p - (1-p)\log(1-p), \\
C_1 &= 2(1-2p)\log\left(\frac{1-p}{p}\right), \\
C_2 &= -2(1-2p)\log\left(\frac{1-p}{p}\right) - \frac{(1-2p)^2}{2p^2(1-p)^2}, \\
C_3 &= -16\,(5\lambda^4 - 10\lambda^2 - 3)\lambda^2 /\, 3(1-\lambda^2)^4, \\
&\vdots \\
C_{11} &= 8192\,(98142\lambda^{30} - 1899975\lambda^{28} + 92425520\lambda^{26} + 3095961215\lambda^{24} \\
&\quad + 25070557898\lambda^{22} + 59810870313\lambda^{20} - 11635283900\lambda^{18} \\
&\quad - 173686662185\lambda^{16} - 120533821070\lambda^{14} + 74948247123\lambda^{12} \\
&\quad + 102982107048\lambda^{10} + 35567469125\lambda^{8} + 4673872550\lambda^{6} \\
&\quad + 217466315\lambda^{4} + 2569380\lambda^{2} + 2277)\lambda^6 /\, 495(1-\lambda^2)^{20},
\end{aligned}$$

where we abbreviate $\lambda = 1 - 2p$. These coefficients are found making use of a one-dimensional random-field Ising model representation [29]. Note that the first three terms involve the log-function, whereas higher terms are rational functions of λ. According to [42] the zeroth and first-order terms were already known in [23, 30], while the second and higher-order terms were not know before.

4.1.2 Settlement coefficients conditional probability

An analogous result of Theorem 4.1 hold for the series expansion of the conditional probability $\mathbb{P}[Y_0 \mid Y_1, \ldots, Y_n]$. In this expansion the coefficients also settle, but only for $n \geq k+1$, as stated in the next theorem.

Theorem 4.2. *Given the series expansions*

$$\mathbb{P}[Y_0 \mid Y_1, \ldots, Y_n] = \sum_{k=0}^{\infty} F_k^{(n)} \delta^k$$

and

$$\mathbb{P}[Y_0 \mid Y_1, Y_2, \ldots] = \sum_{k=0}^{\infty} F_k \delta^k,$$

where $F_k^{(n)}$ *and* F_k *are functions of* P *and* y. *Assume* $P > 0$ *and* $\Pi > 0$. *Then*

$$n \geq k+1 \Rightarrow F_k^{(n)} = F_k.$$

As a result, the settled coefficients F_k only depend on $y_0, \dots, y_{k+1}$. The proof of this theorem will be exactly along the lines of the proof in [39] for the entropy. We give three lemmas which combined together prove the theorem.

First we introduce a more general process $W = \{W_n\}_{n \geq 0}$ with $W_n \in \{-1, 1\}$. For this we let the probability of an erroneous observation of X_i depend on i. So let $\mathbb{P}[Z'_i = 1] = 1 - \delta_i$, and let $W_n = X_n \cdot Z'_n$. Setting all δ_i's equal will give the original process Y.

Denote a vector by

$$v_0^n = (v_0, v_1, \dots, v_n).$$

We abbreviate $\mathbb{P}[X]$ for $\mathbb{P}[X = x]$. Define

$$G_n = G_n(\delta_0^n, w_0^n) = \mathbb{P}[W_0 \mid W_1^n].$$

Note that $G_n((\delta, \delta, \dots, \delta), y_0^n) = \mathbb{P}[Y_0 \mid Y_1^n]$.

Lemma 4.3. *For all $0 < j \leq n$ it holds that if $\delta_j = 0$ then $G_n = G_j$.*

Proof. If $\delta_j = 0$ then W_j is equal to the underlying Markov chain, i.e. $W_j = X_j$. This gives that $\mathbb{P}[W_0^{j-1} \mid W_j^n] = \mathbb{P}[W_0^{j-1} \mid W_j]$, as conditioning on W_{j+1}^n will give no extra information in this case. So let $\delta_j = 0$, then

$$\begin{aligned} G_n &= \mathbb{P}[W_0 \mid W_1^n] \\ &= \frac{\mathbb{P}[W_0^n]}{\mathbb{P}[W_1^n]} \\ &= \frac{\mathbb{P}[W_0^{j-1} \mid W_j]}{\mathbb{P}[W_1^{j-1} \mid W_j]} \\ &= \mathbb{P}[W_0 \mid W_1^j] = G_j. \end{aligned}$$

□

Let $\vec{k} = k_0^n$, where $k_i \in \mathbb{N} \cup \{0\}$. Define the weights of $\vec{k}$ as $w(\vec{k}) = \sum_{i=0}^n k_i$, and define

$$G_n^{\vec{k}} = \left. \frac{\partial^{w(\vec{k})} G_n}{\partial \delta_0^{k_0} \dots \partial \delta_n^{k_n}} \right|_{\vec{\delta}=0}.$$

Lemma 4.4. *Let $\vec{k}^{(c)} = (k_0, k_1, \dots, k_{n-1}, k_n = 0, \underbrace{0, \dots, 0}_{c})$, then, for all $c \in \mathbb{N}$,*

$$G_{n+c}^{\vec{k}^{(c)}} = G_n^{\vec{k}}.$$

Proof. Note that we do not differentiate with respect to δ_n, so setting $\delta_n = 0$ implies, by Lemma 4.3, that $G_{n+c} = G_n$. This gives

$$\begin{aligned} G_{n+c}^{\vec{k}^{(c)}} &= \left. \frac{\partial^{w(\vec{k}^{(c)})} G_{n+c}}{\partial \delta_0^{k_0} \dots \partial \delta_{n-1}^{k_{n-1}}} \right|_{\vec{\delta}=0} \\ &= \left. \frac{\partial^{w(\vec{k})} G_n}{\partial \delta_0^{k_0} \dots \partial \delta_{n-1}^{k_{n-1}}} \right|_{\vec{\delta}=0} = G_n^{\vec{k}}. \end{aligned}$$

□

As $F_k^{(n)}$ is the kth coefficient of the series expansion of $\mathbb{P}[Y_0 \mid Y_1, \ldots, Y_n]$ it can be written as

$$F_k^{(n)} = \frac{1}{k!} \frac{\partial^k}{\partial \delta^k} \mathbb{P}[Y_0 \mid Y_1, \ldots, Y_n]\bigg|_{\delta=0}.$$

Let the vector $\vec{k}$ give the number of times we differentiate to each of the δ_i, $i = 0, \ldots, n$ in the process W, then we can write

$$F_k^{(n)} = \frac{1}{k!} \sum_{\vec{k}\colon w(\vec{k})=k} G_n^{\vec{k}}.$$

Many terms $G_n^{\vec{k}}$ in this sum equal zero, as the next lemma shows.

Lemma 4.5. *If there exists i, j with $0 < j < i \leq n$ for which $k_i > k_j = 0$ then $G_n^{\vec{k}} = 0$.*

Proof. From $k_j = 0$ it follows that $\delta_j = 0$ and so, again using Lemma 4.3, we get $G_n = G_j$. This gives

$$\begin{aligned} G_n^{\vec{k}} &= \frac{\partial^{w(\vec{k})} G_n}{\partial \delta_0^{k_0} \ldots \partial \delta_n^{k_n}}\bigg|_{\vec{\delta}=0} \\ &= \frac{\partial^{w(\vec{k})} G_j}{\partial \delta_0^{k_0} \ldots \partial \delta_n^{k_n}}\bigg|_{\vec{\delta}=0} \\ &= \frac{\partial^{w(\vec{k})-1}}{\partial \delta_0^{k_0} \ldots\ \partial \delta_i^{k_i - 1} \ldots \partial \delta_n^{k_n}} \left(\frac{\partial G_j}{\partial \delta_i} \right)\bigg|_{\vec{\delta}=0} = 0. \end{aligned}$$

The last equality holds as G_j does not depend on δ_i. Note that the one but last step is possible as $k_i \geq 1$. ☐

Now we give the proof of Theorem 4.2:

Proof of Theorem 4.2. Let $\vec{k} = k_0^n$ with $k = w(\vec{k})$. Define the length of $\vec{k}$ as $l(\vec{k}) = \max\{i \mid k_i > 0\}$. Then Lemma 4.5 gives

$$G_n^{\vec{k}} \neq 0 \Rightarrow l(\vec{k}) \leq k,$$

as the maximum length is achieved when $\vec{k} = (0, \underbrace{1, \ldots, 1}_{k}, 0, \ldots, 0)$.

From Lemma 4.4 it follows that for all $\vec{k}$ with $l(\vec{k}) \leq k$

$$G_n^{\vec{k}} = G_{k+1}^{(k_0, \ldots, k_{k+1})},$$

and so

$$G_n^{(k)} = G_{k+1}^{(k)}, \quad \forall\, n \geq k+1.$$

Assuming analyticity of $\mathbb{P}[Y_0 \mid Y_1, Y_2, \ldots]$ and $F^{(n)}$ around $\delta = 0$, we have $\lim_{n \to \infty} F_k^{(n)} = F_k$ and therefore

$$F_k^{(n)} = F_k, \quad \forall\, n \geq k+1.$$

☐

4.1.3 Settlement coefficients entropy

We now give the outline of the proof of Theorem 4.1, as given in [39]. It is along the same lines as the proof of Theorem 4.2. Define

$$\tilde{G}_n = \tilde{G}_n(\delta_0^n) = h(Y_0 \mid Y_1, \ldots, Y_n),$$

then $\tilde{G}_n((\delta, \delta, \ldots, \delta)) = C_n$. The following lemmas can easily be proved along the same lines of the corresponding lemmas in Section 4.1.2. For detailed proofs we refer to [39].

Lemma 4.6. *For all $0 < j \leq n$, if $\delta_j = 0$ then $\tilde{G}_n = \tilde{G}_j$.*

Lemma 4.7. *If there exists i, j with $0 < j < i \leq n$, for which $k_i \geq 1, k_j \leq 1$ then $\tilde{G}_n^{\vec{k}} = 0$.*

Lemma 4.8. *For $\vec{k}^{(c)}$ with $k_n \leq 1$ and for all $c \in \mathbb{N}$: $\tilde{G}_{n+c}^{\vec{k}^{(c)}} = \tilde{G}_n^{\vec{k}}$.*

Note that the conditions in Lemma 4.7 slightly differ from those in Lemma 4.5. From this the different bound follows:

$$\tilde{G}_n^{\vec{k}} \neq 0 \Rightarrow l(\vec{k}) \leq \left\lceil \frac{k+1}{2} \right\rceil,$$

which gives the corresponding result for the settlement of the coefficients $C_k^{(n)}$.

4.1.4 Example settlement coefficients

We will illustrate the settlement of the coefficients in the series expansion of the conditional probability by means of a simple example. Consider the binary symmetric model. We calculate the conditional probabilities for the all-one vector y, and express this as a power series in δ around $\delta = 0$. So we derive

$$\mathbb{P}[Y_0 = 1 \mid Y_1 = 1, \ldots, Y_n = 1] = \sum_{k=0}^{\infty} F_k^{(n)}(p\,; 1, \ldots, 1)\, \delta^k,$$

for $n \geq 0$. For details on the calculation of this, see Section 4.3. The coefficients $F_k^{(n)}(p\,; 1, \ldots, 1)$ are now given by:

n	$F_0^{(n)}$	$F_1^{(n)}$	$F_2^{(n)}$	$F_3^{(n)}$	$F_4^{(n)}$
0	$1/2$				
1	p	$2(1-2p)$	$-2(1-2p)$		
2	p	$\frac{1-2p}{p}$	$\frac{(1-2p)(3p-2)}{p^2}$	$\frac{-4(p-1)(1-2p)^2}{p^3}$	$\frac{-2(1-2p)^2(5p^2-10p+4)}{p^4}$
3	p	$\frac{1-2p}{p}$	$\frac{-(1-p)^2(1-2p)}{p^3}$	$\frac{-(1-2p)^2(2p^2-1)}{p^5}$	$\frac{-(1-2p)^2(5p^4-5p^2+1)}{p^7}$
4	p	$\frac{1-2p}{p}$	$\frac{-(1-p)^2(1-2p)}{p^3}$	$\frac{2(1-p)^2(1-2p)^2}{p^5}$	$\frac{-(1-2p)^2(p^4-4p^3+14p^2-14p+4)}{p^7}$
5	p	$\frac{1-2p}{p}$	$\frac{-(1-p)^2(1-2p)}{p^3}$	$\frac{2(1-p)^2(1-2p)^2}{p^5}$	$\frac{-(1-p)^2(1-2p)^2(p^2-10p+5)}{p^7}$
6	p	$\frac{1-2p}{p}$	$\frac{-(1-p)^2(1-2p)}{p^3}$	$\frac{2(1-p)^2(1-2p)^2}{p^5}$	$\frac{-(1-p)^2(1-2p)^2(p^2-10p+5)}{p^7}$

Here we fixed the values of the y_i. In general the coefficients of the series expansion will depend on these. Because of the settlement the coefficient $F_k = F_k(p\,; y_0, \ldots, y_{k+1})$.

4.2 Analyticity of series expansions

In this section we give the results of Han and Marcus [19, 20] concerning the analyticity of the series expansions for $h(Y)$.

Consider the Markov chain $\{V_n\}_{n\geq 0}$ with state space $\mathcal{V}$ and transition probability matrix Δ. Let $\{Y_n\}_{n\geq 0}$ be a hidden Markov model with state space $\mathcal{Y}$, defined by $Y_n = \Phi(V_n)$ for some function $\Phi : \mathcal{V} \mapsto \mathcal{Y}$. Then the main result is given by:

Theorem 4.9 ([20, Theorem 1.1])**.** *Suppose that the entries of Δ are analytically parameterized by a real variable vector $\vec{\varepsilon}$. If at $\vec{\varepsilon} = \vec{\varepsilon}_0$,*

i for all $y \in \mathcal{Y}$, there is at least one j with $\Phi(j) = y$ such that the j-th column of Δ is strictly positive, and

ii every column of Δ is either all zero or strictly positive,

then $h(Y)$ is a real analytic function of $\vec{\varepsilon}$ at $\vec{\varepsilon}_0$.

If all entries of Δ are strictly positive, both conditions are met.

Real analyticity of a function at a certain point implies that it can be expanded as a convergent power series in a neighborhood of the point. A derivation is given to determine a complex neighborhood of $\vec{\varepsilon}_0$ where the function is analytic. There is no complete set of necessary and sufficient conditions on Δ and Φ known to [20] that guarantees analyticity of the entropy $h(Y)$. Only for a very special case of hidden Markov models these conditions are given, when there exists a y such that $\Phi^{-1}(y)$ contains exactly one element.

For the binary symmetric model, we have $V_n = (X_n, Y_n)$ with transition probability matrix Δ as given in (2.1.1). We have that $h(Y)$ is analytical as a function of δ and p, when both are in $(0, 1)$. But, by Theorem 4.9, this constraint can be relaxed. Also for $\delta = 0$ and $p \in (0, 1)$ analyticity of $h(Y)$ still holds, as it can be easily checked that in this case both conditions hold.

For the binary symmetric case a system of inequalities is given in [20], which involves an r such that the entropy is an analytic function of δ for $|\delta| < r$. From this a lower bound on the convergence radius is estimated for given values of p. We will discuss another idea to derive the convergence radius, for an expansion of the entropy in $\delta(1-\delta)$, in Section 5.7.

4.3 Series expansion in δ

From now on we will focus on the binary symmetric hidden Markov model. We derive a series expansion in δ around $\delta = 0$ for the conditional probabilities $\mathbb{P}[Y_0 \mid Y_1, \ldots, Y_n]$. For this, we first observe that we can write

$$\mathbb{P}[Y_0, Y_1, \ldots, Y_n] = \sum_{k=0}^{n+1} f_k^{(n)} \delta^k.$$

The coefficients $f_k^{(n)} = f_k^{(n)}(p\,; y_0, \ldots, y_n)$ are found by differentiation of this probability on the left-hand side:

$$\frac{1}{k!} \frac{\partial^k}{\partial \delta^k} \mathbb{P}[Y_0 = y_0, \ldots, Y_n = y_n] \bigg|_{\delta=0}.$$

Using this expansion we derive an expansion for the conditional probabilities

$$\mathbb{P}[Y_0 \mid Y_1, \ldots, Y_n] = \sum_{k=0}^{\infty} F_k^{(n)} \delta^k.$$

We give the expressions for the first few terms, from that we show the earlier discussed settlement of the coefficients for the first two: $F_0^{(n)} = F_0$ and $F_1^{(n)} = F_1$ for $n \geq k+1$. We note some structure in the coefficients, but this turns out not to be sufficient to find a general expression for them.

4.3.1 Series expansion $\mathbb{P}[Y_0, Y_1, \ldots, Y_n]$

As mentioned before, the probability $\mathbb{P}[Y_0 = y_0, \ldots, Y_n = y_n]$ is a polynomial in δ of degree $n+1$ and can therefore be written as

$$\mathbb{P}[Y_0 = y_0, \ldots, Y_n = y_n] = \sum_{k=0}^{n+1} f_k^{(n)} \delta^k,$$

where

$$\begin{aligned} f_k^{(n)} &= f_k^{(n)}(p\,; y_0, \ldots, y_n) \\ &= \frac{1}{k!} \frac{\partial^k}{\partial \delta^k} \mathbb{P}[Y_0 = y_0, \ldots, Y_n = y_n] \bigg|_{\delta=0}. \end{aligned} \tag{4.3.1}$$

We write $\mathbb{P}[Y_0 = y_0, \ldots, Y_n = y_n]$ in such a way that we can easily take the derivative of it with respect to δ. For this, we will condition on the number of y_i's that are 'flipped' with respect to the underlying Markov chain x_i. Such a flip will occur when $z_i = -1$, see Section 2.1.4. As the z_i are i.i.d. distributed Bernoulli random variables, the probability of a certain number of flips, say l,is binomially distributed with success probability δ. So for all $z_0, \ldots, z_n$, having exactly l flips:

$$\mathbb{P}[Z_0 = z_0, \ldots, Z_n = z_n] = \delta^l \, (1-\delta)^{n-l+1}.$$

Denoting by $\#\{i : z_i = -1\}$ the number of flips, we can now write

$$\begin{aligned} &\mathbb{P}[Y_0 = y_0, \ldots, Y_n = y_n] \\ &= \sum_{l=0}^{n+1} \sum_{\#\{i:z_i=-1\}=l} \mathbb{P}[Y_0 = y_0, \ldots, Y_n = y_n \mid Z_0 = z_0, \ldots, Z_n = z_n] \, \mathbb{P}[Z_0 = z_0, \ldots, Z_n = z_n] \\ &= \sum_{l=0}^{n+1} \delta^l (1-\delta)^{n-l+1} \sum_{\#\{i:z_i=-1\}=l} \mathbb{P}[Y_0 = y_0, \ldots, Y_n = y_n \mid Z_0 = z_0, \ldots, Z_n = z_n] \\ &= \sum_{l=0}^{n+1} \delta^l (1-\delta)^{n-l+1} \sum_{\#\{i:z_i=-1\}=l} \mathbb{P}[X_0 = y_0 z_0, \ldots, X_n = y_n z_n] \\ &= \sum_{l=0}^{n+1} \delta^l (1-\delta)^{n-l+1} c_l(p\,; n\,; y_0, \ldots, y_n), \end{aligned} \tag{4.3.2}$$

where

$$c_l(p\,;n\,;y_0,\ldots,y_n) := \sum_{\#\{i:z_i=-1\}=l} \mathbb{P}[X_0 = y_0 z_0, \ldots, X_n = y_n z_n]. \tag{4.3.3}$$

This is the probability of observing the sequence $y_0, \ldots, y_n$ when exactly l bits of it are flipped. The sum is over $\binom{n+1}{l}$ terms. In the sequel we will abbreviate this probability by $c_l(n)$. Note that it does not depend on δ.

