DIRECTED INFORMATION AND ITS BIOMEDICAL APPLICATIONS

A Thesis

by

QING WAN

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee,	Xiaoning Qian
Committee Members,	Tie Liu
	Ulisses Braga Neto
	David R. Larson
Head of Department,	Miroslav M. Begovic

May 2016

Major Subject: Electrical Engineering

ABSTRACT

The foundation of Information theory by C. E. Shannon is the entropic measures to quantize the abstract concept of information in random variables. When the interest is to study the relationships among multiple random processes, directed information introduced by Massey in 1990 extends such entropic measures to study the direction of information flow among the random processes, thereafter enabling the causality inference.

Theoretically, directed information is defined on joint or conditional probability distributions of two random processes. In practice, with observed time series data, the estimation of directed information will be dependent on accurate estimates of joint probability distribution functions of random processes. We explore existing directed information estimators, generally based on Context Tree Weighting method and general asymptotic equipartition property, to infer the directed interactive relationships among involved components in neural activity mathematically modeled on Poisson process. Outcome validates causality relationship among a set of neurons. Although challenges remain, particularly finding computational efficient way to distinguish indirect relationship from direct interaction, estimation of directed information can effectively dig out causality among network.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Dr. Qian who supervised my thesis. I would like to thank my committee members, Dr. Liu, Dr. Ulisses and Dr. Larson, for their discussions and suggestions on my research.

Thanks also go to my girlfriend who is currently a PhD student in A&M, for her accompany and wisdom in my critical months of doing thesis especially after my defense. I love you forever!

Finally, I would also like to acknowledge my parents for their financial support and encouragement.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

1. INTRODUCTION

With recent advancements of high-throughput high-content sensing and imaging techniques, time series physiological data becomes feasible to help monitor and understand life behavior so that we can translate these data for more cost-effective healthcare. One of critical challenges of analyzing such data is to understand possible interaction among multiple random processes that drive the dynamics of living systems, so that we can better understand "causality" relationship among constituting components in the system of interest.

Information Theory studies transmission, processing, utilization and extraction of information. Information Theory builds upon entropic measures of involved random variables in a system. C. E. Shannon [1], the founder of Information Theory, introduced Shannon Entropy to quantize the average uncertainty of information in random variables.

Directed information, first introduced by Massey [2] in 1990, an extended information measure that evaluates the direction of information flow in channels. When the interest is to study the relationships among multiple stochastic processes, directed information, studies the direction of information flow among random processes, thereafter enable the causality inference [3].

Theoretically, directed information is defined on joint or conditional probability distributions of two random processes. In practice, with observed time series data, the estimation of directed information will be dependent on accurate estimates of joint probability distribution functions of random processes, which imposes unique analytic challenges. The authors in [4], based on the concept of universal probability assignment, introduced estimation of directed information with L_1 convergence. *Jiao* [5] employed the Context-Tree Weighting method [6], originally used for arithmetic encoding for data compression, as a universal probability assignment, to refine theorems in [4] and analyze causal influence in stock markets.

In this thesis, we explore existing directed information estimators to estimate the directed interactive relationships among random processes. We are specifically interested in applying directed information estimates in inferring causality of involved neurons by analyzing spiking activities from a neural network.

2. BACKGROUND

2.1 Notations

In this thesis, any uppercase letter with or without a subscript such as X or X_1 denotes a random variable. The bold uppercase letter such as $\boldsymbol{X}$ denotes a random process. The uppercase letter with a superscript such as X^N denotes a discrete-time random process with the time index from 1 to N, which constitutes a set of random variables $X_1, X_2, \ldots, X_N$. The lower case letter with or without a subscript such as x or x_1 denotes a feasible value of the random variable X or X_1 respectively. Correspondingly, a bold lowercase such as $\boldsymbol{x}$ denotes a set of feasible values for an observation of random process $\boldsymbol{X}$ or X^N.

$p(X)$ denotes the corresponding probability mass function (PMF) or probability density function (PDF) of a random variable X, depending on weather the state space is discrete or continuous, respectively. $p(X, Y)$ denotes the corresponding joint probability distribution function of (X, Y). For a discrete random variable X, $p(x)$ denotes the probability of random variable X taking a feasible value x, that is $p(x) = p(X = x)$. Similarly, $p(x, y)$ denotes the probability of the pair of random variables (X, Y) taking values (x, y) that is $p(X = x, Y = y)$.

We focus on discrete random variables and processes in the thesis as we are specifically interested in analyzing spiking signals from neurons, which are often modeled as binary sequences in [7].

2.2 Definitions

2.2.1 Entropy [8]

The Shannon ***entropy*** *H(X)* of a discrete random variable *X* is defined by

$$H(X) \triangleq -\sum_{x \in \mathcal{X}} p(x) \log p(x)$$

where x denotes a feasible value that *X* can take and $\mathcal{X}$ denotes the *alphabet set* (state space) of *X*.

2.2.2 Conditional Entropy [8]

The ***conditional entropy*** $H(Y|X)$ of two discrete random variables *X*, *Y* is defined as

$$H(Y|X) \triangleq \sum_{x \in \chi} p(x) H(Y \mid X = x)$$

$$= -\sum_{x \in \chi} p(x) \sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(y|x)$$

$$= -E[\log p(y|x)]$$

where $p(y|x) = \frac{p(x,y)}{p(x)}$ and the expectation $E[\,]$ is taken with respect to the joint probability distribution $p(X,Y)$.

2.2.3 Relative Entropy [8]

The ***relative entropy*** or ***Kullback-Leibler distance*** between two probability mass function *p(X)* and *q(X)* is defined as

$$D(p \,||\, q) \triangleq \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = E_p \log \frac{p(x)}{q(x)}$$

.

2.2.4 Mutual Information [8]

The ***mutual information*** $I(X;Y)$ of two random variables X and Y

$$I(X;Y) \triangleq \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

. Mutual information is always nonnegative and achieve zero if and only if X and Y are independent [8, Corollary 2.6.3].

2.2.5 Conditional mutual information [8]

For random variables X and Y given Z, ***conditional mutual information*** is defined as

$$I(X;Y \mid Z) = H(X|Z) - H(X|Y,Z)$$

$$= E_{p(x,y,z)} \log \frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)}$$

. Straightforward from the definition, conditional mutual information is nonnegative as mutual information [8].

2.2.6 Directed Information [2]

The ***directed information*** $I(X^N \rightarrow Y^N)$ from a random sequence X^N to a random sequence Y^N is defined by

$$I(X^N \to Y^N) \triangleq \sum_{n=1}^{N} I(X^n; Y_n \mid Y^{n-1})$$

. Here $I(X^n; Y_n \mid Y^{n-1})$ is the conditional mutual information defined by

$$I(X^n; Y_n \mid Y^{n-1}) \triangleq E_{p(X^n, Y_n, Y^{n-1})} \log \frac{p(X^n, Y_n \mid Y^{n-1})}{p(X^n \mid Y^{n-1}) p(Y_n \mid Y^{n-1})}$$

$$= \sum_{x^n \in \mathcal{X}^n} \sum_{y^{n-1} \in \mathcal{Y}^{n-1}} \sum_{y_n \in \mathcal{Y}} \log \frac{p(X^n, Y_n \mid Y^{n-1})}{p(X^n \mid Y^{n-1}) p(Y_n \mid Y^{n-1})}$$

. Obviously, $I(X^N \to Y^N)$ is nonnegative.

2.2.7 Causally conditional probability [9]

The ***causally conditional probability*** is defined between two random sequences X^n taking feasible sequence values x^n and Y^n taking feasible sequence values y^n that have equal lengths

$$p(y^n \mid\mid x^n) \triangleq \prod_{i=1}^{n} p\left(y_i \mid y^{i-1}, x^i\right)$$

$$p(y^n \mid\mid x^{n-1}) \triangleq \prod_{i=1}^{n} p\left(y_i \mid y^{i-1}, x^{i-1}\right)$$

.

2.2.8 Causally conditional entropy [9]

The ***causal conditional entropy*** for two random sequences Y^n and X^n is defined as

$$H(Y^n \mid\mid X^n) \triangleq \sum_{i=1}^{n} H\left(Y_i \mid Y^{i-1}, X^i\right)$$

$$= \sum_{i=1}^{n} \sum_{x^i, y^i} -p\left(y_i, y^{i-1}, x^i\right) \log\left(y_i | y^{i-1}, x^i\right)$$

$$= \sum_{i=1}^{n} \sum_{x^i, y^i} -p\left(y^i, x^i\right) \log\left(y_i | y^{i-1}, x^i\right)$$

.

2.2.9 Entropy of a discrete-time stochastic process [8]

The ***entropy of a discrete-time stochastic*** **process** $\boldsymbol{X} = \{X_n | n \in \mathbb{N}^+\}$ is defined by

$$H(\boldsymbol{X}) \triangleq \lim_{n \to \infty} \frac{1}{n} H(X_1, X_2, \dots, X_n)$$

when the limit exists. And

$$\bar{H}_n(\boldsymbol{X}) \triangleq \frac{1}{n} H(X_1, X_2, \dots, X_n)$$

is the ***entropy rate*** of the stochastic process $\boldsymbol{X}$ til time n.

2.2.10 Directed information and Directed information rate between two-discrete time stochastic processes

The ***directed information*** between two discrete-time stochastic processes $\boldsymbol{X} = \{X_i\}$ and $\boldsymbol{Y} = \{Y_j\}$ ($\boldsymbol{X}$, $\boldsymbol{Y}$ have equal length) is defined as

$$I(\boldsymbol{X} \to \boldsymbol{Y}) \triangleq \lim_{n \to \infty} \frac{1}{n} I(X^n \to Y^n)$$

when the limit exists. And

$$\bar{I}_n(\boldsymbol{X} \to \boldsymbol{Y}) \triangleq \frac{1}{n} I(X^n \to Y^n)$$

is the ***directed information rate*** of these two processes till time n. Apparently, both $I(\boldsymbol{X} \to \boldsymbol{Y})$ and $\bar{I}_n(\boldsymbol{X} \to \boldsymbol{Y})$ are nonnegative.

2.2.11 Causally conditional entropy and causally conditional entropy rate between two discrete-time stochastic processes

The ***causally conditional entropy*** between two discrete-time stochastic processes $\boldsymbol{X} = \{X_i\}$ and $\boldsymbol{Y} = \{Y_j\}$ ($\boldsymbol{X}$, $\boldsymbol{Y}$ have equal length) is defined as

$$H(\boldsymbol{Y}||\boldsymbol{X}) \triangleq \lim_{n\to\infty} \frac{1}{n} H(Y^n||X^n)$$

when the limit exists. And

$$\bar{H}_n(\boldsymbol{Y}||\boldsymbol{X}) = \frac{1}{n} H(Y^n||X^n)$$

is the ***causally conditional entropy rate*** of these two till time n.

2.3 Some properties of directed information

2.3.1 Lemma

Given a pair of random sequences X^N and Y^N,

$$I(X^N \to Y^N) \triangleq \sum_{n=1}^{N} I(X^n; Y_n \mid Y^{n-1}) = H(Y^N) - H(Y^N || X^N)$$

where

$$H(Y^N) - H(Y^N || X^N) = \sum_{n=1}^{N} [H(Y_n|Y^{n-1}) - H(Y_n|X^n, Y^{n-1})]$$

.

Proof:

$I(X^N \to Y^N)$

$= \sum_{n=1}^{N} I(X^n, Y_n | Y^{n-1})$

$= \sum_{n=1}^{N} [H(Y_n|Y^{n-1}) - H(Y_n|X^n, Y^{n-1})]$ (2.61) of [8]

$= \sum_{n=1}^{N} H(Y_n|Y^{n-1}) - \sum_{n=1}^{N} H(Y_n|X^n, Y^{n-1})$ A1.1.1

Since

$\sum_{n=1}^{N} H(Y_n|Y^{n-1}) = H(Y_1, Y_2, Y_3, \dots, Y_N) = H(Y^N)$, Chain rule for entropy (2.48) of [8]

By definition of causal conditional entropy $H(Y^n||X^n) \triangleq \sum_{i=1}^{n} H(Y_i|Y^{i-1}, X^i)$,

A1.1.1 => $I(X^N \to Y^N)$

$= \sum_{n=1}^{N} I(X^n, Y_n|Y^{n-1}) = H(Y^N) - H(Y^N||X^N)$.