Note that we can write

$$\mathbb{P}[X_i = x_i \mid X_{i+1} = x_{i+1}] = \frac{1}{2} + \frac{1}{2}(1-2p)x_i x_{i+1},$$

so

$$\mathbb{P}[X_0 = x_0, \ldots, X_n = x_n] = \frac{1}{2}\prod_{i=0}^{n-1}\left(\frac{1}{2} + \frac{1}{2}(1-2p)x_i x_{i+1}\right). \tag{4.3.4}$$

This provides a way to calculate the probabilities in $c_l(n)$. As

$$\mathbb{P}[X_0 = y_0 z_0, \ldots, X_n = y_n z_n] = \frac{1}{2}\prod_{i=0}^{n-1}\left(\frac{1}{2} + \frac{1}{2}(1-2p)y_i\, z_i\, y_{i+1}\, z_{i+1}\right).$$

we have

$$c_l(n) = \frac{1}{2}\sum_{\#\{i:z_i=-1\}=l}\prod_{i=0}^{n-1}\left(\frac{1}{2} + \frac{1}{2}(1-2p)y_i\, z_i\, y_{i+1}\, z_{i+1}\right). \tag{4.3.5}$$

Continuing with (4.3.1), we now have

$$f_k^{(n)} = \frac{1}{k!}\left[\sum_{l=0}^{n+1} c_l(n)\,\frac{\partial^k}{\partial\delta^k}\,\left(\delta^l(1-\delta)^{n-l+1}\right)\right]\Bigg|_{\delta=0}.$$

We can work out this last term. Note that by Leibniz's rule [1]:

$$\frac{\partial^k}{\partial\delta^k}\,\delta^l(1-\delta)^{n-l+1} = \sum_{m=0}^{k}\binom{k}{m}\,\frac{\partial^m}{\partial\delta^m}\delta^l\,\frac{\partial^{(k-m)}}{\partial\delta^{(k-m)}}(1-\delta)^{n-l+1}.$$

The two terms in the right-hand side are given by

$$\frac{\partial^m}{\partial\delta^m}\delta^l = \frac{l!}{(l-m)!}\delta^{(l-m)}, \qquad \text{for } m \le l,$$

and zero otherwise, and

$$\frac{\partial^{(k-m)}}{\partial\delta^{(k-m)}}(1-\delta)^{n-l+1} = (-1)^{(k-m)}\,\frac{(n-l+1)!}{(n-l+1-(k-m))!}(1-\delta)^{(n-l+1-(k-m))}, \qquad \text{for } k-m \le n-l+1,$$

and zero otherwise.

Now we plug in $\delta = 0$. The first term is only non-zero for $m = l$ and then equals one. For $m = l$ the constraint for the second term reduces to $k \leq n+1$, and in this case it equals one for all n, l, k, m. We have:

$$\begin{aligned}
f_k^{(n)} &= \frac{1}{k!} \left[\sum_{l=0}^{n+1} c_l(n) \, \frac{\partial^k}{\partial \delta^k} \, \left(\delta^l (1-\delta)^{n-l+1}\right) \right] \Bigg|_{\delta=0} \\
&= \frac{1}{k!} \sum_{l=0}^{n+1} c_l(n) \sum_{\substack{m=0,\, m\leq l, \\ k-m\leq n-l+1}}^{k} \delta^{(l-m)} \, (-1)^{(k-m)} \binom{k}{m} \frac{l!}{(l-m)!} \frac{(n-l+1)!}{(n-l+1-(k-m))!} \Bigg|_{\delta=0}.
\end{aligned}$$

Working this out gives $f_k^{(n)} = 0$ for $k \geq n+1$, and

$$\begin{aligned}
f_k^{(n)} &= \sum_{l=0}^{n+1} c_l(n) \, (-1)^{(k-l)} \, \frac{1}{k!} \binom{k}{l} \frac{l!}{(l-l)!} \frac{(n-l+1)!}{(n-k+1)!} \\
&= \sum_{l=0}^{n+1} \frac{c_l(n)}{(k-l)!} \, (-1)^{(k-l)} \, \frac{(n-l+1)!}{(n-k+1)!}
\end{aligned} \tag{4.3.6}$$

otherwise. The first $f_k^{(n)}$'s are given by:

$$\begin{aligned}
f_0^{(n)} &= c_0(n), \\
f_1^{(n)} &= -(n+1)c_0(n) + c_1(n), \\
f_2^{(n)} &= \begin{cases} \frac{n(n+1)}{2} c_0(n) - n c_1(n) + c_2(n) & \text{if } n \geq 1; \\ 0 & \text{otherwise,} \end{cases} \\
f_3^{(n)} &= \begin{cases} \frac{-(n-1)n(n+1)}{6} c_0(n) + \frac{(n-1)n}{2} c_1(n) - (n-1)c_2(n) + \frac{1}{6} c_3(n) & \text{if } n \geq 2; \\ 0 & \text{otherwise.} \end{cases}
\end{aligned}$$

4.3.2 Series expansion $\mathbb{P}[Y_0 \mid Y_1, \ldots, Y_n]$

In the last section we wrote $\mathbb{P}[Y_0, Y_1, \ldots, Y_n]$ as a polynomial in δ. Using Bayes' Rule we will use this to express the conditional probability $\mathbb{P}[Y_0 | Y_1, \ldots, Y_n]$ as a series expansion in δ around $\delta = 0$:

$$\begin{aligned}
\mathbb{P}[Y_0 | Y_1, \ldots, Y_n] = \frac{\mathbb{P}[Y_0, Y_1, \ldots, Y_n]}{\mathbb{P}[Y_1, \ldots, Y_n]} \; &= \; \frac{\sum_{k=0}^{n+1} a_k^{(n)} \delta^k}{\sum_{k=0}^{n} b_k^{(n)} \delta^k} \\
&= \frac{a_0}{b_0} + \frac{(a_1 b_0 - a_0 b_1)}{b_0^2} \delta + \frac{\left(a_2 b_0^2 - a_1 b_1 b_0 + a_0 \left(b_1^2 - b_0 b_2\right)\right)}{b_0^3} \delta^2 + \ldots \\
&= \sum_{k=0}^{\infty} F_k^{(n)} \delta^k,
\end{aligned} \tag{4.3.7}$$

where the a_i and b_i in the second line should be read as $a_i^{(n)}$ and $b_i^{(n)}$ respectively, and $F_k^{(n)} = F_k^{(n)}(p\,; y_0, \ldots, y_n)$. The general form for $F_k^{(n)}$ is given in [17]:

$$F_k^{(n)} = \frac{(-1)^k}{b_0^{k+1}} \begin{vmatrix} a_0b_1 - b_0a_1 & b_0 & 0 & \ldots & 0 \\ a_0b_2 - b_0a_2 & b_1 & b_0 & \ldots & 0 \\ a_0b_3 - b_0a_3 & b_2 & b_1 & \ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ a_0b_{k-1} - b_0a_{k-1} & b_{k-2} & b_{k-3} & \ldots & b_0 \\ a_0b_k - b_0a_k & b_{k-1} & b_{k-2} & \ldots & b_1 \end{vmatrix},$$

where again a_i should be read as $a_i^{(n)}$, and b_i as $b_i^{(n)}$. Note that

$$\begin{aligned} a_k^{(n)} &= f_k^{(n)}(p\,; y_0, \ldots, y_n), \\ b_k^{(n)} &= f_k^{(n-1)}(p\,; y_1, \ldots, y_n), \end{aligned} \tag{4.3.8}$$

where the $f_k^{(n)}$ are as given in (4.3.6).

4.3.3 Calculating F_k's

We will now give a few of the coefficients in the series expansion (4.3.7). Because of the settlement of the coefficients, see Theorem 4.2, we have $F_k = F_k^{(n)}$ for $n \geq k+1$. So to calculate F_k it suffices to consider the expansion of $\mathbb{P}[Y_0 \mid Y_1, \ldots, Y_{k+1}]$.

In order to find F_0, we calculate $\mathbb{P}[Y_0 \mid Y_1] = \mathbb{P}[Y_0, Y_1]/\mathbb{P}[Y_0]$. The denominator is trivially 1/2, and for the nominator we have from (4.3.3) and (4.3.6):

$$\begin{aligned} \mathbb{P}[Y_0 = y_0, Y_1 = y_1] &= f_0^{(1)} + f_1^{(1)}\delta + f_2^{(1)}\delta^2 \\ &= c_0(1) + \big(c_1(1) - 2c_0(1)\big)\delta + \frac{1}{2}\big(2c_0(1) - 2c_1(1) + 2c_2(1)\big)\delta^2, \end{aligned}$$

where $c_i(1) = c_i(p\,; 1\,; y_0, y_1)$. These can, using (4.3.5), be determined:

$$\begin{aligned} c_0(1) &= \mathbb{P}[X_0 = y_0, X_1 = y_1] \\ &= \frac{1}{2}\left(\frac{1}{2} + \frac{1}{2}(1-2p)y_0y_1\right), \\ c_1(1) &= \mathbb{P}[X_0 = \bar{y}_0, X_1 = y_1] + \mathbb{P}[X_0 = y_0, X_1 = \bar{y}_1] \\ &= \frac{1}{2} - \frac{1}{2}(1-2p)y_0y_1, \\ c_2(1) &= \mathbb{P}[X_0 = \bar{y}_0, X_1 = \bar{y}_1] \\ &= \frac{1}{2}\left(\frac{1}{2} + \frac{1}{2}(1-2p)y_0y_1\right). \end{aligned}$$

We now have

$$\mathbb{P}[Y_0 = y_0, Y_1 = y_1] = \frac{1}{2}\left(\frac{1}{2} + \frac{1}{2}(1-2p)y_0y_1\right) - (1-2p)y_0y_1\,\delta + (1-2p)y_0y_1\,\delta^2.$$

This gives

$$\mathbb{P}[Y_0 \mid Y_1] = \frac{\mathbb{P}[Y_0, Y_1]}{\mathbb{P}[Y_1]} = \tfrac{1}{2}\left(1 + (1-2p)y_0y_1\right) - 2(1-2p)y_0y_1\delta + 2(1-2p)y_0y_1\delta^2,$$

so

$$F_0 = \tfrac{1}{2}(1 + (1-2p)y_0y_1).$$

In the same way we can derive higher-order coefficients. For F_1 we consider $\mathbb{P}[Y_0 \mid Y_1, Y_2]$, which turns out to be

$$\mathbb{P}[Y_0 \mid Y_1, Y_2] = \tfrac{1}{2}(1 + (1-2p)y_0y_1) - \frac{2(1-2p)y_0y_1}{1 + (1-2p)y_1y_2}\delta + O\left(\delta^2\right)$$

and so

$$F_1 = \frac{-2(1-2p)y_0y_1}{1 + (1-2p)y_1y_2}.$$

Using $\mathbb{P}[Y_0 \mid Y_1, Y_2, Y_3]$ we find F_2:

$$F_2 = \frac{2(1-2p)y_0y_1((1-2p)y_2y_3 - (1-2p)y_1y_2(3 - (1-2p)y_2y_3) + 1)}{((1-2p)y_1y_2 + 1)^2((1-2p)y_2y_3 + 1)},$$

and from $\mathbb{P}[Y_0 \mid Y_1, Y_2, Y_3, Y_4]$ the F_3 follows:

$$F_3 = \frac{16\lambda^2 y_0 y_1^2 y_2 \left(y_1 y_2^2 y_3^2 y_4 \lambda^3 - y_2y_3(y_1(y_2 + y_4) - y_3y_4)\lambda^2 - (y_1y_2 + y_3(y_2 - y_4))\lambda + 1\right)}{(\lambda y_1y_2 + 1)^3(\lambda y_2y_3 + 1)^2(\lambda y_3y_4 + 1)},$$

where $\lambda = 1 - 2p$.

Our aim was to find a general form for these coefficients. From the expressions for F_0, F_1, F_2 and F_3 we see that the denominators have a very nice structure. Unfortunately we are not able to detect a nice structure in the nominators.

Coefficients in λ_i

In the expressions for the F_k in the previous section, we spotted the terms $(1-2p)\, y_i\, y_{i+1}$. These come in because of (4.3.5). This suggests that the expression can become more clear using the terms

$$\lambda_i = (1-2p)\, y_i\, y_{i+1}.$$

This gives for the first four terms:

$$\begin{aligned} F_0 &= \frac{1}{2}(\lambda_0 + 1), \\ F_1 &= -\frac{2\lambda_0}{\lambda_1 + 1}, \\ F_2 &= \frac{2\lambda_0(\lambda_1(\lambda_2 - 3) + \lambda_2 + 1)}{(\lambda_1 + 1)^2(\lambda_2 + 1)}, \\ F_3 &= \frac{16\lambda_0\lambda_1(\lambda_1(\lambda_2(\lambda_3 - 1) - \lambda_3 - 1) + \lambda_2(\lambda_3 - 1) + \lambda_3 + 1)}{(\lambda_1 + 1)^3(\lambda_2 + 1)^2(\lambda_3 + 1)}. \end{aligned} \tag{4.3.9}$$

From this we see again the nice structure of the denominators, but the nominators stay unclear.

4.3.4 Settlement $F_0^{(n)}$

In Theorem 4.2 it is proved that the coefficients $F_k^{(n)}$ settle for $n \geq k+1$. We will show this here for $F_0^{(n)}$. In Appendix B.2 we will show it for $F_1^{(n)}$.

From (4.3.7) it follows that

$$F_0^{(n)} = \frac{a_0^{(n)}}{b_0^{(n)}},$$

where

$$\begin{aligned} a_0^{(n)} &= c_0(p\,;n\,;y_0,\ldots,y_n), \\ b_0^{(n)} &= c_0(p\,;n-1\,;y_1,\ldots,y_n). \end{aligned}$$

According to (4.3.4) we have, writing again $\lambda_i = (1-2p)\,y_i\,y_{i+1}$:

$$\begin{aligned} a_0^{(0)} &= \frac{1}{2}, \\ a_0^{(n)} &= \frac{1}{2^{n+1}}(1+\lambda_0)(1+\lambda_1)\ldots(1+\lambda_{n-1}), \qquad \text{for } n \geq 1. \end{aligned}$$

Furthermore

$$\begin{aligned} b_0^{(0)} &= 1, \qquad b_0^{(1)} = \frac{1}{2}, \\ b_0^{(n)} &= \frac{1}{2^n}(1+\lambda_1)\ldots(1+\lambda_{n-1}), \qquad \text{for } n \geq 2. \end{aligned}$$

Almost all terms cancel out in the division $a_0^{(n)}/\,b_0^{(n)}$, and it follows that:

$$F_0^{(0)} = \frac{1}{2}, \qquad F_0^{(n)} = \frac{1}{2}(1+\lambda_0), \text{ for } n \geq 1.$$

As by (4.3.9) $F_0 = \frac{1}{2}(1+\lambda_0)$, this gives that $F_0^{(n)} = F_0$ for $n \geq 1 = k+1$. Note that F_0 only depends on y_0 and y_1.

The settlement of $F_1^{(n)}$ follows along the same lines. It involves more work, as for this also $a_1^{(n)}$ and $b_1^{(n)}$ need to be calculated. Tedious bookkeeping then gives

$$F_1^{(0)} \neq F_1^{(1)} \neq F_1^{(2)} = F_1^{(3)} = \ldots = F_1,$$

which shows the desired settlement, see Appendix B.2.

4.4 Series expansion in $\xi = \delta/(1-\delta)$

We now consider the series expansion of $\mathbb{P}[Y_0|Y_1,\ldots,Y_n]$ in $\xi = \frac{\delta}{1-\delta}$ around $\xi = 0$:

$$\mathbb{P}[Y_0 \mid Y_1,\ldots,Y_n] = (1-\delta)\sum_{k=0}^{\infty} g_k^{(n)}\xi^k.$$

In the sequel it will turn out that it is convenient to have the term $(1-\delta)$ in front of the summation.

4.4.1 Series expansion

From (4.3.2) we have

$$\mathbb{P}[Y_0, \ldots, Y_n] = \sum_{k=0}^{n+1} \delta^k (1-\delta)^{n-k+1} c_k(n).$$

Let $\xi = \frac{\delta}{1-\delta}$, then we can write:

$$\delta^k (1-\delta)^{n-k+1} = \left(\frac{\delta}{1-\delta}\right)^k (1-\delta)^{n+1} = (1-\delta)^{n+1} \xi^k.$$

This gives

$$\begin{aligned} \mathbb{P}[Y_0 | Y_1, \ldots, Y_n] = \frac{\mathbb{P}[Y_0, Y_1, \ldots, Y_n]}{\mathbb{P}[Y_1, \ldots, Y_n]} &= \frac{(1-\delta)^{n+1} \sum_{k=0}^{n+1} a'^{(n)}_k \xi^k}{(1-\delta)^n \sum_{k=0}^{n} b'^{(n)}_k \xi^k} \\ &= (1-\delta) \sum_{k=0}^{\infty} g_k^{(n)} \xi^k, \end{aligned}$$

where $g_k^{(n)} = g_k^{(n)}(p\,; y_0, \ldots, y_{k+1})$. For these coefficients we see the same settlement for $n \geq k+1$ as we did for the $F_k^{(n)}$.

A simple example of the settlement will be given in the next section.