2.3.2 Theorem

For two sequences X^N and Y^N, $I(X^N \to Y^N) = 0$ if and only if Y_n is independent to X^m for any $n \geq m, n, m \in \{l \in \mathbb{N} \mid 1 \leq l \leq N\}$ given history Y^0.

Proof:

Backward is obvious. Here try to show forward.

Due to nonnegativity of conditional mutual information [8],

$I(X^N \to Y^N) = 0$

$=> I(X^n; Y_n \mid Y^{n-1}) = 0$, any $n \in \{l \in \mathbb{N} \mid 1 \leq l \leq N\}$

$=>$ X^n is independent of Y_n, given Y^{n-1}, any $n \in \{l \in \mathbb{N} \mid 1 \leq l \leq N\}$

Suppose there exist $i \geq j, i, j \in \{l \in \mathbb{N} \mid 1 \leq l \leq N\}$ such that Y_i is dependent to X_j, then contradiction above, q. e. d.

2.3.3 Lemma

Given $X^0 = x^0$, $Y^0 = y^0$, we have

$$p(y^n, x^n \mid y^0, x^0) = p(y^n || x^n) p(x^n || y^{n-1})$$

.

Proof:

By definition of causal conditional probability,

$$p(y^n \mid\mid x^n) \triangleq \prod_{i=1}^{n} p\left(y_i | y^{i-1}, x^i\right)$$

$$p(x^n \mid\mid y^n) \triangleq \prod_{i=1}^{n} p\left(x_i | x^{i-1}, y^i\right)$$

, will have,

$$p(y^n \mid\mid x^n) p(x^n \mid\mid y^n) \triangleq \prod_{i=1}^{n} p\left(y_i | y^{i-1}, x^i\right) p\left(x_i | x^{i-1}, y^i\right)$$

$$= \prod_{i=1}^{n} \frac{p\left(y^i, x^i\right) p\left(x^i, y^{i-1}\right)}{p(x^i,\ y^{i-1}) p(x^{i-1},\ y^{i-1})} = \prod_{i=1}^{n} p(y_i, x_i | y^{i-1}, x^{i-1}) = p(y_n, x_n | y_0, x_0)$$

.

2.3.4 Theorem

Conservation law [5] of directed information

$$I(X^N, Y^N) = I(X^N \to Y^N) + I(Y^{N-1} \to X^N)$$

where

$$I(Y^{N-1} \to X^N)$$

$$\triangleq I(\emptyset \times Y^{N-1} \to X^N)$$

$$= \sum_{n=1}^{N} [H(X_n|X^{n-1}) - H(X_n|Y^{n-1}, X^{n-1})]$$

where $\times$ denotes Cartesian product.

Proof:

$I(X^N \to Y^N) + I(Y^{N-1} \to X^N)$

$= H(Y^N) - H(Y^N||X^N) + H(X^N) - H(X^N||Y^{N-1})$

$= H(Y^N) + H(X^N) - [H(Y^N||X^N) + H(X^N||Y^{N-1})]$

$= H(Y^N) + H(X^N) - H(X^N, Y^N)$

$= I(X^N, Y^N)$ q. e. d.

3. DIRECTED INFORMATION AND CAUSALITY

Indicating the directivity of information flow is one of the most significant characteristics directed information has. For two random sequences X^N and Y^N, their mutual information is

$$I(X^N;Y^N) = H(Y^N) - H(Y^N|X^N)$$

, compared to their directed information which is

$$I(X^N \to Y^N) = H(Y^N) - H(Y^N||X^N)$$

, we can easily check the difference between causally conditional entropy

$$H(Y^N||X^N) \triangleq \sum_{i=1}^{N} H\left(Y_i|Y^{i-1},X^i\right) = \sum_{i=1}^{N} E_{Y^i,X^i}\left(Y_i|Y^{i-1},X^i\right)$$

and conditional entropy

$$H(Y^N|X^N) = -\sum_{y^N \in y^N, x^N \in \mathcal{X}^N} p(y^N,x^N)\log\frac{p(y^N,x^N)}{p(y^N)\,p(x^N)}$$

is causally conditional entropy reflects the conditional average uncertainty conditioning on *past* and *present* while conditional entropy doesn't have, which make the directivity. In general, $I(X^N \to Y^N) \neq I(Y^N \to X^N)$. 2.3.4 Theorem unveils relationship between directed information and mutual information.

Causality [10] is the relation between one process and another, where the first is understood to be partly responsible for the second. Definition of causality restricts one's past or present has influence on the other's future, not vice versa. Definition of directed

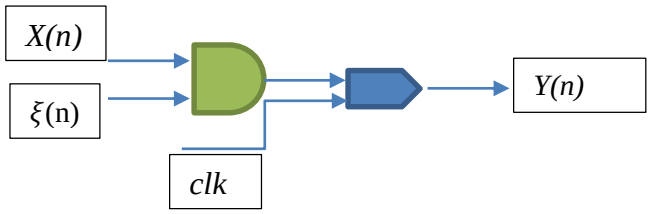
information between two random sequences satisfies this restriction perfectly, which empower directed information to infer causality between them two.

Directed information introduces an information measure, converting average uncertainty of causality into bits (logarithm of 2). 2.3.2 Theorem mathematically proved this directivity of causal influence. By 2.3.2 theorem, two random sequence $X^N$and Y^N, X^N has causal influence on Y^N if $I(X^N \to Y^N) > 0$.

Fig 3.1 demonstrates an example of theoretical analyses of causal influence of input $X(n)$ and output $Y(n)$ of AND gate contaminated by noise with length $N = 3$.

Blue block is D trigger that simulate a delay in channel. Green block is an AND gate. $Y(n+1) = X(n)\ \&\ \xi(n), n \in \{0,1,2,3\}$ and X(n) are i. i. d for $n \in \{0,1,2,3\}$ and so is $\xi(n)$. $\xi(n)$ actually plays a role of binary noise.

Fig 3.1 Illustration of AND gate contaminated by noise



Probability distributions are below.

$\xi(n)$	0	1
P	0.5	0.5

$X(n)$	0	1
P	$\frac{7}{8}$	$\frac{1}{8}$

Initial conditions are $X(0) = 0, \xi(0) = 0, Y(0) = 0$.

Calculate $I(X^3 \to Y^3) = \sum_{n=1}^{N=3}[H(Y_n|Y^{n-1}) - H(Y_n|X^n, Y^{n-1})]$.

1) For n=1

$H(Y_n|Y^{n-1}) - H(Y_n|X^n, Y^{n-1})$

$= H(Y_1|Y^0) - H(Y_1|X_0^1, Y^0) = H(Y_1) - H(Y_1|X_0) = 0$

2) For n=2

$H(Y_n|Y^{n-1}) - H(Y_n|X^n, Y^{n-1})$

$= H(Y_2|Y_0^1) - H(Y_2|X_0^2, Y_0^1) = H(Y_2) - H(Y_2|X_1)$

(Y_2, X_1)	$p(X_1)$	$p(Y_2, X_1)$
(0, 0)	$p(X_1 = 0) = \frac{7}{8}$	$p(Y_2 = 0, X_1 = 0)$ $= p(Y_2 = 0\|X_1 = 0)p(X_1 = 0)$ $= \frac{7}{8}$

(0,1)	$p(X_1 = 1) = \frac{1}{8}$	$p(Y_2 = 0, X_1 = 1)$ $= p(Y_2 = 0 \mid X_1 = 1)p(X_1 = 1)$ $= \frac{1}{2} \times \frac{1}{8} = \frac{1}{16}$
(1,0)	$p(X_1 = 0) = \frac{7}{8}$	$p(Y_2 = 1, X_1 = 0) = 0$
(1,1)	$p(X_1 = 1) = \frac{1}{8}$	$p(Y_2 = 1, X_1 = 1)$ $= p(Y_2 = 1 \mid X_1 = 1)p(X_1 = 1)$ $= \frac{1}{2} \times \frac{1}{8} = \frac{1}{16}$

$H(Y_2) = 0.3373, H(Y_2|X_1) = 2 \times \frac{1}{16} \log_2 2 = \frac{1}{8}$

⇨ $H(Y_2|Y_0^1) - H(Y_2|X_0^2, Y_0^1) = 0.2123$

3) For n=3

$H(Y_n|Y^{n-1}) - H(Y_n|X^n, Y^{n-1})$

$= H(Y_3|Y_0^2) - H(Y_3|X^3, Y_0^2) = H(Y_3) - H(Y_3|X_2)$

$= 0.2123$

4) Combine

$I(X^3 \to Y^3) = 0.4246 > 0;$

Calculate $I(Y^3 \to X^3)$

$$= \sum_{n=1}^{N=3} [H(X_n|X^{n-1}) - H(X_n|Y^n, X^{n-1})] = \sum_{n=1}^{N=3} [H(X_n) - H(X_n)] = 0$$

.

Information flow from X to Y merely. In other words, X has causal influence on Y, not vice versa.

4. ESTIMATION OF DIRECTED INFORMATION

Theoretical definition of directed information comes down to using joint and conditional probability distributions to calculate DI between two random sequences. In practice, with observed time series data, the estimates of directed information are divided into two procedures: 1. model and estimate joint or conditional probabilities of samples; 2. estimate directed information or directed information rate with these estimated probabilities in first step.

4.1 Step1: Modeling for probability estimation

Paper [5] employed Context Tree Weighting [6] method, assigning probabilities sequentially to time series. Mathematical model of CTW is bounded memory tree source that the next-symbol probabilities depend on a finite number of most recent symbols.

4.1.1 Binary Bounded Memory Tree Source Definition

This chapter cites [6] with some modifications.

4.1.1.1 String

A string s is a concatenation of binary symbols, hence $s = q_{1-l}q_{2-l} \dots q_0$ (string from right to left, starting with 0 and going negative) with $q_{-i} \in \{0,1\}$, for $i = 0, 1, \dots, l-1$.

4.1.1.2 Length of string

Length of a string s denotes $l(s)$.

Examples:

Semi-infinite string $s = \cdots q_{-1}q_0$ has length $l(s) = \infty$.

The empty string λ has length $l(\lambda) = 0$.

4.1.1.3 Concatenation of two strings

If there are two strings

$$s' = q'_{1-p}q'_{2-p} \ldots q'_0$$

and

$$s = q_{1-l}q_{2-l} \ldots q_0$$

then

$$s's = q'_{1-p}q'_{2-p} \ldots q'_0q_{1-l}q_{2-l} \ldots q_0$$

is the concatenation of both.

4.1.1.4 Suffix

We say that a string $s = q_{1-l}q_{2-l} \ldots q_0$ is a suffix of string $s' = q'_{1-l}q'_{2-l} \ldots q'_0$ if $l \leq l'$

and $q_{-i} = q'_{-i}$ for $i = 0, 1, \ldots, l-1$.

The empty string λ is a suffix of all strings.

4.1.1.5 Binary tree source

A binary tree source generates a sequence $x_{-\infty}^{\infty}$ of digits assuming values in the alphabet $\{0, 1\}$. $x_{\mathrm{m}}^{\mathrm{n}}$ denotes the sequence $x_m x_{m+1} \ldots x_n$, and allow *m* and *n* to be infinitely large.

For $n < m$ the sequence $x_{\mathrm{m}}^{\mathrm{n}}$ is empty, denoted by $\emptyset$.

4.1.1.6 Suffix set

A suffix set S is defined by $\{s(k)|s(k)\ is\ a\ string, k \in \mathbb{N}^+\}$.