4.4.2 Example settlement $g_k^{(n)}$

For the case $y = \{1, 1, \ldots\}$ we calculate

$$\mathbb{P}[Y_0 = 1 \mid Y_1 = 1, \ldots, Y_n = 1] = (1-\delta) \sum_{k=0}^{\infty} g_k^{(n)}(p\,; 1, \ldots, 1) \xi^k,$$

where the coefficients are given by

n	$g_0^{(n)}$	$g_1^{(n)}$	$g_2^{(n)}$	$g_3^{(n)}$	$g_4^{(n)}$
0	1	1			
1	$1-p$	$3p-1$	$2(1-2p)$	$-2(1-2p)$	$2(1-2p)$
2	$1-p$	$\frac{p^2}{1-p}$	$\frac{(1-2p)(3p-1)}{(1-p)^2}$	$\frac{(1-2p)\left(5p^2-1\right)}{-(1-p)^3}$	$\frac{(1-2p)\left(7p^3+7p^2-7p+1\right)}{(1-p)^4}$
3	$1-p$	$\frac{p^2}{1-p}$	$\frac{p^2(1-2p)}{(1-p)^3}$	$-\frac{(1-2p)\left(p^2-3p+1\right)^2}{(1-p)^5}$	$\frac{(1-2p)\left(p^3-p^2-2p+1\right)^2}{(1-p)^7}$
4	$1-p$	$\frac{p^2}{1-p}$	$\frac{p^2(1-2p)}{(1-p)^3}$	$-\frac{p^2(1-2p)\left(p^2+2p-1\right)}{(1-p)^5}$	$\frac{(1-2p)\left(p^6+6p^5-15p^4+28p^3-23p^2+8p-1\right)}{(1-p)^7}$
5	$1-p$	$\frac{p^2}{1-p}$	$\frac{p^2(1-2p)}{(1-p)^3}$	$-\frac{p^2(1-2p)\left(p^2+2p-1\right)}{(1-p)^5}$	$\frac{p^2(1-2p)\left(p^4+6p^3+p^2-4p+1\right)}{(1-p)^7}$
6	$1-p$	$\frac{p^2}{1-p}$	$\frac{p^2(1-2p)}{(1-p)^3}$	$-\frac{p^2(1-2p)\left(p^2+2p-1\right)}{(1-p)^5}$	$\frac{p^2(1-2p)\left(p^4+6p^3+p^2-4p+1\right)}{(1-p)^7}$

In Appendix B.3 we give the coefficients g_k for general y. Also we determine the coefficients for the expansion of the logarithm of the conditional probability. We remark some structure for the coefficients in both cases, but we are not able to express them in a general form.

Chapter 5

Recurrence relations

In this chapter we will derive a power series expansion for the entropy of the binary symmetric hidden Markov model, making use of two recurrence relations for the conditional probability $\mathbb{P}[Y_0 = 1|Y_1, \ldots, Y_n]$. This expansion will be in $\zeta = \delta(1-\delta)$ around $\zeta = 0$, where the coefficients are functions of p. For these recurrence relations one only has to keep track of the previous transition probabilities of the process, instead of the entire history of it.

We start by giving and proving the two recurrence relations. Iterating these enables us to compute the conditional probability $\mathbb{P}[Y_0|Y_1, \ldots, Y_n]$, and we find a strict upper and lower bound for it. The main part of this chapter will be the method to derive a power series expansion for the entropy, based on the use of the two relations. The expansion will be derived by substituting one expansion into another. We will give a conjecture for the domain on which it converges. Finally we give a small but efficient simulation to estimate the entropy for given parameters p and δ. We end this chapter by comparing the expansion we found with the result of Zuk et al. [42]. Note that in this chapter we entirely focus on the binary symmetric hidden Markov model.

5.1 Recurrence relations f_1 and f_{-1}

For the conditional probability that $Y_0 = 1$ given the past, we define:

$$w_n(y_1, \ldots, y_n) := \mathbb{P}[Y_0 = 1|Y_1 = y_1, \ldots, Y_n = y_n].$$

This w_n can be expressed in w_{n-1} by two recursive relations given in the following theorem.

Theorem 5.1. *We have*

$$\begin{aligned} w_n(1, y_2, \ldots, y_n) &= f_1(w_{n-1}(y_2, \ldots, y_n)), \\ w_n(-1, y_2, \ldots, y_n) &= f_{-1}(w_{n-1}(y_2, \ldots, y_n)), \end{aligned} \tag{5.1.1}$$

where

$$\begin{aligned} f_1(x) &:= 1 - p - \frac{\delta(1-\delta)(1-2p)}{x}, \\ f_{-1}(x) &:= p + \frac{\delta(1-\delta)(1-2p)}{1-x}. \end{aligned} \tag{5.1.2}$$

Both f_1 and f_{-1} are defined on the interval around $x = \frac{1}{2}$ on which they are strictly between 0 and 1, as will be commented later.

The theorem states that for this hidden Markov process, the transition probabilities form a Markov process, as

$$w_n(y_1, \dots, y_n) = \begin{cases} f_1(w_{n-1}(y_2, \dots, y_n)) & \text{with prob. } w_{n-1}(y_2, \dots, y_n), \\ f_{-1}(w_{n-1}(y_2, \dots, y_n)) & \text{with prob. } 1 - w_{n-1}(y_2, \dots, y_n). \end{cases} \tag{5.1.3}$$

In this way the transition probabilities depend only on the previous ones. This enables us to simulate the process very efficiently without having to keep track of the entire history of it, see Section 5.9.

5.2 Proofs recurrence relations

We will give two proofs of Theorem 5.1. For the first one we express the probability $\mathbb{P}[Y_0, \dots, Y_n]$ in terms of $\mathbb{P}[Y_1, \dots, Y_n]$ and $\mathbb{P}[Y_2, \dots, Y_n]$. This proof is given below. For the second proof we condition on the state of X_1, and this proof is given in Appendix A.7.

Define

$$p_n(y_0, \dots, y_n) := \mathbb{P}[Y_0 = y_0, Y_1 = y_1, \dots, Y_n = y_n].$$

The next proposition gives a recurrence relation for this probability:

Proposition 5.2.

$$p_n(y_0, \dots, y_n) = \frac{\lambda\, y_0\, y_1 + 1}{2}\, p_{n-1}(y_1, \dots, y_n) - \delta(1-\delta)\, \lambda\, y_0\, y_1\, p_{n-2}(y_2, \dots, y_n), \tag{5.2.1}$$

where $\lambda := 1 - 2p$.

The proof of this proposition is given in Appendix A.8. We now prove the theorem.

Proof of Theorem 5.1. Note that by definition and by Bayes' Law:

$$\begin{aligned} w_n(y_1, \dots, y_n) &= \mathbb{P}[Y_0 = 1 \mid Y_1, \dots, Y_n] \\ &= \frac{\mathbb{P}[Y_0 = 1, Y_1 = y_1 \dots Y_n = y_n]}{\mathbb{P}[Y_1 = y_1 \dots Y_n = y_n]} \\ &= \frac{p_n(1, y_1, \dots, y_n)}{p_{n-1}(y_1, \dots, y_n)}. \end{aligned} \tag{5.2.2}$$

Using this and (5.2.1), we get

$$\begin{aligned} w_n(1, y_2, \dots, y_n) &= \frac{p_n(1, 1, y_2, \dots, y_n)}{p_{n-1}(1, y_2, \dots, y_n)} \\ &= \frac{\frac{\lambda+1}{2} p_{n-1}(1, y_2, \dots, y_n) - \delta(1-\delta)\lambda p_{n-2}(y_2, \dots, y_n)}{p_{n-1}(1, y_2, \dots, y_n)} \\ &= \frac{\lambda + 1}{2} - \frac{\delta(1-\delta)\,(1-2p)}{w_{n-1}(y_2, \dots, y_n)}, \end{aligned}$$

where the last equality holds, as by (5.2.2) we have that $w_{n-1}(y_2,\ldots,y_n) = \frac{p_{n-1}(1,y_2,\ldots,y_n)}{p_{n-2}(y_2,\ldots,y_n)}$. As $\lambda = 1-2p$ we have $\frac{\lambda+1}{2} = 1-p$, and we find

$$w_n(1,y_2,\ldots,y_n) = 1-p-\frac{\delta(1-\delta)(1-2p)}{w_{n-1}(y_2,\ldots,y_n)},$$

which proves the first equation of the theorem. Analogously we can derive

$$w_n(-1,y_2,\ldots,y_n) = p+\frac{\delta(1-\delta)(1-2p)}{1-w_{n-1}(y_2,\ldots,y_n)},$$

which proves the second one. □

5.3 Symmetry

By symmetry, it holds that

$$w_n(y_1,\ldots,y_n) = 1-w_n(-y_1,\ldots,-y_n),$$

because

$$\begin{aligned} w_n(y_1,\ldots,y_n) &= \mathbb{P}[Y_0=1|Y_1=y_1,\ldots,Y_n=y_n] \\ &= \mathbb{P}[Y_0=-1|Y_1=-y_1,\ldots,Y_n=-y_n] \\ &= 1-\mathbb{P}[Y_0=1|Y_1=-y_1,\ldots,Y_n=-y_n] \\ &= 1-w_n(-y_1,\ldots,-y_n). \end{aligned}$$

Furthermore the process is symmetric in δ, which directly follows from the fact that f_1 and f_{-1} only depend on δ via the term $\delta(1-\delta)$. It also holds that $f_1(x,p,\delta) = f_{-1}(1-x,1-p,\delta)$ and

$$f_{\pm 1}(x,p,\delta) = 1-f_{\pm 1}(x,1-p,\delta).$$

This last relation will turn out to be important when investigating the radius of convergence of the series expansion given in Section 5.6.

5.4 Iteration of f_1 and f_{-1}

We can express $w_n(y_1,\ldots,y_n)$ in terms of f_1 and f_{-1}:

$$w_n(y_1,\ldots,y_n) = f_{y_1}\left(f_{y_2}\left(\ldots\left(f_{y_{n-1}}\left(f_{y_n}\left(\tfrac{1}{2}\right)\right)\right)\right)\right), \tag{5.4.1}$$

which directly follows from (5.1.1). The fraction $\frac{1}{2}$ comes in because

$$w_0 = \mathbb{P}[Y_0=1] = \tfrac{1}{2}.$$

We can illustrate this by plotting f_1, f_{-1} and x, see Figure 5.1. This plot is for the values $p=0.3$ and $\delta=0.1$, but the shape of the curves will essentially be the same for other choices of $0<p<1/2$ and $0<\delta<1$, $\delta\neq 1/2$. For $1/2<p<1$ the graphs are mirrored in the line $x=1/2$.

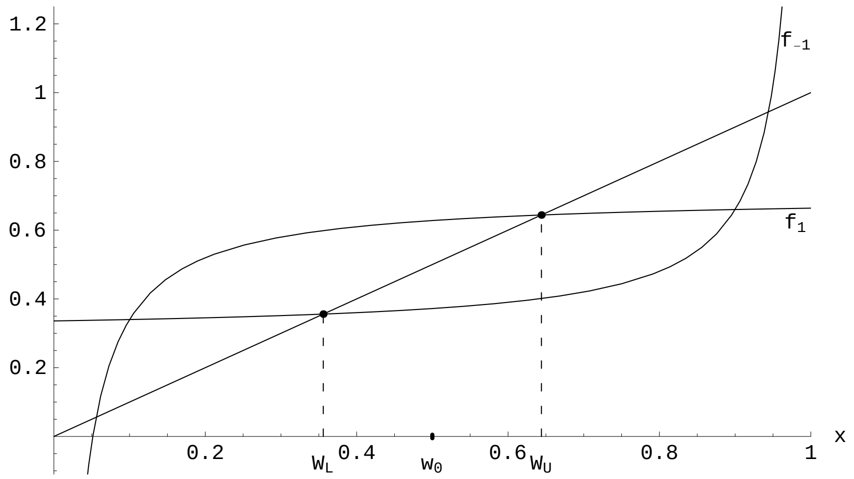


Figure 5.1: Plot of f_1, f_{-1} and x, for $p = 0.3$ and $\delta = 0.1$, with the lower and upper bounds W_L and W_U indicated.

For the iteration we start at $x = w_0 = \frac{1}{2}$, and repeatedly apply either f_1 or f_{-1}, depending on the realization of y_i. From the plot it directly follows that $w_1, w_2, \ldots$ will always be in the interval between the two indicated intersections, those closest to $x = 1/2$. Repeatedly applying f_{-1} will give convergence to the left intersection, and repeatedly applying f_1 to the right one. This holds as the derivatives of f_{-1} respectively f_1 in these points are smaller than 1. The two indicated intersections are the solutions of $f_{-1}(x) = x$ and $f_1(x) = x$. They will be closer investigated in the next section.

5.5 Upper and lower bound

From the plot and the reasoning in the previous section, it followed that w_n is bounded. We will derive tight uniform lower and upper bounds, denoted by W_L respectively W_U. These bounds are tight, so W_L is the largest and W_U the smallest value such that

$$\forall\, n\ \forall\, \{y_1, \ldots, y_n\}: \quad W_L \le w_n(y_1, \ldots, y_n) \le W_U.$$

First we state a result which follows from expression (5.4.1).

Lemma 5.3. *For $p \in (0, \frac{1}{2})$ it holds that, for all n:*

$$\begin{aligned} W_L(n) := f_{-1}\left(f_{-1}\left(\ldots\left(f_{-1}\left(\tfrac{1}{2}\right)\right)\right)\right) \\ \le \quad & w_n(y_1, \ldots, y_n) \quad \le \\ & f_1\left(f_1\left(\ldots\left(f_1\left(\tfrac{1}{2}\right)\right)\right)\right) =: W_U(n), \end{aligned}$$

i.e. n times f_{-1} applied to $\frac{1}{2}$, c.q. n times f_1.

This follows from the fact that $f_{-1}(x) \le f_1(x)$ for all x such that $W_L(n) \le x \le W_U(n)$, for all n. The proof is given in Appendix A.9. By symmetry, we have for these bounds

$$W_L(n) = 1 - W_U(n),$$

for all n. These bounds are tight: For $w_n(-1,\dots,-1)$ and $w_n(1,\dots,1)$ the lower respectively upper bounds hold with equality. For $p > \frac{1}{2}$ the upper and lower bounds are switched. Assume throughout the sequel that $0 < p < \frac{1}{2}$. From the lemma it follows that tight uniform lower and upper bounds are:

$$\begin{aligned} W_L &= \lim_{n\to\infty} W_L(n), \\ W_U &= \lim_{n\to\infty} W_U(n). \end{aligned}$$

Lemma 5.4. *Both* $\lim\limits_{n\to\infty} W_L(n)$ *and* $\lim\limits_{n\to\infty} W_U(n)$ *exist and are finite.*

Proof. We first prove by induction that $W_L(n)$ is decreasing in n.

$$\begin{aligned} W_L(1) = f_{-1}(\tfrac{1}{2}) &= p + 2\delta(1-\delta)(1-2p) \\ &\le p + \tfrac{1}{2}(1-2p) = \tfrac{1}{2} = w_0 = W_L(0), \end{aligned}$$

where the inequality holds as $\delta(1-\delta) \le \frac{1}{4}$. Now assume that $W_L(n+1) \le W_L(n)$ then

$$\begin{aligned} W_L(n+2) &= p + \frac{\delta(1-\delta)(1-2p)}{1-W_L(n+1)} \\ &\le p + \frac{\delta(1-\delta)(1-2p)}{1-W_L(n)} = W_L(n+1). \end{aligned}$$

We have $W_L(n) \le \frac{1}{2}$, and $W_L(n) \ge 0$ as it is a probability. By completeness of the real numbers, it follows that $\lim_{n\to\infty} W_L(n)$ exists and is finite.

By the same reasoning as above it follows that $W_U(n)$ is increasing in n. As $\frac{1}{2} \le W_U(n) \le 1$, we have that $\lim_{n\to\infty} W_U(n)$ exists and is finite. Note that this also follows from the equality $W_L(n) = 1 - W_U(n)$. □

Proposition 5.5. *As tight uniform lower and upper bounds for* $w_n(y_1,\dots,y_n)$*, we have:*

$$\forall n\ \forall\{y_1,\dots,y_n\}:\ W_L \le w_n(y_1,\dots,y_n) \le W_U.$$

Proof. The statement directly follows from Lemma 5.3 and Lemma 5.4. □

For the limit W_L it holds that $W_L = f_{-1}(W_L)$, which is the intersection of $f_1(x)$ and the line $y = x$ in the interval $[0, \frac{1}{2}]$, see Figure 5.1. So we have

$$W_L = p + \frac{\delta(1-\delta)(1-2p)}{1-W_L}.$$

This gives a quadratic equation in W_L, from which W_L can be solved in terms of p and δ:

$$W_L = \frac{1 + p - \sqrt{(1-p)^2 - 4\delta(1-\delta)(1-2p)}}{2}, \tag{5.5.1}$$

where we took the solution of the quadratic equation the gives $W_L \in [0, \frac{1}{2}]$ for all δ and all $p < \frac{1}{2}$. Analogously it holds that $W_U = f_1(W_U)$ and we find

$$W_U = \frac{1 - p + \sqrt{(1-p)^2 - 4\delta(1-\delta)(1-2p)}}{2}. \tag{5.5.2}$$

This is the intersection of f_1 and $y = x$ in $[\frac{1}{2}, 1]$. Note that $W_U = 1 - W_L$.

By the relation $\mathbb{P}[Y_0 = -1 \mid Y_1, \ldots, Y_n] = 1 - w_n(y_1, \ldots, y_n)$, the given bounds are by symmetry also bounds for this probability and hence for $\mathbb{P}[Y_0 \mid Y_1, \ldots, Y_n]$. These bounds are much better than the bounds found in the proof of Proposition 2.5, which gave δ and $1 - \delta$.

The functions f_1 and f_{-1} map the interval $[W_L, W_U]$ to itself. So both functions are in $(0, 1)$ on this domain, which should be true for a probability. The next proposition proves the claim made in Section 5.4.

Proposition 5.6. *The derivatives of f_{-1} and f_1 in W_L respectively W_U are in $(0, 1)$.*

This gives that both W_L and W_U are attracting fixed points of f_{-1} respectively f_1. The proof of this is given in Appendix A.10.

5.6 Expansion entropy

We now derive a power series expansion for $h_Y = h_Y(p, \delta)$ in $\zeta = \delta(1 - \delta)$ around $\zeta = 0$:

$$h_Y = \sum_{k=0}^{\infty} h_{Y,k}(p)\, \zeta^k, \tag{5.6.1}$$

where the $h_{Y,k}$'s depend only on p. The outline of the approach used to derive expressions for the $h_{Y,k}$'s will be as follows. First we consider the expansion

$$h_Y = c_0(p) + 2\, c_1(p)\, \zeta + \sum_{n=2}^{\infty} c_n(p)\, d_{n-1}(p, \zeta)\, \zeta^n, \tag{5.6.2}$$

for some coefficients c_n and d_n, where the d_n depend on ζ. The factor 2 in front of the second coefficient will become clear later. Then we give an expansion for these coefficients:

$$d_n(p, \zeta) = r_{n,0}(p) + 2\, r_{n,1}(p)\, \zeta + \sum_{k=2}^{\infty} r_{n,k}(p)\, d_{k-1}(p, \zeta)\, \zeta^k. \tag{5.6.3}$$

Here all but the first two coefficients depend on ζ. By repeadetly plugging in $d_n(p, \zeta)$ into its own expansion, we find

$$d_n(p, \zeta) = \sum_{k=0}^{\infty} R_{n,k}(p)\, \zeta^k, \tag{5.6.4}$$

where the $R_{n,k}$ depend only on p. This expansion we plug in into (5.6.2), to find an expansion for h_Y where the coefficients do not depend on ζ any more. This is the desired power series expansion (5.6.1).