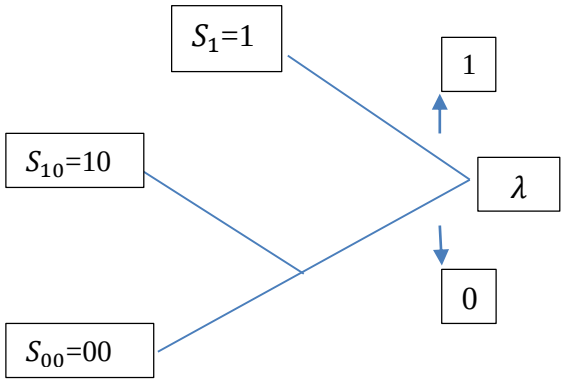
4.1.1.7 Tree model

Each suffix set S (alphabet $\{0, 1\}$) is equivalent to a tree model that λ is its root and each path $s = q_{1-l}q_{2-l} \dots q_0 \in S$ is a node of tree.

Especially, a suffix set S of binary alphabet $\{0, 1\}$ is equivalent to a binary tree model that λ is its root and each path $s = q_{1-l}q_{2-l} \dots q_0 \in S$, q_{1-l} equal to 1 or 0 is a left or right leaf of the tree respectively and others $q_{i-l}, i \in \{2, 3, .., l\}$ equal to 1 or 0 is a left or right node of the tree model respectively.

An example (Fig 4.1) for suffix set $S = \{00, 10, 1\}$

Fig 4.1 An example of suffix set



.

4.1.1.8 Properness of a suffix set

No string in S is a suffix of any other string in S.

4.1.1.9 Completeness of a suffix set

Each semi-infinite sequence (string) $\ldots x_{n-2}x_{n-1}x_n$ has a suffix that belongs to S.

4.1.1.10 Binary bounded memory tree source

A binary bounded memory tree source is a binary tree source with memory no larger than D (a suffix set S that satisfies $l(s) \leq D, D \in \mathbb{N}^+, 0 \leq D < \infty$ for all $s \in S$), D is the *depth* of tree.

A special case is each element in sequence is an i. i. d random variable, which named *memoryless source*. Depth of memoryless source is zero.

4.1.1.11 Suffix function

A suffix function β_S: $\ldots x_{n-2}x_{n-1}x_n \to S$ where $\ldots x_{n-2}x_{n-1}x_n$is semi-infinite and S is a suffix set. A suffix parameter is a set Θ_S of suffix functions θ_s: $s \to [0,1], \forall s \in S$.

4.1.1.12 Next symbol probability

The *actual* next symbol probability at time point t for a bounded memory (depth of D) tree source based on past $(t-D)$ to $(t-1)$ with suffix set S and parameter set Θ_S

$$p_a(X_t = 1|x_{t-D}^{t-1}, \beta_S, S, \Theta_S)$$

$$= 1 - p(X_t = 0|x_{t-D}^{t-1}, \beta_S, S, \Theta_S) \triangleq \theta_S(x_{t-D}^{t-1}), \beta_S(x_{t-D}^{t-1}) = \ s \in S, \forall t$$

where a denotes actual probability.

For memoryless source, parameter set Θ_S has only one suffix function that is $\theta_s, s \in \mathcal{X}$, where $\mathcal{X}$ is the alphabet set of the memoryless source.

4.1.1.13 Theorem

A sequence x_1^t generated by a binary tree source with bounded memory of depth D has probability

$$p_a(X_1^t = x_1^t|x_{1-D}^0, S, \Theta_S)) = \prod_{\tau=1}^{t} p_a(X_\tau = x_\tau|x_{\tau-D}^{\tau-1}, S, \Theta_S)\,, t \geq 1$$

.

Proof:

For $t = 1$, $p_a(X_1^t = x_1^t|x_{1-D}^0, S, \Theta_S) = p_a(X_1 = x_1|x_{1-D}^0, S, \Theta_S)$;

For $t = 2$, $p_a(X_1^t = x_1^t|x_{1-D}^0, S, \Theta_S) = p_a(X_2 = x_2|x_{2-D}^1, S, \Theta_S)p_a(X_1 = x_1|x_{1-D}^0, S, \Theta_S)$

$= p_a(X_2 = x_2|x_{1-D}^1, S, \Theta_S)p_a(X_1 = x_1|x_{1-D}^0, S, \Theta_S)$

$= \frac{p_a(x_2, x_{1-D}^1, S, \Theta_S)}{p_a(x_{1-D}^1, S, \Theta_S)} \frac{p_a(x_1, x_{1-D}^0, S, \Theta_S)}{p_a(x_{1-D}^0, S, \Theta_S)}$

$= \frac{p_a(x_{1-D}^2, S, \Theta_S)}{p_a(x_{1-D}^1, S, \Theta_S)} \frac{p_a(x_{1-D}^1, S, \Theta_S)}{p_a(x_{1-D}^0, S, \Theta_S)}$

$$= \frac{p_a(x_{1-D}^2, S, \Theta_S)}{1} \frac{1}{p_a(x_{1-D}^0, S, \Theta_S)}$$

$$= p_a(X_1^2 = x_1^2 | x_{1-D}^0, S, \Theta_S);$$

Suppose $t = k$ have

$$p_a\left(X_1^k = x_1^k | x_{1-D}^0, S, \Theta_S\right) = \prod_{\tau=1}^{k} p_a(X_\tau = x_\tau | x_{\tau-D}^{\tau-1}, S, \Theta_S);$$

Then

$$p_a\left(X_{k+1} = x_{k+1} | x_{k+1-D}^k, S, \Theta_S\right) \prod_{\tau=1}^{k} p_a(X_\tau = x_\tau | x_{\tau-D}^{\tau-1}, S, \Theta_S)$$

$$= p_a\left(X_{k+1} = x_{k+1} | x_{1-D}^k, S, \Theta_S\right) p_a\left(X_1^k = x_1^k | x_{1-D}^0, S, \Theta_S\right)$$

$$= \frac{p_a(x_{k+1}, x_{1-D}^k, S, \Theta_S)}{p_a(x_{1-D}^k, S, \Theta_S)} \frac{p_a(x_1^k, x_{1-D}^0, S, \Theta_S)}{p_a(x_{1-D}^0, S, \Theta_S)}$$

$$= \frac{p_a(x_{k+1}, x_{1-D}^k, S, \Theta_S)}{p_a(x_{1-D}^k, S, \Theta_S)} \frac{p_a(x_1^k, x_{1-D}^0, S, \Theta_S)}{p_a(x_{1-D}^0, S, \Theta_S)}$$

$$= \frac{p_a\left(x_{k+1}, x_{1-D}^k, S, \Theta_S\right)}{p_a(x_{1-D}^0, S, \Theta_S)}$$

$$= p_a\left(X_1^{k+1} = x_1^{k+1} | x_{1-D}^0, S, \Theta_S\right) \text{ q. e. d.}$$

4.1.1.14 Model class

The set of all tree models having memory not larger than D is called the model class $\mathcal{C}_D$.

4.1.1.15 Cost of a tree model

The cost of a tree model (or a suffix set) S with respect to model class $\mathcal{C}_D$ is

$$\Gamma_D(S) \triangleq |S| - 1 + |\{s : s \in S, l(s) \neq D\}|$$

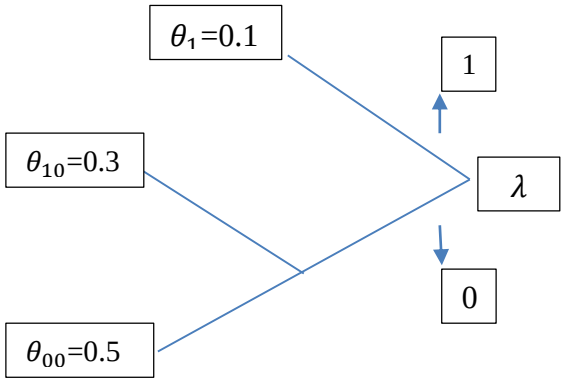
where it is assumed that $S \in \mathcal{C}_D$.

4.1.2 Probability of code word generated by binary bounded memory tree source (parameter known)

Given a binary tree model and its parameter, probability of a code word can be easily predicted.

Here cite the example in [6] as Fig 4.2.

Fig 4.2 An example of suffix set with parameters



Code word $x_1^7 = 0110100$ conditioned on its most recent past of depth $D = 3$ of $x_{-2}^0 = 010$ with suffix set $S = \{00, 10, 1\}$, suffix function $\beta_S(x_{t-3}^{t-1}) = s: \beta(000) = \beta(100) = 00, \beta(001) = \beta(011) = \beta(101) = \beta(111) = 1, \beta(010) = \beta(110) = 10$, tree model is Fig 4.1 and parameter $\Theta_S = \{\theta_{00} = 0.5,\ \theta_{10} = 0.3, \theta_1 = 0.1\}$.

Calculate the actual probability P_a of receiving $x_1^7 = 0110100$ conditioned on history $x_{-2}^0 = 010$. That is,

$P_a(x_1^7 = 0110100 | x_{-2}^0 = 010)$

$= P_a(X_1 = 0 | x_{-2}^0 = 010, s = 10) P_a(X_2 = 1 | x_{-1}^1 = 100, s = 00) P_a(X_3 = 1 | x_0^2$

$= 001, s = 1) P_a(X_4 = 0 | x_1^3 = 011, s = 1) P_a(X_5 = 1 | x_2^4 = 110, s$

$= 10) P_a(X_6 = 0 | x_3^5 = 101, s = 1) P_a(X_7 = 0 | x_4^6 = 010, s = 10)$

$= (1 - \theta_{10})\theta_{00}\theta_1(1 - \theta_1)\theta_{10}(1 - \theta_1)(1 - \theta_{10})$, by 4.1.1.13 theorem

$= 0.7 \times 0.5 \times 0.1 \times 0.9 \times 0.3 \times 0.9 \times 0.7$

$= 0.0059535$.

4.1.3 Probability of code word generated by binary bounded memory tree source (parameter unknown)

Fig 4.2 is a bounded memory tree with deterministic parameter. Generally, for unknown memory tree model and unknown parameters, Context Tree Weighting method introduced a weighted tree model, estimating the probability of source code received based on a maximum memory depth of D. The core of CTW is *K-T* (Krichevski-Trofimov [11]) estimator which smooths a probability estimated based on a mathematical model of binary memoryless source with Dirichlet distribution as a prior. A binary memoryless source with parameter set $\Theta = \{\theta_1\}$ generates a sequence with m zeros and n ones. Then the probability of such a sequence is $p_a(m, n) = (1 - \theta_1)^m \theta_1^n$. *K-T* estimator smooths $p_a(m, n)$ by a $(\frac{1}{2}, \frac{1}{2})$-Dirichlet distribution, that is

$$p_e(m,n) = \int_0^1 \frac{1}{\sqrt{(1-\theta_1)\theta_1}} (1-\theta_1)^m \theta_1^n \, d\theta_1$$

where $p_e(m,n)$ denotes estimated probability of a sequence of m zeros and n ones generated by a binary memoryless source. Recursion of *K-T* estimator can be found either [6] or [11] that is

$$p_e(a+1,b) = \frac{a+\frac{1}{2}}{a+b+1} p_e(a,b)$$

$$p_e(a,b+1) = \frac{b+\frac{1}{2}}{a+b+1} p_e(a,b)$$

$$p_e(0,0) = 1$$

. Proof is in [6, appendix Ⅱ].

4.1.3.1 Context Tree

Definition of context tree [6] $\mathcal{T}_D$ is a set of nodes labeled by s, where s is a binary string with length $l(s)$ such that $0 \leq l(s) \leq D$. Each node s is called the parent of the node $0s$ and $1s$. The node s is called the parent of the node $0s$ and $1s$, who in turn are the children of s. To each node $s \in \mathcal{T}_D$, there correspond counts $as \geq 0$ and $bs \geq 0$. For the children $0s$ and $1s$ of parent node s, the counts must satisfy $a_{0s} + a_{1s} = a_s$ and $b_{0s} + b_{1s} = b_s$.