5.6.1 Expansion entropy, coefficients depending on ζ

The entropy of the process Y is given by

$$h_Y := \lim_{n \to \infty} \mathbb{E}[-\log \mathbb{P}[Y_0 \mid Y_1, \ldots, Y_n]].$$

A series expansion for this as in (5.6.2) will be given by the next theorem. For this we introduce the random variable W, which is the limit of w_n as n tends to infinity:

$$W := \lim_{n\to\infty} w_n(Y_1, \dots, Y_n).$$

By Proposition 2.6 this limit exists.

Theorem 5.7. *The entropy of the process Y is given by*

$$h_Y = h(p) + 2h'(p)\,(1-2p)\,\zeta + \sum_{k=2}^{\infty} \frac{h^{(k)}(p)}{k!}(1-2p)^k \zeta^k\, \mathbb{E}\left[g_{k-1}(W)\right], \tag{5.6.5}$$

where

$$h(p) = -\left(p\log p + (1-p)\log(1-p)\right),$$

and

$$g_n(W) := \frac{1}{W^n} + \frac{1}{(1-W)^n}.$$

Proof. From (5.1.3) it follows that

$$\begin{aligned}\mathbb{E}[w_n(Y_1,\dots,Y_n)] = \mathbb{E}[&w_{n-1}(Y_2,\dots,Y_{n-1})\, f_1\left(w_{n-1}(Y_2,\dots,Y_{n-1})\right) \\ &+ (1-w_{n-1}(Y_2,\dots,Y_{n-1}))\, f_{-1}\left(w_{n-1}(Y_2,\dots,Y_{n-1})\right)].\end{aligned}$$

Let

$$h(p) = -\left(p\log p + (1-p)\log(1-p)\right).$$

This gives for the entropy h_Y

$$\begin{aligned} h_Y &= \lim_{n\to\infty} \mathbb{E}[h\left(w_n(Y_1,\dots,Y_n)\right)] \\ &= \lim_{n\to\infty} \mathbb{E}\big[w_{n-1}(Y_2,\dots,Y_n) h\left(f_1\left(w_{n-1}(Y_2,\dots,Y_n)\right)\right) \\ &\qquad + \left(1-w_{n-1}(Y_2,\dots,Y_n)\right) h\left(f_{-1}\left(w_{n-1}(Y_2,\dots,Y_n)\right)\right)\big]. \end{aligned}$$

Recall that $w_n(Y_1,\dots,Y_n) = \mathbb{P}[Y_0 = 1 | Y_1,\dots,Y_n]$, and that the random variable W is the limit of this:

$$W = \lim_{n\to\infty} w_n(Y_1,\dots,Y_n).$$

By Lebesgue's Bounded Convergence Theorem [2] we can interchange the limit and the expectation in the expression for h_Y, which gives

$$h_Y = \mathbb{E}[W h(f_1(W)) + (1-W) h(f_{-1}(W))].$$

We plug in the expressions for f_{-1} and f_1, and use that $h(x) = h(1-x)$:

$$\begin{aligned} h_Y &= \mathbb{E}\big[W h\left(1-p-\frac{\zeta(1-2p)}{W}\right) + (1-W) h\left(p + \frac{\zeta(1-2p)}{1-W}\right)\big] \\ &= \mathbb{E}\big[W h\left(p+\frac{\zeta(1-2p)}{W}\right) + (1-W) h\left(p + \frac{\zeta(1-2p)}{1-W}\right)\big]. \end{aligned}$$

We now replace both h by its series expansion in ζ around $\zeta = 0$, and collect the powers of ζ. This gives

$$\begin{aligned}
h_Y &= \mathbb{E}\big[Wh(p) + (1-W)h(p) \\
&\qquad + \zeta(1-2p)h'(p) + \zeta(1-2p)h'(p) \\
&\qquad + \zeta^2 \frac{(1-2p)^2}{2W} h''(p) + \zeta^2 \frac{(1-2p)^2}{2(1-W)} h''(p) + \ldots\big] \\
&= \sum_{k=0}^{\infty} \frac{h^{(k)}(p)}{k!} (1-2p)^k \zeta^k \, \mathbb{E}\left[\frac{1}{W^{k-1}} + \frac{1}{(1-W)^{k-1}}\right] \\
&= h(p) + 2h'(p)\,(1-2p)\,\zeta \\
&\qquad + \sum_{k=2}^{\infty} \frac{h^{(k)}(p)}{k!} (1-2p)^k \zeta^k \, \mathbb{E}\left[\frac{1}{W^{k-1}} + \frac{1}{(1-W)^{k-1}}\right].
\end{aligned}$$

The last step holds as for $k = 0$ respectively $k = 1$ we have, for all W:

$$\frac{1}{W^{-1}} + \frac{1}{(1-W)^{-1}} = 1, \qquad \frac{1}{W^0} + \frac{1}{(1-W)^0} = 2.$$

Defining

$$g_n(W) := \frac{1}{W^n} + \frac{1}{(1-W)^n},$$

gives the statement of the theorem. □

The given series expansion (5.6.5) corresponds to (5.6.2) with

$$\begin{aligned}
c_k(p) &= \frac{h^{(k)}(p)}{k!}(1-2p)^k, \\
d_k(p,\zeta) &= \mathbb{E}\left[g_k(W)\right],
\end{aligned} \tag{5.6.6}$$

where W depends on p and ζ. The $h^{(k)}(p)$ denote the kth derivative of h in p. It is straightforward to derive that they are given by

$$\begin{aligned}
h'(p) &= \log\frac{1-p}{p}, \\
h^{(k)}(p) &= (k-2)!\left(\frac{(-1)^{k-1}}{p^{k-1}} - \frac{1}{(1-p)^{k-1}}\right), \text{ for } k \geq 2.
\end{aligned}$$

Denote by $h_{Y,k}$ be the kth term in the series expansion of h_Y, so

$$h_Y = \sum_{k=0}^{\infty} h_{Y,k}(p)\,\zeta^k.$$

Then from (5.6.5) it directly follows that

$$\begin{aligned}
h_{Y,0} &= h(p) \\
&= -p\log p - (1-p)\log(1-p), \\
h_{Y,1} &= 2\,h'(p)\,(1-2p) \\
&= 2(1-2p)\log\frac{1-p}{p}.
\end{aligned}$$

In order to find the $h_{Y,k}$ for $k \geq 2$, we will derive a series expansion for $\mathbb{E}[g_n(W)]$, as will be done in the next section.

5.6.2 Expansion $\mathbb{E}\left[g_n(W)\right]$

To find higher-order terms in the expansion (5.6.5), we express

$$\mathbb{E}[g_n(W)] = \mathbb{E}\left[\frac{1}{W^n} + \frac{1}{(1-W)^n}\right]$$

as a series expansion in ζ around $\zeta = 0$.

Proposition 5.8. *A series expansion of* $\mathbb{E}\left[g_n(W)\right]$ *is given by*

$$\mathbb{E}\left[g_n(W)\right] = g_n(p) + 2\zeta(1-2p)g_n'(p) + \sum_{k=2}^{\infty} \frac{g_n^{(k)}(p)}{k!}\,(1-2p)^k\,\zeta^k \mathbb{E}[g_{k-1}(W)].$$

Proof. We will prove this statement in a way similar to the proof of Theorem 5.7. Note that $g_n(x) = g_n(1-x)$. We have for $n \geq 1$:

$$\begin{aligned}
\mathbb{E}\left[g_n(W)\right] &= \mathbb{E}\left[W g_n(f_1(W)) + (1-W)g_n(f_{-1}(W))\right] \\
&= \mathbb{E}\left[W g_n\left(p + \frac{\zeta(1-2p)}{W}\right) + (1-W)g_n\left(p + \frac{\zeta(1-2p)}{1-W}\right)\right] \\
&= \mathbb{E}\big[W\left(g_n(p) + \frac{\zeta(1-2p)}{W}g_n'(p) + \dots\right) \\
&\qquad + (1-W)\left(g_n(p) + \frac{\zeta(1-2p)}{1-W}g_n'(p) + \dots\right)\big] \\
&= g_n(p) + 2\zeta(1-2p)g_n'(p) + \frac{1}{2}\zeta^2(1-2p)^2 g_n''(p)\mathbb{E}[g_1(W)] + \dots \\
&= \sum_{k=0}^{\infty} \frac{g_n^{(k)}(p)}{k!}\,(1-2p)^k\,\zeta^k \mathbb{E}[g_{k-1}(W)] \\
&= g_n(p) + 2\zeta(1-2p)g_n'(p) + \sum_{k=2}^{\infty} \frac{g_n^{(k)}(p)}{k!}\,(1-2p)^k\,\zeta^k \mathbb{E}[g_{k-1}(W)],
\end{aligned}$$

where for the last step we used that $g_{-1}(W) = 1$ and $g_0(W) = 2$ for all W. $\square$

This expresses $\mathbb{E}\left[g_n(W)\right]$ in terms of $\mathbb{E}\left[g_k(W)\right]$, for $k = 1, 2, \ldots$. Note that the first two coefficients only depend on p. We can repeatedly plug in the expansion into itself. In that way we can find coefficients for the series expansion only depending on p. This will be demonstrated in the sequel of this section. Doing this, we find a power series expansion for $\mathbb{E}\left[g_n(W)\right]$, where an arbitrary coefficient can be found in finite time.

To simplify notation, write as in (5.6.3):

$$\mathbb{E}\left[g_n(W)\right] = r_{n,0}(p) + 2\,r_{n,1}(p)\,\zeta + \sum_{k=2}^{\infty} r_{n,k}(p)\,\zeta^k\,\mathbb{E}\left[g_{k-1}(W)\right], \tag{5.6.7}$$

where

$$r_{n,k}(p) = \frac{g_n^{(k)}(p)}{k!}\,(1-2p)^k.$$

We want to derive a power series expansion like (5.6.4):

$$\mathbb{E}\left[g_n(W)\right] = \sum_{k=0}^{\infty} R_{n,k}(p)\,\zeta^k.$$

Writing out the first few terms of (5.6.7) gives

$$\begin{aligned}\mathbb{E}\left[g_n(W)\right] = r_{n,0} + 2r_{n,1}\,\zeta \\ + r_{n,2}\,\zeta^2\,\mathbb{E}[g_1(W)] \\ + r_{n,3}\,\zeta^3\,\mathbb{E}[g_2(W)] \\ + r_{n,4}\,\zeta^4\,\mathbb{E}[g_3(W)] + \dots .\end{aligned}$$

Now plug in this expansion for $n = 1$ into $\mathbb{E}[g_1(W)]$, and collect the powers of ζ:

$$\begin{aligned}\mathbb{E}\left[g_n(W)\right] &= r_{n,0} + 2r_{n,1}\,\zeta \\ &\quad + r_{n,2}\,\zeta^2\left[r_{1,0} + 2\,r_{1,1}\,\zeta^1 + r_{1,2}\,\zeta^2\,\mathbb{E}[g_1(W)] + \dots\right] \\ &\quad + r_{n,3}\,\zeta^3\,\mathbb{E}[g_2(W)] \\ &\quad + r_{n,4}\,\zeta^4\,\mathbb{E}[g_3(W)] + \dots \\ &= r_{n,0} + 2\,r_{n,1}\,\zeta \\ &\quad + r_{n,2}r_{1,0}\,\zeta^2 \\ &\quad + (r_{n,3}\mathbb{E}[g_2(W)] + 2\,r_{n,2}r_{1,1})\,\zeta^3 \\ &\quad + (r_{n,4}\mathbb{E}[g_3(W)] + r_{n,2}r_{1,2}\,\mathbb{E}[g_1(W)])\,\zeta^4\ + \dots .\end{aligned}$$

Plugging the expansion for $n = 2$ into $\mathbb{E}[g_2(W)]$ gives, after collecting the powers of ζ:

$$\begin{aligned}\mathbb{E}\left[g_n(W)\right] &= r_{n,0} + 2\,r_{n,1}\,\zeta \\ &\quad + r_{n,2}r_{1,0}\,\zeta^2 \\ &\quad + (r_{n,3}r_{2,0} + 2\,r_{n,2}r_{1,1})\,\zeta^3 \\ &\quad + (r_{n,4}\mathbb{E}[g_3(W)] + r_{n,2}r_{1,2}\,\mathbb{E}[g_1(W)] + 2\,r_{n,3}r_{2,1})\,\zeta^4\ + \dots .\end{aligned}$$

In the next step, we have to replace $\mathbb{E}[g_3(W)]$ and $\mathbb{E}[g_1(W)]$ by their expansions. We can keep doing this until we find all coefficients up to a desired order.

An arbitrary coefficient $R_{n,k}$ can be found in finite time. For $R_{n,k}$ one or more terms $\mathbb{E}[g_i(W)]$ have to be replaced in the coefficients of $\zeta^2, \dots, \zeta^k$. This are $k-1$ coefficients. In coefficient $R_{n,i}$ there are a maximum of $i-1$ replacements, which gives that the number of replacements cannot exceed

$$\sum_{i=2}^{k}(i-1) = \frac{k\,(k-1)}{2} = O(k^2).$$

So it takes $O(k^2)$ time to derive the expression for the coefficient $R_{n,k}$.

The first $R_{n,k}$ are given by:

$$\begin{aligned}
R_{n,0} &= r_{n,0}, \\
R_{n,1} &= 2r_{n,1}, \\
R_{n,2} &= r_{1,0}r_{n,2}, \\
R_{n,3} &= 2r_{1,1}r_{n,2} + r_{2,0}r_{n,3}, \\
R_{n,4} &= r_{1,0}r_{1,2}r_{n,2} + 2r_{2,1}r_{n,3} + r_{3,0}r_{n,4}, \\
R_{n,5} &= 2r_{1,1}r_{1,2}r_{n,2} + r_{1,3}r_{2,0}r_{n,2} \\
&\qquad + r_{1,0}r_{2,2}r_{n,3} + 2r_{3,1}r_{n,4} + r_{4,0}r_{n,5}.
\end{aligned}$$

Close investigation of the way in which the expansions are plugged into each other gives us the following system for $R_{n,k}$:

$$\begin{aligned}
R_{n,0} &= r_{n,0},\ n \geq 1, \quad R_{n,1} = 2\,r_{n,1},\ n \geq 1, \\
R_{n,k+1} &= \sum_{i=1}^{k} r_{n,\,i+1}\, R_{i,\,k-i}, \quad n \geq 1, k \geq 1.
\end{aligned} \tag{5.6.8}$$

In this way we can express each $R_{n,k}$ in terms of only p. This leads to the following proposition.

Proposition 5.9. *The power series expansion for $\mathbb{E}[g_n(W)]$ for $n \geq 1$ in terms of ζ around $\zeta = 0$ is given by*

$$\mathbb{E}\left[g_n(W)\right] = \sum_{k=0}^{\infty} R_{n,k}(p)\,\zeta^k, \tag{5.6.9}$$

where the $R_{n,k}$ are as given in (5.6.8).

5.6.3 Power series expansion entropy

We now combine the results of the previous sections to find the desired power series expansion for h_Y. For this we plug in the expansion for $\mathbb{E}[g_n(W)]$ as given in (5.6.9) into the expansion (5.6.5). Collecting the powers of ζ gives

$$\begin{aligned}
h_Y &= c_0(p) + 2\,c_1(p)\,\zeta + \sum_{n=2}^{\infty} c_n(p)\,\zeta^n\,\mathbb{E}[g_{n-1}(W)] \\
&= c_0(p) + 2\,c_1(p)\,\zeta + \sum_{n=2}^{\infty}\left[c_n(p)\,\zeta^n\left(\sum_{k=0}^{\infty} R_{n-1,k}(p)\,\zeta^k\right)\right] \\
&= c_0(p) + 2\,c_1(p)\,\zeta + \sum_{n=2}^{\infty}\left[\zeta^n\left(\sum_{m=2}^{n} c_m(p)R_{m-1,n-m}(p)\right)\right].
\end{aligned}$$

This gives us the main result of this chapter, which is stated in the next theorem.

Theorem 5.10. *The entropy of a binary symmetric hidden Markov model Y, expanded in $\zeta = \delta(1-\delta)$ around $\zeta = 0$ is given by*

$$h_Y = c_0(p) + 2\,c_1(p)\,\zeta + \sum_{n=2}^{\infty}\left[\zeta^n\left(\sum_{m=2}^{n} c_m(p)R_{m-1,n-m}(p)\right)\right] \tag{5.6.10}$$

where the c_n are as given in (5.6.6) *and the $R_{n,k}$ as in* (5.6.8).

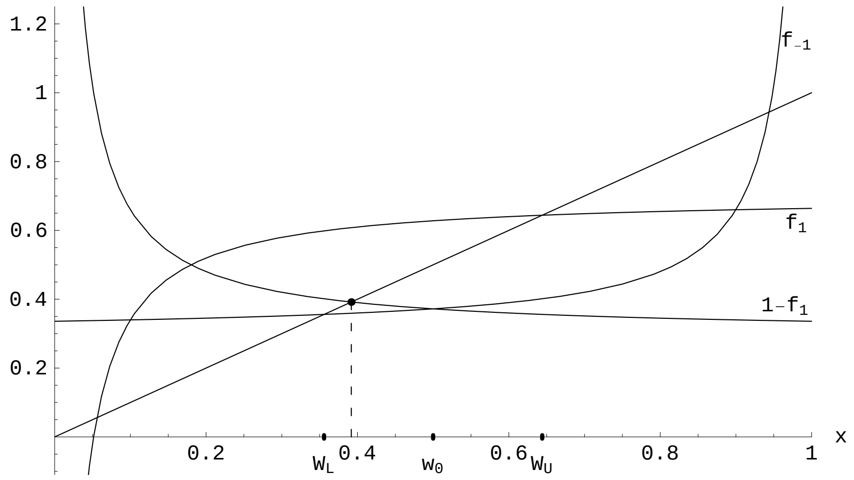


Figure 5.2: Plot of f_1, f_{-1}, x and $1-f_1$, for $p = 0.3$ and $\delta = 0.1$, with the intersection $1-f_1(x) = x$ indicated.

The first ten coefficients are given in Appendix C. As will be shown in Section 5.10, these coefficients coincide with the coefficients found by Zuk et al. [42], which give the power series expansion of h_Y in δ.

5.7 Radius of convergence

We state a conjecture for the interval on which the power series expansion of h_Y converges. First we explain the idea that suggested this conjecture. The given conjecture is supported by numerical results.

Consider

$$\mathbb{E}\left[g_n(W)\right] = \mathbb{E}\left[W g_n(f_1(W)) + (1-W) g_n(f_{-1}(W))\right].$$

As $g_n(x) = g_n(1-x)$ we have

$$g_n(f_{-1}(W)) = g_n(\min\{f_{-1}(W), 1 - f_1(W)\}).$$

We have that $1 - f_1$, see (5.1.2), is given by

$$1 - f_1(x) = p + \frac{\zeta(1-2p)}{x}.$$

Its graph is given in Figure 5.2.

In this way, we get in each step of the iteration actually four terms: $f_1(W), f_{-1}(W), 1 - f_1(W)$ and $1 - f_{-1}(W)$. This leads to four fixed points. Iterating with f_{-1} and f_1 gives respectively W_L

and W_U, see (5.5.1) and (5.5.2). The series expansion for these will converge for

$$|\zeta| < \frac{(1-p)^2}{4(1-2p)},$$

as in general the series expansion for $\sqrt{a - b\,\zeta}$ will converge for $|\zeta| < a/b$. We have that this fraction is larger than 1/4 for all $p \in (0, 1/2]$, so we have convergence of the expansions for W_L and W_U for all $\zeta \in [0, 1/4]$, i.e. for all ζ for which the series expansion of h_Y has an interpretation as entropy.

Now consider the two fixed points which follow from iterating with $1-f_1$ and $1-f_{-1}$. By symmetry we only have to consider one of these. Denote the solution of $1 - f_1(x) = x$ by W^*. It is given by

$$W^* = \frac{p + \sqrt{p^2 + 4(1-2p)\zeta}}{2}.$$

For all ζ and $p \in (0, 1/2]$ we have $W_L \leq W^* \leq 1/2$. The expansion for W^* will converge for

$$|\zeta| < \zeta_{W^*} = \frac{p^2}{4(1-2p)},$$

for $p \in (0, 1/2]$. Only for $p > \sqrt{2} - 1$ we have that $\zeta_{W^*} > 1/4$. This gives that for smaller values of p the expansion will not converge for all ζ. Based on this, we state the following conjecture concerning the radius of convergence for the expansion of h_Y:

Conjecture 5.11. *The interval for p and ζ for which the series expansion* (5.6.10) *converges to h_Y is given by*

$$4\zeta < \begin{cases} p^2/(1-2p) & \text{if } 0 < p < \sqrt{2} - 1, \\ 1 & \text{if } \sqrt{2} - 1 < p < 2 - \sqrt{2}, \\ (1-p)^2/(2p-1) & \text{if } 2 - \sqrt{2} < p < 1. \end{cases}$$

Here the results for $p > 1/2$ followed by symmetry. This conjecture gives that there is at least an interval with positive length where the expansion will converges. The graph corresponding to this area is given in Figure 5.3.

The radius of convergence ζ_r of an arbitrary power series $\sum_{k=0}^{\infty} a_k\, \zeta^k$ is given by

$$\zeta_r = \lim_{k\to\infty} \left| \frac{a_k}{a_{k+1}} \right|,$$

when this limit exists or is ∞. The series then converges for $|\zeta| < \zeta_r$. As the coefficients of the expansion for h_Y are too complex to straightforwardly take this limit, we approximated it by calculating the fraction for increasing k, for given values of $p \in (0, 1)$. Although the convergence of this is very slow, the results of this look like to support the conjecture.

5.8 Plots entropy

In Figure 5.4 we plot the entropy h_Y against p, for $p \in [0, 1]$ and three values of δ: 0.01, 0.1 and 0.5. We give the series expansion using up to the first eighteen orders, so we give:

$$h_Y \approx \sum_{k=0}^{k_{max}} h_{Y,k} \zeta^k,$$

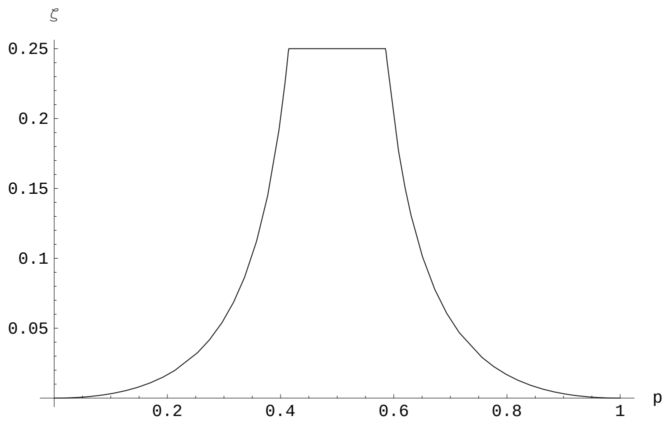


Figure 5.3: The interval for p and ζ for which the series expansion (5.6.10) converges to h_Y, as given in Conjecture 5.11.

for $k_{max} = 0, 1, \ldots, 17$. Moreover we display the estimated convergence interval, which follows from Conjecture 5.11. We also plot the approximation for h_Y found using the simulation given in Section 5.9.

Note that for the case $\delta = 0.5$ the entropy does not depend on the value of p, as every realization of Y can be seen as a fair coin flip. In this case the entropy equals $\log 2$. This is also the value for the entropy in case $p = 0.5$. So $h_Y(p, \delta = 0.5) = h_Y(p = 0.5, \delta) = \log 2$.

5.9 Simulation

Using equations (5.1.2) we can efficiently find a numerical approximation of the entropy by simulation. This makes use of the fact that the transition probabilities of the process Y form a Markov chain, see (5.1.3).

5.9.1 Idea simulation

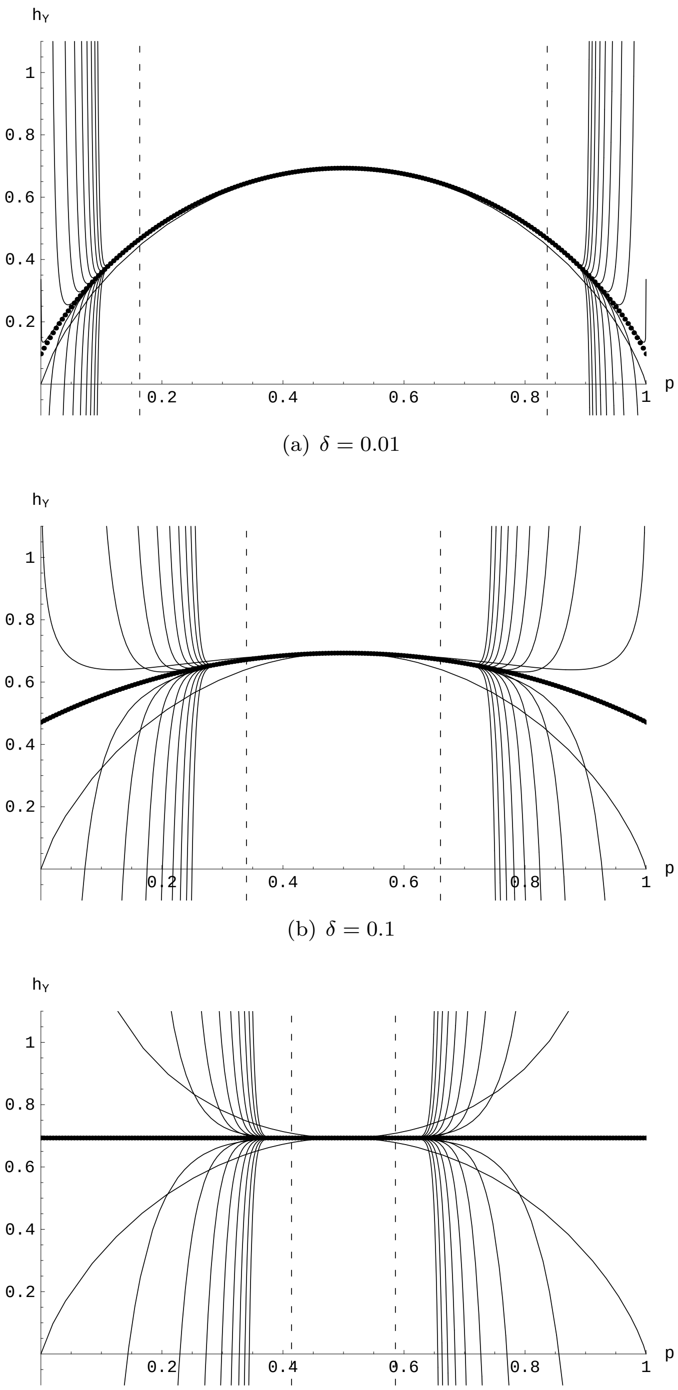
We will simulate a realization of $\{Y_n\}_{n\geq 0}$ by drawing a y randomly from $\{1, -1\}$, where with probability x it will be a 1, so $\mathbb{P}[y = 1] = x$. We only keep track of the probability x, and not of the history of outcomes. We start with $x = w_0 = 1/2$ and update x depending on the outcome of y, by applying either f_1 in case $y = 1$, or f_{-1} in case $y = -1$:

$$\begin{aligned} x &\leftarrow f_1(x) \quad &\text{if } y = 1, \\ x &\leftarrow f_{-1}(x) \quad &\text{if } y = -1. \end{aligned}$$

After every draw we calculate the conditional entropy $H(Y_i \mid Y_{i-1}, \ldots, Y_0)$. This is

$$H(Y_i \mid Y_{i-1}, \ldots, Y_0) = -x \log(x) - (1 - x) \log(1 - x).$$

For this we only need the probability x, and not the history of the process. We keep the running sum over the conditional entropy. At the end we divide it by $n + 1$, the number of realizations



(a) $\delta = 0.01$

(b) $\delta = 0.1$

(c) $\delta = 0.5$

Figure 5.4: Plots of the first eighteen orders in the series expansion of the entropy h_Y, against $p \in [0, 1]$. Also shown the convergence interval and the approximation found by simulation.

that were simulated. The sum is equal to $H(Y_0, \dots, Y_n)$, as, by the chain rule for entropy, see Lemma A.2, we have:

$$H(Y_0, \dots, Y_n) = \sum_{i=0}^{n} H(Y_i \mid Y_{i-1}, \dots, Y_0).$$

In this way we find an approximation for h_Y, as by (2.2.1):

$$h_Y = \lim_{n\to\infty} \frac{1}{n+1} H(Y_0, \dots, Y_n).$$

The approximation becomes better as n increases.

5.9.2 Program

The code of the program:

```
{n = 100, x = 0.5, sum = 0}

For[i = 0, i <= n, i++,
   sum = sum + ( -x Log[x] - (1-x) Log[1-x]);
   If[Random[] < x, x = fp[x], x = fm[x]];
];
sum/(n+1)
```

where `fp` $= f_1$, `fm` $= f_{-1}$, and `Random` draws a random number uniformly on $[0, 1]$.

5.9.3 Results

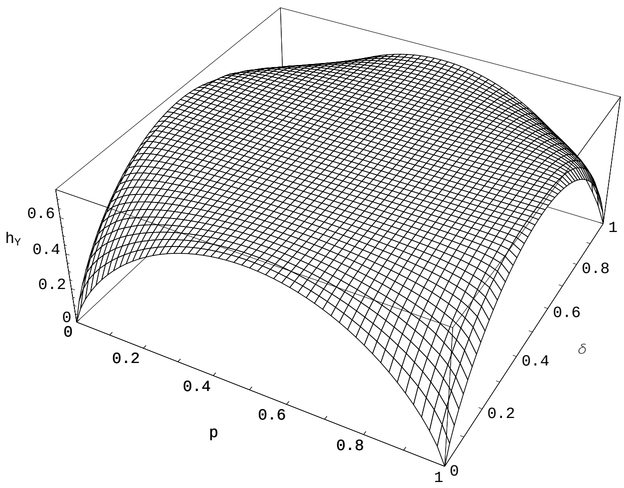
We use the program to approximate the entropy for both p and δ in $[0, 1]$ in steps of 0.02, for $n = 10,000$. The results are given in Figure 5.5. The maximum entropy is achieved in case $\delta = 1/2$ or $p = 1/2$ and is equal to $\log 2 \approx 0.69314\dots$; the minimum is 0 for p and δ both either 0 or 1. Note that, as is to be expected, there is symmetry in both p and δ, i.e. the value of h_Y is equal for p and $1-p$, as well as for δ and $1-\delta$. This does not hold for interchanging p and δ, although from Figure 5.5(a) this may look like to be the case.

5.10 Coefficients series expansions

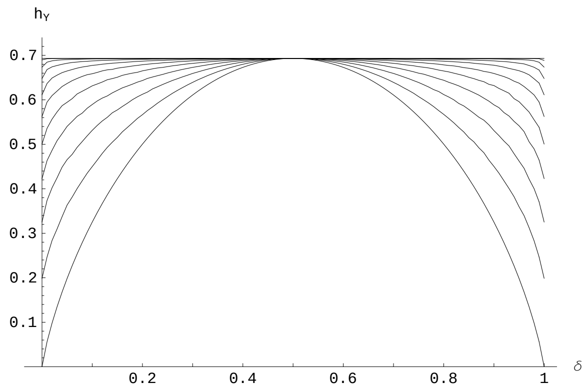
In [42] Zuk et al. express $h_Y(p, \delta)$ as a power series expansion in δ around $\delta = 0$. They give the first twelve coefficients of this expansion, see Section 4.1.1. We show that these are implied by our result from Section 5.6.3: the expansion in $\zeta = \delta(1-\delta)$.

Write $\tilde{f}_k = \tilde{f}_k(p) := h_{Y,k}(p)$, i.e.

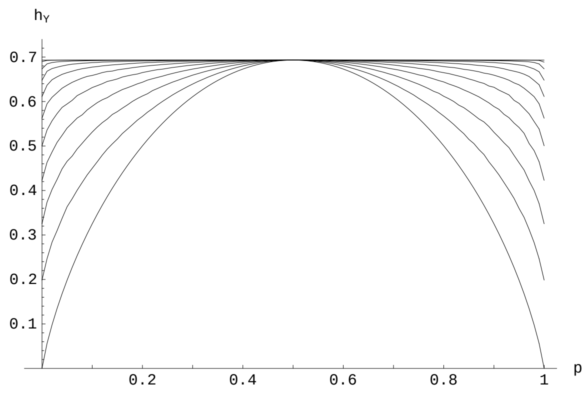
$$h_Y(p, \delta) = \sum_{k=0}^{\infty} \tilde{f}_k (\delta(1-\delta))^k. \tag{5.10.1}$$



(a) h_Y as a function of p and δ.



(b) h_Y as a function of δ for $p \in \{0, 0.05, \ldots, 0.5\}$.



(c) h_Y as a function of p for $\delta \in \{0, 0.05, \ldots, 0.5\}$.

Figure 5.5: Results of the simulation for the entropy h_Y.

Let $f_k = f_k(p)$ be the coefficients of the expansion in δ:

$$h_Y(p,\delta) = \sum_{k=0}^{\infty} f_k \delta^k,$$

found in [42]. We want to express the coefficients $\tilde{f}_k$ in f_k and vice versa.
By the Binomial Theorem [1] we can write

$$(1-\delta)^k = \sum_{l=0}^{k} (-1)^l \binom{k}{l} \delta^l.$$

Plugging in this expansion for $(1-\delta)^k$ in (5.10.1) gives:

$$\begin{aligned}\sum_{k=0}^{\infty} f_k \delta^k &= \sum_{k=0}^{\infty} \tilde{f}_k (\delta(1-\delta))^k \\ &= \sum_{k=0}^{\infty} \sum_{l=0}^{k} (-1)^l \binom{k}{l} \delta^{l+k} \tilde{f}_k.\end{aligned}$$

Let $m = k + l$ then this is

$$\sum_{k=0}^{\infty} f_k \delta^k = \sum_{k=0}^{\infty} \sum_{m=k}^{2k} (-1)^{m-k} \binom{k}{m-k} \delta^m \tilde{f}_k.$$

Now interchange the sums to get

$$\sum_{k=0}^{\infty} f_k \delta^k = \sum_{m=0}^{\infty} \delta^m \sum_{k=\lceil \frac{m}{2} \rceil}^{m} (-1)^{m-k} \binom{k}{m-k} \tilde{f}_k,$$

so the general expression for f_m in terms of $\tilde{f}_k$, $k \leq m$ is:

$$f_m = \sum_{k=\lceil \frac{m}{2} \rceil}^{m} (-1)^{m-k} \binom{k}{m-k} \tilde{f}_k. \tag{5.10.2}$$

The first five f_k's are given by:

$$\begin{array}{lll} f_0 = \tilde{f}_0, & & f_3 = \tilde{f}_3 - 2\tilde{f}_2, \\ f_1 = \tilde{f}_1, & & f_4 = \tilde{f}_4 - 3\tilde{f}_3 + \tilde{f}_2, \\ f_2 = \tilde{f}_2 - \tilde{f}_1, & & \dots \end{array}$$

which can be easily checked by plugging in the expressions for $\tilde{f}_k$.

We can express the result in matrix form notation. Let $\underline{f} = \{f_0,\ f_1,\ \dots\}^T$ and $\underline{\tilde{f}} = \{\tilde{f}_0,\ \tilde{f}_1,\ \dots\}^T$. Then

$$\underline{f} = L\,\underline{\tilde{f}}$$

where the lower diagonal matrix L is, using (5.10.2), given by:

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & -2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 1 & -3 & 1 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 3 & -4 & 1 & 0 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & -1 & 6 & -5 & 1 & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & -4 & 10 & -6 & 1 & 0 & 0 & \dots \\ 0 & 0 & 0 & 0 & 1 & -10 & 15 & -7 & 1 & 0 & \dots \\ 0 & 0 & 0 & 0 & 0 & 5 & -20 & 21 & -8 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Even though the matrix L has infinite dimension, in this case we can define its inverse. This is possible as L is lower-triangular and has only 1's on the diagonal. For an arbitrary dimension, say n, the inverse of L with dimension $n \times n$ is the ordinary inverse of L restricted to be an $n \times n$ matrix. Denoting the inverse of L found in this way by L^{-1}, it holds that

$$\underline{\tilde{f}} = L^{-1}\underline{f}.$$

This gives that the coefficients we have found are implied by the coefficients found in [42]. The matrix L^{-1} is lower diagonal again, and given by:

$$L^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 2 & 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 5 & 5 & 3 & 1 & 0 & 0 & 0 & 0 & 0 & \dots \\ 0 & 14 & 14 & 9 & 4 & 1 & 0 & 0 & 0 & 0 & \dots \\ 0 & 42 & 42 & 28 & 14 & 5 & 1 & 0 & 0 & 0 & \dots \\ 0 & 132 & 132 & 90 & 48 & 20 & 6 & 1 & 0 & 0 & \dots \\ 0 & 429 & 429 & 297 & 165 & 75 & 27 & 7 & 1 & 0 & \dots \\ 0 & 1430 & 1430 & 1001 & 572 & 275 & 110 & 35 & 8 & 1 & \dots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Chapter 6

Conclusion and discussion

In this thesis we considered the entropy of hidden Markov models. First we gave different bounds for the convergence rate of the conditional probability $\mathbb{P}[Y_0 \mid Y_1, \ldots, Y_n]$ for these models. It turned out that the best rate was given by Fernandez, Ferrari and Galves [15].