4.1.3.2 Definition of weighted probability [6] p_w^s is

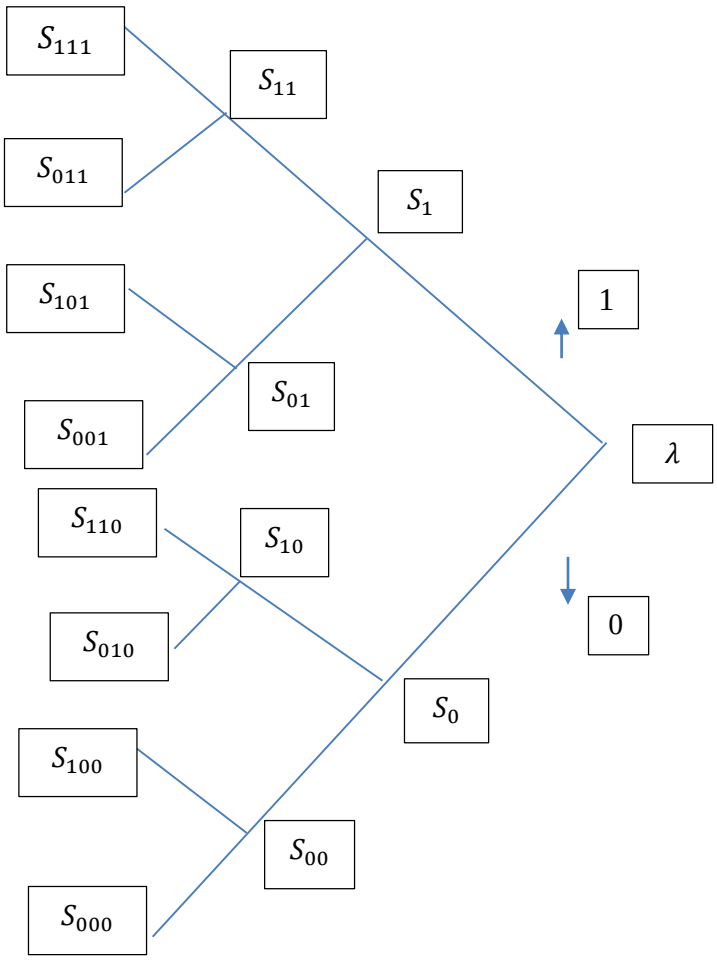
$$p_w^s \triangleq \begin{cases} kp_e(a_s,b_s) + (1-k)p_w^{0s}p_w^{1s}, 0 \leq l(s) < 0, k \in [0,1] \\ p_e(a,b), l(s) = D \end{cases}$$

where k is the weight.

4.1.3.3 Definition of Context Tree Weighting [6]

The context tree together with the weighted probabilities of the nodes is called a *weighted context tree*. [6, Fig 2] demonstrates a CTW probability assigned on a binary tree with $D = 3$. Redraw it in Fig 4.3.

Fig 4.3 An example of CTW probability assigned on a binary tree with depth=3



Received $x_1^7 = 0110100, x\_{-2}^0 = 010, D = 3$. Calculate $P_e^s(m,n)$ and $P_w^s(m,n), k = \frac{1}{2}$ by 4.1.3.2 definition for each node or leaf s as follows.

For $s = 111$, $a_{111} = 0, b_{111} = 0, => p_w^{111} = p_e^{111}(0,0) = 1$;

For $s = 011$, $a_{011} = 1, b_{011} = 0, => p_w^{011} = p_e^{011}(1,0) = \frac{0+\frac{1}{2}}{0+0+1} = \frac{1}{2}$;

For $s = 101$, $a_{101} = 1, b_{101} = 0, => p_w^{101} = p_e^{101}(1,0) = \frac{1}{2}$;

For $s = 001$, $a_{001} = 0, b_{001} = 1, => p_w^{001} = p_e^{001}(0,1) = \frac{1}{2}$;

For $s = 110$, $a_{110} = 0, b_{110} = 1, => p_w^{110} = p_e^{110}(0,1) = \frac{1}{2}$;

For $s = 010$, $a_{010} = 2, b_{101} = 0, => p_w^{010} = p_e^{010}(2,0) = \frac{3}{4}p_e(1,0) = \frac{3}{8}$;

For $s = 100$, $a_{100} = 1, b_{100} = 0, => p_w^{100} = p_e^{100}(1,0) = \frac{1}{2}$;

For $s = 000$, $a_{000} = 0, b_{101} = 0, => p_w^{000} = p_e^{000}(0,0) = 1$;

For $s = 11$, $p_e^{11}(1,0) = \frac{1}{2}$, $p_w^{11} = \frac{1}{2}p_e^{11} + \frac{1}{2}p_w^{011}p_w^{111} = \frac{1}{2}$;

For $s = 01$, $p_e^{01}(1,1) = \frac{1}{8}$, $p_w^{01} = \frac{1}{2}p_e^{01} + \frac{1}{2}p_w^{001}p_w^{101} = \frac{3}{16}$;

For $s = 10$, $p_e^{10}(2,1) = \frac{1}{16}$, $p_w^{10} = \frac{1}{8}$;

For $s = 00$, $p_e^{00}(0,1) = \frac{1}{2}$, $p_w^{00} = \frac{1}{2}$;

For $s = 1$, $p_e^{1}(2,1) = \frac{1}{16}$, $p_w^{1} = \frac{5}{64}$;

For $s = 0$, $p_e^{0}(2,2) = \frac{3}{128}$, $p_w^{0} = \frac{11}{256}$;

For $s = \lambda$, $p_e^{\lambda}(4,3) = \frac{5}{2048}$, $p_w^{\lambda} = \frac{95}{32768}$.

Redundancy [6] of CTW is bounded by $\frac{1}{2}logt$ where t is the length of sequence received.

Context Tree Weighting method can work on multi-alphabet [12] with redundancy for finite sequence received.

$$\log\frac{p_a(x^T)}{p_e(t||\mathcal{A}|)} \leq \frac{|\mathcal{A}|-1}{2}\log T + |\mathcal{A}| - 1$$

.

4.1.3.4 Sequential probability assignment [4]

A **sequential probability assignment** Q consists of a set of conditional probabilities $\left\{Q_{X_i|x^{i-1}}|\forall x^{i-1} \in \chi^{i-1}\right\}_{i=1}^{\infty}$.