We proved the settlement of the coefficients of the series expansion for this conditional probability, in the same way as Zuk et al. [39]. We tried to find a general form for the coefficients in these expansions for the binary symmetric case. Although we tried different strategies to do so, we did not succeed in this. The coefficients found by the different methods, showed all some structure, but it turned out that we were not able to spot enough structure to give a general form. So this remains as a major challenge.

In the last chapter we derived a method to obtain a power series expansion for the entropy in the binary symmetric case. This gave an expansion in ζ for the entropy h_Y in this case. Using this method one can generate an arbitrary number of coefficients of this expansion. We also gave an efficient way to simulate this entropy. Next to that we stated a conjecture concerning the radius of convergence for the series expansion, but the proof that this is indeed the correct radius is still open.

Appendix A

Proofs

In this appendix we give the proofs which were left out from the main part of this thesis.

A.1 Proposition 2.1

In this section we prove that a grouped Markov chain can be written as a hidden Markov model, and vice versa.

Proof of Proposition 2.1. ∘ Given a grouped Markov chain $\hat{Y}$ as in Section 2.1.5. To write this as a hidden Markov model, take for the underlying Markov process X the underlying Markov chain of $\hat{Y}$, which is $\hat{X}$. Take $Y = \{Y_n\}_{n\geq 0}$ to be the process defined by $\mathbb{P}[Y_n = k \mid X_n = j] = N_{jk}$, where the emission probability matrix N is given by

$$N_{jk} = \begin{cases} 1 & \text{if } j \in \mathcal{B}_k, \\ 0 & \text{otherwise,} \end{cases}$$

for $j \in \hat{\mathcal{S}}$ and $k \in \hat{\mathcal{S}}'$. Now this hidden Markov model Y gives the same process as the grouped Markov chain $\hat{Y}$.

∘ Given a hidden Markov model: X the hidden Markov process with transition probability matrix P, and Y the observed process with emission probability matrix Π. To write this as a grouped Markov chain, define the process $V = \{V_n\}_{n\geq 0}$ by $V_n = (X_n, Y_n)$. As Y_n only depends on X_n, this is a Markov chain. Its state space is given by $\mathcal{S} \times \mathcal{S}'$. Let Δ be the transition probability matrix of this process, given by

$$\Delta_{\{i,k\},\{i',k'\}} = P_{ii'}\Pi_{i'k'}, \tag{A.1.1}$$

for $i, i' \in \mathcal{S}$ and $k, k' \in \mathcal{S}'$. Let $\mathcal{B}_1, \ldots, \mathcal{B}_{|\mathcal{S}'|}$ be mutually exclusive and exhaustive nonempty subsets of $\mathcal{S} \times \mathcal{S}'$, such that

$$V_n = (X_n, Y_n) \in \mathcal{B}_i \Leftrightarrow Y_n = i.$$

Now this grouped Markov chain V gives the same process as the hidden Markov model Y. □

A.2 Lemma 2.2

In this section we give the proof of Lemma 2.2 as in [8]. This makes use of the following two lemmas

Lemma A.1 (Cesáro mean). *Let $\{a_n\}$ be a sequence of real numbers. If $a_n \overset{n\to\infty}{\longrightarrow} a$ and $b_n = \frac{1}{n}\sum_{i=1}^{n} a_n$, then $b_n \overset{n\to\infty}{\longrightarrow} a$.*

The proof of this can be found in, for instance, [2].

Lemma A.2 (Chain rule for entropy). *For the random variable $(U_0, \ldots, U_n)$ it holds that*

$$H(U_0, \ldots, U_n) = \sum_{i=0}^{n} H(U_i \mid U_{i-1}, \ldots, U_0).$$

We give the proof of this as in [8].

Proof. First we show that it holds that

$$H(U_0, U_1) = H(U_0) + H(U_0 \mid U_1).$$

This follows by just writing out the entropy and condition on U_0:

$$\begin{aligned}
H(U_0, U_1) &= -\sum_{U_0}\sum_{U_1} \mathbb{P}[U_0, U_1] \log \mathbb{P}[U_0, U_1] \\
&= -\sum_{U_0}\sum_{U_1} \mathbb{P}[U_0, U_1] \log \mathbb{P}[U_0]\mathbb{P}[U_1 \mid U_0] \\
&= -\sum_{U_0}\sum_{U_1} \mathbb{P}[U_0, U_1] \log \mathbb{P}[U_0] - \sum_{U_0}\sum_{U_1} \mathbb{P}[U_0, U_1] \log \mathbb{P}[U_1 \mid U_0] \\
&= -\sum_{U_0} \mathbb{P}[U_0] \log \mathbb{P}[U_0] - \sum_{U_0}\sum_{U_1} \mathbb{P}[U_0, U_1] \log \mathbb{P}[U_1 \mid U_0] \\
&= H(U_0) + H(U_0 \mid U_1).
\end{aligned}$$

Equivalently it holds that $H(U_0, U_1) = H(U_0) + H(U_1 \mid U_0)$. Repeadetly applying this gives the statement of the lemma. □

We now give the proof of Lemma 2.2.

Proof of Lemma 2.2. By Lemma A.2 we have:

$$H(Y_0, \ldots, Y_n) = \sum_{i=0}^{n} H(Y_i \mid Y_{i-1}, \ldots, Y_0).$$

Dividing by $n+1$ and taking the limit gives:

$$\lim_{n\to\infty} \frac{1}{n+1} H(Y_0, \ldots, Y_n) = \lim_{n\to\infty} \frac{1}{n+1} \sum_{i=0}^{n} H(Y_i \mid Y_{i-1}, \ldots, Y_0).$$

Now we apply Lemma A.1 to the right-hand side of this to get the desired result:

$$\lim_{n\to\infty} \frac{1}{n+1} H(Y_0, \ldots, Y_n) = \lim_{n\to\infty} H(Y_n \mid Y_{n-1}, \ldots, Y_0).$$ □

A.3 Lemma 2.4

In this section we prove the subadditivity lemma, from which the main argument in due to Fekete [13].

Proof of Lemma 2.4. Assume that condition (2.2.2) holds for the sequence $\{x_n\}$. With induction on k it follows that $x_{k\,m} \leq k\,x_m$, for all $m, k \in \mathbb{N}$. Note that every $n \in \mathbb{N}$ can be written as $n = k\,m + r$ with $0 \leq r \leq m-1$. Let $C_m = \max_{0 \leq r < m} x_r$. Then for all $r \in [0, 1, \ldots, m-1]$ and all $n, k \in \mathbb{N}$ we have

$$x_n = x_{k\,m+r} \leq x_{k\,m} + x_r \leq x_{k\,m} + C_m \leq k\,x_m + C_m.$$

Hence

$$\begin{aligned} \frac{x_n}{n} &\leq \frac{k\,x_m}{n} + \frac{C_m}{n} \\ &= \frac{k\,m}{n}\,\frac{x_m}{m} + \frac{C_m}{n}. \end{aligned}$$

Let $n \to \infty$, then we get

$$\limsup_{n\to\infty} \frac{x_n}{n} \leq \frac{x_m}{m}, \text{ for all } m \geq 1,$$

as $k\,m$ and C_m are constants not depending on n. So

$$\limsup_{n\to\infty} \frac{x_n}{n} \leq \inf_{m\geq 1} \frac{x_m}{m}.$$

But on the other hand

$$\frac{x_n}{n} \geq \inf_{m\geq 1} \frac{x_m}{m}, \text{ for all } n \geq 1,$$

so it follows that

$$\lim_{n\to\infty} \frac{x_n}{n} = \inf_{m\geq 1} \frac{x_m}{m}.$$ □

A.4 Proposition 2.5

In this section we give two alternative proofs of Proposition 2.5. The first one is along the same lines as the first proof:

Second proof of Proposition 2.5. We have:

$$
\begin{aligned}
&\mathbb{P}[Y_0|Y_1,\dots,Y_n] \\
&= \frac{\mathbb{P}[Y_0,Y_1,\dots,Y_n]}{\mathbb{P}[Y_1,\dots,Y_n]} \\
&= \frac{\displaystyle\sum_{X_0,X_1,\dots,X_n} \mathbb{P}[Y_0,Y_1,\dots,Y_n|X_0,X_1,\dots,X_n]\mathbb{P}[X_0,X_1,\dots,X_n]}{\displaystyle\sum_{X_1,\dots,X_n} \mathbb{P}[Y_1,\dots,Y_n|X_1,\dots,X_n]\mathbb{P}[X_1,\dots,X_n]} \\
&= \frac{\displaystyle\sum_{X_0,X_1,\dots,X_n} \mathbb{P}[X_0]\prod_{i=0}^{n-1}\mathbb{P}[X_{i+1}|X_i]\prod_{i=0}^{n}\mathbb{P}[Y_i|X_i]}{\displaystyle\sum_{X_1,\dots,X_n} \mathbb{P}[X_1]\prod_{i=1}^{n-1}\mathbb{P}[X_{i+1}|X_i]\prod_{i=1}^{n}\mathbb{P}[Y_i|X_i]} \\
&= \frac{\displaystyle\sum_{X_1,\dots,X_n} \prod_{i=1}^{n-1}\mathbb{P}[X_{i+1}|X_i]\prod_{i=1}^{n}\mathbb{P}[Y_i|X_i]\left(\sum_{X_0}\mathbb{P}[X_0]\mathbb{P}[X_1|X_0]\mathbb{P}[Y_0|X_0]\right)}{\displaystyle\sum_{X_1,\dots,X_n} \mathbb{P}[X_1]\prod_{i=1}^{n-1}\mathbb{P}[X_{i+1}|X_i]\prod_{i=1}^{n}\mathbb{P}[Y_i|X_i]}.
\end{aligned}
$$

Using Lemma 3.1 this gives, assuming $\delta \le \frac{1}{2}$,

$$
\begin{aligned}
\mathbb{P}[Y_0|Y_1,\dots,Y_n] &\ge \min_{X_1}\frac{\sum_{X_0}\mathbb{P}[X_0]\mathbb{P}[X_1|X_0]\mathbb{P}[Y_0|X_0]}{\mathbb{P}[X_1]} \\
&\ge \min_{X_0,Y_0}\mathbb{P}[Y_0|X_0] = a,
\end{aligned}
$$

and analogously

$$
\begin{aligned}
\mathbb{P}[Y_0|Y_1,\dots,Y_n] &\le \max_{X_1}\frac{\sum_{X_0}\mathbb{P}[X_0]\mathbb{P}[X_1|X_0]\mathbb{P}[Y_0|X_0]}{\mathbb{P}[X_1]} \\
&\le \max_{X_0,Y_0}\mathbb{P}[Y_0|X_0] = b.
\end{aligned}
$$

As $\Pi > 0$ we have $a > 0$ and $b < 0$, and the statement of the proposition follows. $\square$

We will give the third proof only for the binary symmetric case, although it can be easily extended to the general case. It is based on conditioning only on X_0.

Third proof of Proposition 2.5 (for binary symmetric hidden Markov model). We have

$$
\begin{aligned}
&\mathbb{P}[Y_0 = 1 \mid Y_1,\dots,Y_n] \\
&= \mathbb{P}[Y_0 = 1 \mid X_0 = 1, Y_1,\dots,Y_n]\mathbb{P}[X_0 = 1 \mid Y_1,\dots,Y_n] \\
&\qquad + \mathbb{P}[Y_0 = 1 \mid X_0 = -1, Y_1,\dots,Y_n]\mathbb{P}[X_0 = -1 \mid Y_1,\dots,Y_n] \\
&= \mathbb{P}[Y_0 = 1 \mid X_0 = 1]\mathbb{P}[X_0 = 1 \mid Y_1,\dots,Y_n] \\
&\qquad + \mathbb{P}[Y_0 = 1 \mid X_0 = -1]\mathbb{P}[X_0 = -1 \mid Y_1,\dots,Y_n] \\
&= (1-\delta)q + \delta(1-q) \ \in [\delta, 1-\delta],
\end{aligned}
$$

as $q := \mathbb{P}[X_0 = 1 \mid Y_1, \ldots, Y_n] \in [0,1]$, and assuming $\delta \leq \frac{1}{2}$. The analogous result holds for $\mathbb{P}[Y_0 = -1 \mid Y_1, \ldots, Y_n]$. As $\delta > 0$ the proposition now follows. $\square$

A.5 Proposition 3.3

In this section we prove that if the coupling $\tilde{X}$ is successful, then X is weakly ergodic.

Proof of Proposition 3.3. We will prove this statement in the same way as in [18]. Observe that for an arbitrary coupling it holds that

$$\tilde{P}_{g,h}(\tilde{x}_n = (k,k)) \leq \min\{p_{gk}^{(n)}, p_{hk}^{(n)}\}, \quad \forall\, n.$$

Summing over $k \in \mathcal{S}$ gives

$$\tilde{P}_{g,h}(\tilde{x}_n \in \mathcal{D}) \leq \sum_k \min\left\{p_{gk}^n, p_{hk}^n\right\} =: \alpha_{gh}^{(n)}, \quad \forall\, n.$$

Note that $\{\tilde{x}_n \in \mathcal{D}\} = \{T \leq n\}$, and $\alpha_{gh}^{(n)} \leq 1$ for all n. It now follows that

$$\liminf_{n\to\infty} \alpha_{gh}^{(n)} \geq \lim_{n\to\infty} \tilde{P}_{gh}(T \leq n) = \tilde{P}_{gh}(T < \infty).$$

For a successful coupling $\tilde{P}_{gh}(T < \infty) = 1$, which gives

$$\lim_{n\to\infty} \alpha_{gh}^{(n)} = 1. \tag{A.5.1}$$

Using the identity $\min\{a,b\} = \frac{1}{2}(a + b - |a-b|)$ for any real a, b we get

$$\begin{aligned} \alpha_{gh}^{(n)} &= \sum_k \min\left\{p_{gk}^{(n)}, p_{hk}^{(n)}\right\} \\ &= \frac{1}{2}\sum_k \left(p_{gk}^{(n)} + p_{hk}^{(n)} - \left|p_{gk}^{(n)} - p_{hk}^{(n)}\right|\right) \\ &= 1 - \frac{1}{2}\sum_k \left|p_{gk}^{(n)} - p_{hk}^{(n)}\right|. \end{aligned}$$

Taking the limit of n to infinity gives, using (A.5.1)

$$\lim_{n\to\infty} \sum_k |p_{gk}^{(n)} - p_{hk}^{(n)}| = 0,$$

which was to be proved. $\square$

A.6 Proposition 3.4

Proof of Proposition 3.4. Let

$$\lambda = \min_{m,n} \frac{p_{im}p_{jn}}{p_{jm}p_{in}}.$$

Then we have to prove that for any i, j

$$\sum_k p_{ik} p_{jk} \geq \frac{\lambda}{K^2}.$$

Note that

$$\lambda \leq \frac{p_{ik} p_{jk}}{p_{jk} p_{ik}} = 1.$$

Moreover, for any i, j, m, n

$$\frac{p_{im} p_{jn}}{p_{jm} p_{in}} \geq \lambda \Leftrightarrow p_{im} \geq p_{jm} \lambda \frac{p_{in}}{p_{jn}}. \tag{A.6.1}$$

We now claim that, for all $i, j \in \mathcal{S}$ there exists a $n \in \mathcal{S}$, such that

$$\frac{p_{in}}{p_{jn}} \geq 1.$$

For this, suppose that for all n: $\frac{p_{in}}{p_{jn}} < 1$. Then $\sum_n p_{in} < \sum_n p_{jn}$. But as $\sum_n p_{in} = \sum_n p_{jn} = 1$ this gives a contradiction, so the claim holds.

Now take such an n. Then, continuing at A.6.1, we conclude that for every m: $p_{im} \geq p_{jm} \lambda$. This gives

$$\sum_k p_{ik} p_{jk} \geq \sum_k \lambda p_{jk} p_{jk} = \lambda \sum_k p_{jk}^2.$$

We now claim that

$$\sum_{k=1}^{K} p_{jk}^2 \geq \frac{1}{K}.$$

Indeed, by the Cauchy–Schwartz inequality [21] we have

$$1 = \sum_{k=1}^{K} p_{jk} \cdot 1 \leq \left(\sum_{k=1}^{K} p_{jk}^2 \right)^{\frac{1}{2}} \left(\sum_{k=1}^{K} 1 \right)^{\frac{1}{2}} \quad \Rightarrow \quad \left(\sum_{k=1}^{K} p_{jk}^2 \right) K \geq 1 \quad \Rightarrow \quad \sum_{k=1}^{K} p_{jk}^2 \geq \frac{1}{K}.$$

Combining this gives

$$\sum_k p_{ik} p_{jk} \geq \lambda \sum_k p_{jk}^2 \geq \frac{\lambda}{K} \geq \frac{\lambda}{K^2},$$

which finishes the proof. □

A.7 Theorem 5.1

In this section we give a second proof of Theorem 5.1, which is based on conditioning on X_0.

Second proof of Theorem 5.1. By Bayes' Law and conditioning on X_1 in both the numerator and the denominator, we have

$$\begin{aligned}
w_n(1, y_2, \dots, y_n) &= \mathbb{P}[Y_0 = 1 \mid Y_1 = 1, Y_2, \dots, Y_n] \\
&= \frac{\mathbb{P}[Y_0 = 1, Y_1 = 1 \mid Y_2, \dots, Y_n]}{\mathbb{P}[Y_1 = 1 \mid Y_2, \dots, Y_n]}. \\
&= \Big(\mathbb{P}[Y_0 = 1, Y_1 = 1 \mid X_1 = 1]\mathbb{P}[X_1 = 1 \mid Y_2, \dots, Y_n] \\
&\qquad + \mathbb{P}[Y_0 = 1, Y_1 = 1 \mid X_1 = -1]\mathbb{P}[X_1 = -1 \mid Y_2, \dots, Y_n]\Big) \,/ \\
&\quad \Big(\mathbb{P}[Y_1 = 1 \mid X_1 = 1]\mathbb{P}[X_1 = 1 \mid Y_2, \dots, Y_n] \\
&\qquad + \mathbb{P}[Y_1 = 1 \mid X_1 = -1]\mathbb{P}[X_1 = -1 \mid Y_2, \dots, Y_n]\Big)
\end{aligned}$$

Write $q := \mathbb{P}[X_1 = 1 \mid Y_2, \dots, Y_n]$. The other probabilities can be calculated straightforwardly. By conditioning on X_0 we have:

$$\begin{aligned}
&\mathbb{P}[Y_0 = 1, Y_1 = 1 \mid X_1 = 1] \\
&= \mathbb{P}[Y_0 = 1, Y_1 = 1 \mid X_0 = 1, X_1 = 1]\, \mathbb{P}[X_0 = 1 \mid X_1 = 1] \\
&\qquad + \mathbb{P}[Y_0 = 1, Y_1 = 1 \mid X_0 = -1, X_1 = 1]\, \mathbb{P}[X_0 = -1 \mid X_1 = 1] \\
&= (1-\delta)^2(1-p) + \delta(1-\delta)p,
\end{aligned}$$

and analogously

$$\mathbb{P}[Y_0 = 1, Y_1 = 1 \mid X_1 = -1] = (1-\delta)\delta p + \delta^2(1-p).$$

Plugging this in gives:

$$w_n(1, y_2, \dots, y_n) = \frac{q(1-\delta)\left((1-p)(1-\delta) + p\,\delta\right) + (1-q)\delta\left(p\,(1-\delta) + (1-p)\delta\right)}{q(1-\delta) + (1-q)\delta}.$$

We rearrange the terms in the numerator and cancel out common factors:

$$\begin{aligned}
w_n(1, y_2, \dots, y_n) &= \frac{(1-p)\left(q(1-\delta) + (1-q)\delta\right) + \delta(1-\delta)(1-2p)}{q(1-\delta) + (1-q)\delta} \\
&= 1 - p - \frac{\delta(1-\delta)(1-2p)}{\mathbb{P}[Y_1 = 1 \mid Y_2, \dots, Y_n]} \\
&= 1 - p - \frac{\delta(1-\delta)(1-2p)}{w_{n-1}(y_2, \dots, y_n)}.
\end{aligned}$$

So

$$w_n(1, y_2, \dots, y_n) = f_1\left(w_{n-1}(y_2, \dots, y_n)\right),$$

and analogously

$$\begin{aligned}
w_n(-1, y_2, \dots, y_n) &= p + \frac{\delta(1-\delta)(1-2p)}{1 - w_{n-1}(y_2, \dots, y_n)} \\
&= f_{-1}\left(w_{n-1}(y_2, \dots, y_n)\right).
\end{aligned}$$

□

A.8 Proposition 5.2

In order to prove Proposition 5.2 we need the following lemma, which will be proved afterwards.

Lemma A.3. *For the binary symmetric hidden Markov model Y it holds that for all $y_0, \ldots, y_n$:*

$$p_n(y_0, \ldots, y_n) = \frac{1}{2}\, p_{n-1}(y_1, \ldots, y_n) + \frac{(1-2\delta)^2}{2^{n+1}}\, y_0 \sum_{l=1}^{n} 2^{n-l}\, (1-2p)^l\, y_l\, p_{n-1-l}(y_{l+1}, \ldots, y_n). \tag{A.8.1}$$

Proof of Proposition 5.2. The proof of the proposition follows by manipulating (A.8.1). First we take out $l = 1$ from the summation, and then use the identity $\frac{1}{4}\,(1-2\delta)^2 = \frac{1}{4} - \delta(1-\delta)$:

$$\begin{aligned}
&p_n(y_0, \ldots, y_n)\\
&= \frac{1}{2}\, p_{n-1}(y_1, \ldots, y_n) + \frac{1}{4}\,(1-2\delta)^2\, \lambda\, y_0\, y_1\, p_{n-2}(y_2, \ldots, y_n)\\
&\quad + (1/2)^{n+1}\,(1-2\delta)^2\, y_0 \sum_{l=2}^{n} 2^{n-l}\, \lambda^l\, y_l\, p_{n-1-l}(y_{l+1}, \ldots, y_n)\\
&= \frac{1}{2}\, p_{n-1}(y_1, \ldots, y_n) + \frac{1}{4}\,\lambda\, y_0\, y_1\, p_{n-2}(y_2, \ldots, y_n) - \delta(1-\delta)\,\lambda\, y_0\, y_1\, p_{n-2}(y_2, \ldots, y_n)\\
&\quad + (1/2)^{n+1}\,(1-2\delta)^2\, y_0 \sum_{l=2}^{n} 2^{n-l}\, \lambda^l\, y_l\, p_{n-1-l}(y_{l+1}, \ldots, y_n).
\end{aligned}$$

Now we take out the factor $y_0 y_1 \lambda/2$ from the second and fourth term, noting that $y_1 y_1 = 1$:

$$\begin{aligned}
&p_n(y_0, \ldots, y_n)\\
&= \frac{1}{2}\, p_{n-1}(y_1, \ldots, y_n) - \delta(1-\delta)\,\lambda\, y_0\, y_1\, p_{n-2}(y_2, \ldots, y_n)\\
&\quad + y_0 y_1 \frac{\lambda}{2}\, \Bigg[\frac{1}{2}\, p_{n-2}(y_2, \ldots, y_n)\\
&\quad + (1/2)^n\,(1-2\delta)^2\, y_1 \sum_{l=2}^{n} 2^{n-l}\, \lambda^{l-1}\, y_l\, p_{n-1-l}(y_{l+1}, \ldots, y_n)\Bigg].
\end{aligned}$$

By (A.8.1) the part between the large braclets is just $p_{n-1}(y_1, \ldots, y_n)$. This gives

$$p_n(y_0, \ldots, y_n) = \frac{\lambda\, y_0\, y_1 + 1}{2}\, p_{n-1}(y_1, \ldots, y_n) - \delta(1-\delta)\,\lambda\, y_0\, y_1\, p_{n-2}(y_2, \ldots, y_n),$$

which was to be proved. □

It remains to prove Lemma A.3.

Proof of Lemma A.3. As shown in the proof of Proposition 2.5 we have

$$p_n(y_0, \ldots, y_n) = \sum_{X_0, X_1, \ldots, X_n} \mathbb{P}[X_n] \prod_{i=0}^{n-1} \mathbb{P}[X_i | X_{i+1}] \prod_{i=0}^{n} \mathbb{P}[Y_i | X_i].$$

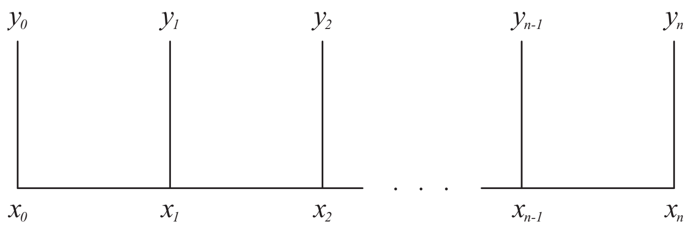


Figure A.1: Graphical interpretation of the expansion of the product in (A.8.2). An edge between x_i and y_i represents the term $(1-2\delta)x_iy_i$ and an edge between x_i and x_{i+1} the term $(1-2p)x_ix_{i+1}$.

For the binary symmetric model, we can write

$$\mathbb{P}[X_i = x_i \mid X_{i+1} = x_{i+1}] = \frac{1}{2} + \frac{1}{2}(1-2p)x_ix_{i+1},$$
$$\mathbb{P}[Y_i = y_i \mid X_i = x_i] = \frac{1}{2} + \frac{1}{2}(1-2\delta)x_iy_i,$$

which can be easily checked by distinguishing whether $x_i = x_{i+1}$ or $x_i \neq x_{i+1}$, respectively whether $x_i = y_i$ or $x_i \neq y_i$. As $\mathbb{P}[X_n] = 1/2$, this gives

$$\begin{aligned} p_n(y_0,\dots,y_n) &= \sum_{X_0,X_1,\dots,X_n} \frac{1}{2}\prod_{i=0}^{n-1}\left(\frac{1}{2}+\frac{1}{2}(1-2p)x_ix_{i+1}\right)\prod_{i=0}^{n}\left(\frac{1}{2}+\frac{1}{2}(1-2\delta)x_iy_i\right) \\ &= \frac{1}{2^{2n+2}}\sum_{X_0,X_1,\dots,X_n}\prod_{i=0}^{n-1}(1+(1-2p)x_ix_{i+1})\prod_{i=0}^{n}(1+(1-2\delta)x_iy_i)\,. \end{aligned} \tag{A.8.2}$$

Writing out the product behind the summation sign gives all possible combinations of choosing from each term either the 1 or the $1+(1-2p)x_ix_{i+1}$ c.q. $1+(1-2\delta)x_iy_i$. We can give an interpretation of this using Figure A.1. Writing out the product gives exactly all possible combinations of subsets of the edges. Here an edge between x_i and y_i represents the term $(1-2\delta)x_iy_i$ and an edge between x_i and x_{i+1} represents the term $(1-2p)x_ix_{i+1}$. This gives the sum over 2^{2n+1} terms. But, as the summation is over $x_i \in \{-1,1\}$ most of the terms will cancel out. To be more precise, this will happen to all terms that contain a factor x_i an odd number of times, or actually exactly once as $x_ix_i = 1$. Left over are the terms where each x_i, $i = 0,\dots,n$ is included an even number of times and thus has disappeared. Now the terms that are left over are all possible combinations of 'U–forms' in the figure. For a given n there are $1,2,\dots,\lceil\frac{n+1}{2}\rceil$ U–forms possible, so the number of terms left over after the summation is equal to

$$\sum_{j=1}^{\lceil\frac{n+1}{2}\rceil}\binom{n+1}{2j}.$$

Now we want to express $p_n(y_0,\dots,y_n)$ in terms of $p_{n-1}(y_1,\dots,y_n), p_{n-2}(y_2,\dots,y_n),\dots,p_0(y_n)$. For this observe that $p_n(y_0,\dots,y_n)$ consists of the U–forms that do contain y_0 and that do not contain y_0. The latter ones are equal to the term $p_{n-1}(y_1,\dots,y_n)$. For the others we have that y_0 is connected to y_l for some $l \in 1,\dots,n$. If it is connected to y_l, this gives the term

$(1-2\delta)^2\,(1-2p)^l\,y_0 y_l$, and the possibilities left over for the other U–forms are equal to that of $p_{n-1-l}(y_{l+1},\dots,y_n)$. Multiplying by $\frac{1}{2}$ the correct number of times gives the desired result:

$$\begin{aligned} p_n(y_0,\dots,y_n) &= \frac{1}{2}\,p_{n-1}(y_1,\dots,y_n) \\ &\quad + \frac{(1-2\delta)^2}{2^{n+1}}\,y_0 \sum_{l=1}^{n} 2^{n-l}\,(1-2p)^l\,y_l\,p_{n-1-l}(y_{l+1},\dots,y_n). \end{aligned}$$ □

A.9 Lemma 5.3

Proof of Lemma 5.3. Assuming $p \in (0, 1/2)$ we will prove that for all x such that $W_L \le x \le W_U$:

$$f_{-1}(x) \le f_1(x).$$

So applying f_{-1} will always give a smaller result than applying f_1. From this it follows that the smallest result is achieved by repeatedly applying f_{-1}, and the largest by repeatedly applying f_1. We have

$$f_{-1}(x) - f_1(x) = -(1-2p) + \frac{\delta(1-\delta)(1-2p)}{x(1-x)}.$$

This is maximal for $x(1-x)$ closest to zero, so for $x \in \{W_L, W_U\}$, as $W_U = 1 - W_L$. This gives

$$\begin{aligned} f_{-1}(x) - f_1(x) &\le -(1-2p) + \frac{\delta(1-\delta)(1-2p)}{W_L(1-W_L)} \\ &= (1-2p)\left(\frac{\delta(1-\delta)}{W_L(1-W_L)} - 1\right). \end{aligned}$$

Note that W_L depends on δ and p, see 5.5.1. It is straightforward but tedious work to check that $\delta(1-\delta)/(W_L(1-W_L))$ is maximal for $\delta = 1/2$, and then equals 1 for all p. Using this and noting that $1-2p$ is positive, we find

$$f_{-1}(x) - f_1(x) \le (1-2p)\left(\frac{\delta(1-\delta)}{W_L(1-W_L)} - 1\right) \le 0.$$

One could easily check that for $\delta \ne 1/2$ we have strict inequality. So for all $\delta \in [0,1]\backslash\{1/2\}$ and for all $p \in (0, 1/2)$ it holds that

$$f_{-1}(x) - f_1(x) < 0.$$ □

A.10 Proposition 5.6

In this section we prove Proposition 5.6, which stated that the derivatives of f_{-1} and f_1 in W_L respectively W_U are in $(0,1)$.

Proof of Proposition 5.6. It is straightforward that both derivatives are strictly larger than 0, as both f_{-1} and f_1 are increasing:

$$\frac{\partial f_{-1}(x)}{\partial x} = \frac{\delta(1-\delta)(1-2p)}{(1-x)^2} > 0, \qquad \frac{\partial f_1(x)}{\partial x} = \frac{\delta(1-\delta)(1-2p)}{x^2} > 0.$$

For the derivative of f_{-1} in W_L we have

$$\left.\frac{\partial f_{-1}(x)}{\partial x}\right|_{x=W_L} = \frac{\delta(1-\delta)(1-2p)}{(1-W_L)^2}.$$

Note that W_L depends on p and δ, see (5.5.1). It is straightforward to check that the right-hand side of the equation is maximal for $p = 0$, and

$$\left.\frac{\partial f_{-1}(x)}{\partial x}\right|_{x=W_L} \leq \frac{4\delta(1-\delta)}{(1+|1-2\delta|)^2} = \begin{cases} \delta/(1-\delta) & \text{if } \delta \in (0, 1/2], \\ (1-\delta)/\delta & \text{if } \delta \in [1/2, 1), \end{cases}$$

with strict inequality when $p > 0$. Both cases evaluate to be smaller or equal to 1 on the indicated domain of δ, so for $p > 0$:

$$\left.\frac{\partial f_{-1}(x)}{\partial x}\right|_{x=W_L} < 1.$$

By symmetry it follows that

$$\left.\frac{\partial f_{1}(x)}{\partial x}\right|_{x=W_U} < 1.$$

□

Appendix B

Series Expansion

In this appendix we give some of the derivations and tables of coefficients which were left out of Chapter 4.

B.1 Function of $\mathbb{P}[Y_0 \mid Y_1, \ldots, Y_n]$

As the series expansion of $\mathbb{P}[Y_0 \mid Y_1, \ldots, Y_n]$ did not lead us to a general form for its coefficients, we try the logarithm of this probability and consider the series expansion of that.

In general

$$\log\left(\sum_{k=0}^{\infty} c_k x^k\right) = \log(c_0) + \frac{c_1}{c_0}\,x + \frac{2c_0c_2 - c_1^2}{2c_0^2}\,x^2 + \frac{c_1^3 - 3c_0c_2c_1 + 3c_0^2c_3}{3c_0^3}\,x^3 + O(x^4).$$

Let

$$\log\left(\mathbb{P}[Y_0 \mid Y_1, \ldots, Y_n]\right) = \sum_{k=0}^{\infty} \tilde{F}_k^{(n)} \delta^k,$$

then

$$\begin{aligned}
&\log\left(\mathbb{P}[Y_0 \mid Y_1, \ldots, Y_n]\right)\\
&= \log\left(\frac{\mathbb{P}[Y_0, Y_1, \ldots, Y_n]}{\mathbb{P}[Y_1, \ldots, Y_n]}\right)\\
&= \log\left(\sum_{k=0}^{n+1} a_k^{(n)}(y_0, \ldots, y_n)\,\delta^k\right) - \log\left(\sum_{k=0}^{n} b_k^{(n)}(y_1, \ldots, y_n)\,\delta^k\right)\\
&= \log a_0 + \frac{a_1}{a_0}\delta + \frac{2a_0a_2 - a_1^2}{2a_0^2}\delta^2 + O(\delta^3)\\
&\qquad - \log b_0 - \frac{b_1}{b_0}\delta - \frac{2b_0b_2 - b_1^2}{2b_0^2}\delta^2 + O(\delta^3)\\
&= \underbrace{\log\left(\frac{a_0}{b_0}\right)}_{\tilde{F}_0} + \underbrace{\left(\frac{a_1}{a_0} - \frac{b_1}{b_0}\right)}_{\tilde{F}_1}\delta + \underbrace{\frac{1}{2}\left(\frac{2a_0a_2 - a_1^2}{2a_0^2} - \frac{2b_0b_2 - b_1^2}{2b_0^2}\right)}_{\tilde{F}_2}\delta^2 + O(\delta^3).
\end{aligned}$$

This shows that only the term $\tilde{F}_0$ involves a logarithmic function. The other $\tilde{F}_k$'s are rational functions. They first three are given by

$$\begin{aligned}
\tilde{F}_0 &= \log\left(\frac{1}{2} + \frac{1}{2}(1-2p)y_0y_1\right), \\
\tilde{F}_1 &= \frac{-\lambda y_0 y_1}{\left(\frac{1}{2}+\frac{1}{2}\lambda y_0 y_1\right)\left(\frac{1}{2}+\frac{1}{2}\lambda y_1 y_2\right)}, \\
\tilde{F}_2 &= \frac{4\lambda y_0 y_1\left(y_0 y_1^2 y_2^2 y_3 \lambda^3 + y_1 y_2(y_2 y_3 - y_0(3y_1+y_3))\lambda^2 + (y_2(y_3-3y_1) - y_0 y_1)\lambda + 1\right)}{(\lambda y_0 y_1+1)^2(\lambda y_1 y_2+1)^2(\lambda y_2 y_3+1)}.
\end{aligned}$$

We calculated many more terms. Unfortunately also for these coefficients were are not able to spot that much structure that we can give a general form for them.

B.2 Settlement $F_1^{(n)}$

In Section 4.3.4 the settlement of $F_0^{(n)}$ for $n \geq 1 = k+1$ was shown. In this appendix we do the same for $F_1^{(n)}$. To do this, by (4.3.7) we should show the settlement of:

$$F_1^{(n)} = \frac{a_1^{(n)} b_0^{(n)} - a_0^{(n)} b_1^{(n)}}{\left(b_0^{(n)}\right)^2},$$

for $n \geq 2 = k+1$, where from (4.3.8):

$$\begin{aligned}
a_1^{(n)} &= -(n+1)\, c_0(p\,; n\,; y_0, \ldots, y_n) + c_1(p\,; n\,; y_0, \ldots, y_n), \\
b_1^{(n)} &= -n\, c_0(p\,; n-1\,; y_1, \ldots, y_n) + c_1(p\,; n-1\,; y_1, \ldots, y_n).
\end{aligned}$$

Note that $c_0(n) = a_0^{(n)}$ and $c_0(n-1) = b_0^{(n)}$. Now consider $c_1(n)$. Recall that this is the probability of observing the sequence $y_0, \ldots, y_n$ when exactly one y_i is flipped. So, by (4.3.3) this is:

$$\begin{aligned}
c_1(n) = \; & \mathbb{P}[X_0 = \bar{y_0}, X_1 = y_1, \ldots, X_n = y_n] \\
& + \mathbb{P}[X_0 = y_0, X_1 = \bar{y_1}, \ldots, X_n = y_n] \\
& \ldots \\
& + \mathbb{P}[X_0 = y_0, X_1 = y_1, \ldots, X_n = \bar{y_n}],
\end{aligned}$$

where $\bar{y}_i = -y_i$. We can write this as the product of terms $1 \pm \lambda_i$. When y_0 is flipped, this gives that λ_0 comes with a minus sign, but when y_1 is flipped, both λ_0 and λ_1 come with a minus sign. This gives the following structure for the minus signs:

λ_0	λ_1	λ_2	$\ldots$	λ_{n-2}	λ_{n-1}
$-$					
$-$	$-$				
	$-$	$-$			
			$\ddots$	$\ddots$	
				$-$	$-$
					$-$

This gives

$$\begin{aligned} c_1(n) = \frac{1}{2^{n+1}} \Big(& (1-\lambda_0)\,(1+\lambda_1)\dots(1+\lambda_{n-1}) \\ & + (1-\lambda_0)\,(1-\lambda_1)\dots(1+\lambda_{n-1}) \\ & + \quad \dots \\ & + (1+\lambda_0)\,(1+\lambda_1)\dots(1-\lambda_{n-1}) \Big). \end{aligned}$$

Along the same lines we can derive

$$\begin{aligned} c_1(n-1) = \frac{1}{2^{n}} \Big(& (1-\lambda_1)\,(1+\lambda_2)\dots(1+\lambda_{n-1}) \\ & + (1-\lambda_1)\,(1-\lambda_2)\dots(1+\lambda_{n-1}) \\ & + \dots \\ & + (1+\lambda_1)\,(1+\lambda_2)\dots(1-\lambda_{n-1}) \Big). \end{aligned}$$

Recalling that $F_0^{(n)} = a_0^{(n)} / \, b_0^{(n)}$, we now have

$$F_1^{(n)} = \frac{a_1^{(n)} b_0^{(n)} - a_0^{(n)} b_1^{(n)}}{\left(b_0^{(n)}\right)^2} = \frac{a_1^{(n)}}{b_0^{(n)}} - F_0^{(n)} \frac{b_1^{(n)}}{b_0^{(n)}}.$$

We calculate both terms in the right-hand side separately. For the second one we have

$$\begin{aligned} F_0^{(0)} \frac{b_1^{(0)}}{b_0^{(0)}} &= 0, \qquad F_0^{(1)} \frac{b_1^{(1)}}{b_0^{(1)}} = 0, \\ F_0^{(n)} \frac{b_1^{(n)}}{b_0^{(n)}} &= -nF_0 + \frac{(1+\lambda_0)(1-\lambda_1)}{2(1+\lambda_1)} \\ & \quad + \frac{1+\lambda_0}{2} \left(\frac{(1-\lambda_1)(1-\lambda_2)}{(1+\lambda_1)(1+\lambda_2)} + \dots + \frac{(1-\lambda_{n-2})(1-\lambda_{n-1})}{(1+\lambda_{n-2})(1+\lambda_{n-1})} + \frac{(1-\lambda_{n-1})}{(1+\lambda_{n-1})} \right), \end{aligned}$$

for $n \geq 2$. The first one is

$$\begin{aligned} \frac{a_1^{(0)}}{b_0^{(0)}} &= 0, \qquad \frac{a_1^{(1)}}{b_0^{(1)}} = -2\lambda_0, \\ \frac{a_1^{(n)}}{b_0^{(n)}} &= -(n+1)F_0 + \frac{1-\lambda_0}{2} \left(1 + \frac{1-\lambda_1}{1+\lambda_1} \right) \\ & \quad + \frac{1+\lambda_0}{2} \left(\frac{(1-\lambda_1)(1-\lambda_2)}{(1+\lambda_1)(1+\lambda_2)} + \dots + \frac{(1-\lambda_{n-2})(1-\lambda_{n-1})}{(1+\lambda_{n-2})(1+\lambda_{n-1})} + \frac{(1-\lambda_{n-1})}{(1+\lambda_{n-1})} \right). \end{aligned}$$

for $n \geq 2$.

Note that the second lines of both are the same and they will cancel out, as well as the term nF_0, as we calculate the difference of the two. So we find:

$$\begin{aligned}
F_1^{(0)} &= 0,\\
F_1^{(1)} &= -2\lambda_0,\\
F_1^{(n)} &= \frac{a_1^{(n)}}{b_0^{(n)}} - F_0^{(n)}\frac{b_1^{(n)}}{b_0^{(n)}}\\
&= -F_0 + \frac{1-\lambda_0}{2}\left(1+\frac{1-\lambda_1}{1+\lambda_1}\right) - \frac{(1+\lambda_0)(1-\lambda_1)}{2(1+\lambda_1)}.\\
&= \frac{-2\lambda_0}{1+\lambda_1} = F_1, \quad \text{for } n \geq 2.
\end{aligned}$$

This shows the settlement of $F_1^{(n)}$ for $n \geq 2 = k+1$. Note that $F_1 = F_1(p\,; y_0, y_1, y_2)$.

B.3 Coefficients g_k

In Section 4.4 we gave the series expansion

$$\mathbb{P}[Y_0 \mid Y_1, \ldots, Y_n] = (1-\delta)\sum_{k=0}^{\infty} g_k^{(n)}\xi^k.$$

Here we will give the first four coefficients expressed in products of y_i, and we give the first three coefficients for the series expansion for the logarithm of this probability.

B.3.1 Coefficients in products of y_i

We want to express the g_k as a linear combination of products of y_0 and y_i's:

$$g_k = c(p) + \sum f_{i_1,\ldots,i_m}(p) \cdot y_0 y_{i_1} \ldots y_{i_m},$$

where $c(p)$ and $f_{i_1,\ldots,i_m}(p)$ are some functions of p. Doing this, we find that for a given k all f have the same denominator, which is, for $k \geq 1$:

$$2^k(p-1)^{2k-1}p^{2k-1}.$$

For better readability, we write this on the left-hand side. Note that only g_0 and g_1 have a constant term. The coefficients are given by:

$$g_0 = \frac{1}{2} + \frac{1}{2}(1-2p)y_0y_1,$$

$$\begin{aligned}
2(p-1)p \cdot g_1 = \frac{1}{2} &- y_0y_1 \cdot (2p-1)\left(p^2-p+1\right)\\
&- y_0y_2 \cdot (2p-1)^2,
\end{aligned}$$

$$\begin{aligned}
g_2 \cdot (4(p-1)^3p^3) = {} & y_0y_1 \cdot (2p-1)\left(2p^4 - 4p^3 + 6p^2 - 4p + 1\right) \\
& + y_0y_2 \cdot (2p-1)^2\left(2p^4 - 4p^3 + 4p^2 - 2p + 1\right) \\
& + y_0y_3 \cdot (2p-1)^3\left(2p^2 - 2p + 1\right) \\
& + y_0y_1y_2y_3 \cdot (2p-1)^4,
\end{aligned}$$

$$\begin{aligned}
& g_3 \cdot (8(p-1)^5p^5) = \\
& \quad - y_0y_1 \cdot (2p-1)\left(4p^8 - 16p^7 + 96p^6 - 232p^5 + 294p^4 - 220p^3 + 100p^2 - 26p + 3\right) \\
& \quad - y_0y_2 \cdot (2p-1)^2\left(4p^8 - 16p^7 + 52p^6 - 100p^5 + 130p^4 - 112p^3 + 62p^2 - 20p + 3\right) \\
& \quad - y_0y_3 \cdot (2p-1)^3\left(12p^6 - 36p^5 + 54p^4 - 48p^3 + 32p^2 - 14p + 3\right) \\
& \quad - y_0y_4 \cdot (2p-1)^4\left(2p^2 - 2p + 1\right)^2 \\
& \\
& \quad - y_0y_1y_2y_3 \cdot (2p-1)^4\left(6p^4 - 12p^3 + 14p^2 - 8p + 3\right) \\
& \quad - y_0y_1y_2y_4 \cdot (2p-1)^5\left(2p^2 - 2p + 1\right) \\
& \quad - y_0y_1y_3y_4 \cdot (2p-1)^6 \\
& \quad - y_0y_2y_3y_4 \cdot (2p-1)^5\left(2p^2 - 2p + 1\right).
\end{aligned}$$

We calculated many more coefficients with the aim to find a general form for them. Unfortunately we did not succeed in this. However, we can make a number of observations. First of all, there are no terms with an odd number of y_i multiplied. We have that $1-2p = (1-p)^2 - p^2$ and $2p^2 - 2p + 1 = (1-p)^2 + p^2$. For the terms with y_0y_i the power of the term $(1-2p)$ is equal to i. If y_{n+1} is part of the product of the coefficient g_n, then f is given by

$$\pm((1-p)^2 - p^2)^{2n-j}((1-p)^2 + p^2)^j$$

for some j, but we are not able to determine an expression for this j.

B.3.2 Coefficients for log

We also look at $\log \mathbb{P}[Y_0 \mid Y_1, \ldots, Y_n]$ expanded in ξ:

$$\begin{aligned}
\log\ \mathbb{P}[Y_0 \mid Y_1, \ldots, Y_n] &= \log\left((1-\delta) \cdot \sum_{k=0}^{\infty} g_k^{(n)} \xi^k\right) \\
&= \log(1-\delta) + \sum_{k=0}^{\infty} G_k^{(n)} \xi^k,
\end{aligned}$$

where $G_k^{(n)} = G_k$ for $n \geq k+1$, and $G_k = G_k(p; y_0, \ldots, y_{k+1})$. It turns out that it now is not possible any more to write G_k as a linear combination of products of y_0 and y_i's. Also products are included which do not contain the factor y_0.

$$\begin{aligned}
G_0 = {} & \frac{1}{2}(\log(1-p) + \log(p)) \\
& + y_0y_1 \cdot \frac{1}{2}(\log(1-p) - \log(p)),
\end{aligned}$$

$$
\begin{aligned}
G_1\left(4(p-1)^2p^2\right) = {} & \left(2p^2-2p+1\right)^2 \\
& + y_0y_1 \cdot (2p-1) \\
& + y_0y_2 \cdot (2p-1)^2 \\
& + y_1y_2 \cdot (2p-1)^3,
\end{aligned}
$$

$$
\begin{aligned}
G_2\left(4(p-1)^4p^4\right) = {} & (-2p^8+8p^7-28p^6+56p^5-70p^4+56p^3-28p^2+8p-1) \\
& - y_0y_1 \cdot (2p-1)\left(2p^2-2p+1\right)\left(3p^2-3p+1\right) \\
& - y_0y_2 \cdot (2p-1)^2\left(p^2-p+1\right)\left(2p^2-2p+1\right) \\
& - y_0y_3 \cdot \frac{1}{2}(2p-1)^3\left(2p^2-2p+1\right) \\
& - y_1y_2 \cdot (2p-1)^3\left(2p^4-4p^3+4p^2-2p+1\right) \\
& - y_1y_3 \cdot \frac{1}{2}(2p-1)^4\left(2p^2-2p+1\right) \\
& - y_2y_3 \cdot \frac{1}{2}(2p-1)^5 \\
& - y_0y_1y_2y_3 \cdot \frac{1}{2}(2p-1)^4.
\end{aligned}
$$

Again we see some structure, but this does not give us a general form for the coefficients either.

Appendix C

Coefficients power series expansion h_Y in ζ

The first ten coefficients of the power series expansion

$$h_Y = \sum_{k=0}^{\infty} h_{Y,k} \cdot \zeta^k,$$

where $h_{Y,k} = h_{Y,k}(p)$, are given by:

$$\begin{aligned}
h_{Y,0} &= -(1-p)\log(1-p) - p\log(p),\\
h_{Y,1} &= 2(1-2p)\log\left(\frac{1-p}{p}\right),\\
h_{Y,2} &= -\frac{(1-2p)^2}{2(1-p)^2p^2},\\
h_{Y,3} &= -\frac{(1-2p)^4\left(4p^2-4p-1\right)}{6(1-p)^4p^4},\\
h_{Y,4} &= \frac{(1-2p)^4\left(32p^6-96p^5+145p^4-130p^3+57p^2-8p-1\right)}{12(1-p)^6p^6},\\
h_{Y,5} &= -\frac{(1-2p)^6\left(56p^6-168p^5+268p^4-256p^3+124p^2-24p-1\right)}{20(1-p)^8p^8},\\
h_{Y,6} &= -\Big((1-2p)^6\big(464p^{10}-2320p^9+4770p^8-5160p^7+2436p^6+1008p^5-2250p^4\\
&\qquad +1440p^3-440p^2+52p+1\big)\Big)/(30(1-p)^{10}p^{10}),\\
h_{Y,7} &= \Big((1-2p)^8\big(448p^{12}-2688p^{11}+9512p^{10}-22920p^9+36943p^8-39820p^7+27792p^6\\
&\qquad -10702p^5+330p^4+1776p^3-787p^2+116p+1\big)\Big)/(42(1-p)^{12}p^{12}),
\end{aligned}$$

$$\begin{aligned}
h_{Y,8} = -\Big(&(1-2p)^8\big(30336p^{14} - 212352p^{13} + 777777p^{12} - 1906086p^{11} + 3330383p^{10} \\
&- 4240516p^9 + 3952433p^8 - 2657486p^7 + 1230229p^6 - 342608p^5 + 25403p^4 \\
&+ 18326p^3 - 6559p^2 + 720p + 3\big)\Big)/(168(1-p)^{14}p^{14}), \\
h_{Y,9} = -\Big(&(1-2p)^{10}\big(3072p^{16} - 24576p^{15} + 52912p^{14} + 59696p^{13} - 631512p^{12} + 1894816p^{11} \\
&- 3520624p^{10} + 4585368p^9 - 4349324p^8 + 3017504p^7 - 1497304p^6 + 498216p^5 \\
&- 91824p^4 + 712p^3 + 3364p^2 - 496p - 1\big)\Big)/(72(1-p)^{16}p^{16}).
\end{aligned}$$

References

[1] M. Abramowitz and I.A. Stegun, editors. *Handbook of mathematical functions with formulas, graphs, and mathematical tables.* Dover New York, 1965.

[2] T.M. Apostol. *Mathematical analysis: A modern approach to advanced calculus.* Addison-Wesley, Reading, 1957.

[3] L.E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, 37(6):1554–1563, 1966.

[4] L.E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, 41(1):164–171, 1970.

[5] J.J. Birch. Approximations for the entropy for functions of Markov chains. *The Annals of Mathematical Statistics*, 33(3):930–938, 1962.

[6] J.J. Birch. On information rates for finite-state channels. *Information and Control*, 6:372–380, 1963.

[7] C. Couvreur. Hidden Markov models and their mixtures. *Département de mathématique, Université Catholique de Louvain, Louvain, Belgium*, 1996.

[8] T.M. Cover and J.A. Thomas. *Elements of information theory.* Wiley, New York, 1991.

[9] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[10] W. Doeblin. Exposé de la théorie des chaînes simples constantes de Markov à un nombre fini d'états. *Revue Mathematique de l'Union Interbalkanique*, 2:77–105, 1938.

[11] R. Durbin, A. Krogh, G. Mitchison, and S.R. Eddy. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids.* Cambridge University Press, 1998.

[12] Y. Ephraim and N. Merhav. Hidden Markov processes. *IEEE Transactions on Information Theory*, 48(6):1518–1569, 2002.

[13] M. Fekete. Über die Verteilung der Wurzeln bei gewissen algebraischen Gleichungen mit ganzzahligen Koeffizienten. *Mathematische Zeitschrift*, 17(1):228–249, 1923.

[14] J.D. Ferguson. Hidden Markov analysis: An introduction. *Proceedings of the Symposium on the Applications of Hidden Markov Models to Text and Speech*, pages 8–15, 1980.

[15] R. Fernández, P.A. Ferrari, and A. Galves. *Coupling, renewal and perfect simulation of chains of infinite order*. Lecture Notes for the Vth Brazilian school of Probability, Ubatuba, August, 2001.

[16] G.D. Forney Jr. The Viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973.

[17] I.S. Gradshteyn and I.M. Ryzhik. *Table of integrals, series, and products.* Academic Press, New York, 1980.

[18] D. Griffeath. A maximal coupling for Markov chains. *Probability Theory and Related Fields*, 31(2):95–106, 1975.

[19] G. Han and B. Marcus. Analyticity of entropy rate in families of hidden Markov chains. *Proceedings of International Symposium on Information Theory*, pages 2193–2197, 2005.

[20] G. Han and B. Marcus. Analyticity of entropy rate of hidden Markov chains. *Arxiv preprint math/0507235*, 2006. Submitted to *IEEE Transactions on Information Theory.*

[21] G.H. Hardy, J.E. Littlewood, and G. Pólya. *Inequalities.* Cambridge University Press, 1988.

[22] T.E. Harris. On chains of infinite order. *Pacific Journal of Mathematics*, 5:707–724, 1955.

[23] P. Jacquet, G. Seroussi, and W. Szpankowski. On the entropy of a hidden Markov process. *Proceedings of Data Compression Conference*, pages 362–371, 2004.

[24] B.H. Juang and L.R. Rabiner. The segmental K-means algorithm for estimating parameters of hidden Markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(9):1639–1641, 1990.

[25] M. Keane. Strongly mixing g-measures. *Inventiones Mathematicae*, 16(4):309–324, 1972.

[26] T. Koski. *Hidden Markov models for bioinformatics.* Kluwer Academic Publishers, 2001.

[27] T. Lindvall. *Lecture on the coupling method.* Wiley, New York, 1992.

[28] I.L. MacDonald. *Hidden Markov and other models for discrete-valued time series.* CRC Press, 1997.

[29] T. Nattermann. Theory of the random field Ising model. In A.P. Young, editor, *Spin glasses and random fields.* World Scientific Singapore, 1998.

[30] E. Ordentlich and T. Weissman. New bounds on the entropy rate of hidden Markov processes. *IEEE Information Theory Workshop*, pages 117–122, 2004.

[31] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[32] C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(I):379–423, 1948.

[33] C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(I):623–656, 1948.

[34] P.C. Shields. *The ergodic theory of discrete sample paths.* American Mathematical Society, 1996.

[35] H. Thorisson. *Coupling, stationarity, and regeneration*. Springer Verlag, 2000.

[36] L.N. Vasershtein. Markov processes on countable product space describing large systems of automata. *Problemy Peredachi Informatsii*, 5:64–73, 1969.

[37] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, 13(2):260–269, 1967.

[38] O. Zuk. The relative entropy rate for two hidden Markov processes. *ITG Munich*, 2006.

[39] O. Zuk, E. Domany, I. Kanter, and M. Aizenman. From finite-system entropy to entropy rate for a hidden Markov process. *IEEE Signal Processing Letters*, 13:517–520, 2006.

[40] O. Zuk, E. Domany, I. Kanter, and M. Aizenman. Taylor series expansions for the entropy rate of hidden Markov processes. *Proceedings of IEEE International Conference on Communications*, 2006.

[41] O. Zuk, I. Kanter, and E. Domany. Asymptotics of the entropy rate for a hidden Markov process. *Proceedings of Data Compression Conference*, pages 173–182, 2005.

[42] O. Zuk, I. Kanter, and E. Domany. The entropy of a binary hidden Markov process. *Journal of Statistical Physics*, 121(3):343–360, 2005.