4.1.3.5 Universal probability assignment [5]

Let $\mathcal{P}$ be a class of probability measures. A probability assignment Q is said to be **universal for the class $\mathcal{P}$** if the normalized relative entropy satisfies

$$\lim_{n\to\infty}\frac{1}{n}D\big(p(x^n)||Q(x^n)\big) = 0$$

for every probability measure p in $\mathcal{P}$. A probability assignment Q is said to be **universal** (without a qualifier) if it is universal for the class of stationary probability measures. CTW is one of universal probability assignments (Proof see [6]).

An example of CTW sequentially assigning probabilities to Fig 4.3 is given in Appendix.

Generally, CTW is to estimate $p(x_{n+1}|x^n) = \frac{p(x^{n+1})}{p(x^n)} \approx \frac{p_w^{\lambda}(x^{n+1}|x_{-D+1}^0)}{p_w^{\lambda}(x^n|x_{-D+1}^0)} = Q(x_{n+1}|x^n, x_{-D+1}^0)$. Actually if a stochastic process $\boldsymbol{X}$ is stationary, irreducible aperiodic and finite-alphabet Markov process, then $\lim_{n\to\infty} Q(x_{n+1}|x^n, x_{-D+1}^0) = p(x_{n+1}|x^n)\ a.e.$ [5, Lemma2]. CTW can work on estimation of joint probabilities of two finite-alphabet random sequences X, Y as well. Just map the alphabet sets $\mathcal{X}$ and $\mathcal{Y}$ to $\mathcal{X} \times \mathcal{Y}$ (one to one and onto).

4.2 Step2: Estimation of directed information

Let us go over the 2.2.6 definition and 2.3.1 Lemma of directed information, that is

$$I(X^N \to Y^N) \triangleq \sum_{n=1}^{N} I(X^n; Y_n \mid Y^{n-1})$$

$$= \sum_{n=1}^{N} [H(Y_n|Y^{n-1}) - H(Y_n|X^n, Y^{n-1})]$$

. Directed information between two random sequences is not a guaranteed convergence due to non-negative mutual information $I(X^n; Y_n \mid Y^{n-1})$. However, for a finite-alphabet stationary ergodic process, entropy rate will converge by general asymptotic

equipartition property (AEP) of Shannon-McMillan-Breiman Theorem [14], which makes convergence of directed information rate possible.

4.2.1 Lemma

If two discrete time stochastic processes $\boldsymbol{X}$ and $\boldsymbol{Y}$ are finite-alphabet stationary and ergodic, then directed information rate

$$\bar{I}_n(\boldsymbol{X} \to \boldsymbol{Y}) \triangleq \frac{1}{n} I(X^n \to Y^n)$$

will converge when n goes to infinity that is direct information

$$I(\boldsymbol{X} \to \boldsymbol{Y}) \triangleq \lim_{n\to\infty} \frac{1}{n} I(X^n \to Y^n)$$

does exist.

Proof:

$$\lim_{n\to\infty} \frac{1}{n} I(X^n \to Y^n)$$

$$= \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} [H(Y_n|Y^{n-1}) - H(Y_n|X^n, Y^{n-1})]$$

$$= \lim_{N\to\infty} \frac{1}{N} [H(Y^N) - H(Y^N||X^N)]$$

By Shannon-McMillan-Breiman-Theorm in [14], $\lim_{N\to\infty} \frac{1}{N} H(Y^N)$ exists and

$$\lim_{N\to\infty} \frac{1}{N} H(Y^N||X^N)$$

$$= \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \sum_{\substack{y^i \in \mathcal{Y} \\ x^i \in \mathcal{X}}} p\left(Y_i, Y^{i-1}, X^i\right) \log p\left(Y_i|Y^{i-1}, X^i\right) = \lim_{N\to\infty} H[Y_N|Y^{N-1}, X^N] \text{ in [9]}$$

, which makes $\lim_{N\to\infty}\frac{1}{N}(Y^N||X^N)$ exist as well. Q. e. d.

There are four estimators in [5]. Before, define

$$\widehat{H}_1(y^n||x^n) \triangleq -\frac{1}{n}\log Q(y^n||x^n) = -\frac{1}{n}\sum_{i=1}^{n}\log Q\left(y_i|y^{i-1},x^i\right),$$

$$\widehat{H}_1(y^n) \triangleq \widehat{H}_1(y^n||\emptyset) = -\frac{1}{n}\sum_{i=1}^{n}\log Q\left(y_i|y^{i-1}\right),$$

$$\widehat{H}_2(y^n||x^n) \triangleq \frac{1}{n}\sum_{i=0}^{n} f\left(Q\left(x_{i+1},y_{i+1}|x^i,y^i\right)\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{y_i\in\mathcal{Y}} Q\left(y_i|x^i,y^{i-1}\right)\log\frac{1}{Q\left(y_i|x^i,y^{i-1}\right)},$$

$$\widehat{H}_2(y^n) \triangleq \frac{1}{n}\sum_{i=1}^{n}\sum_{y_{i+1}\in\mathcal{Y}} Q\left(y_{i+1}|y^i\right)\log\frac{1}{Q(y_{i+1}|y^i)}$$

; Then define four types of estimators below,

$$\widehat{I_1}(x^n\to y^n) \triangleq \widehat{H}_1(y^n) - \widehat{H}_1(y^n||x^n)$$

$$= -\frac{1}{n}\sum_{i=1}^{n}\log Q\left(y_i|y^{i-1}\right) - \left[-\frac{1}{n}\sum_{i=1}^{n}\log Q\left(y_i|y^{i-1},x^i\right)\right];$$

$$\widehat{I_2}(x^n\to y^n) \triangleq \widehat{H}_2(y^n) - \widehat{H}_2(y^n||x^n);$$

$$\widehat{I_3}(x^n\to y^n) \triangleq \frac{1}{n}\sum_{i=1}^{n} D\left(Q\left(y_i|x^i,y^{i-1}\right)||Q\left(y_i|y^{i-1}\right)\right);$$

$$\widehat{I_4}(x^n\to y^n) \triangleq \frac{1}{n}\sum_{i=1}^{n} D\left(Q\left(x_{i+1},y_{i+1}|x^i,y^i\right)||Q\left(y_{i+1}|y^i\right)Q\left(x_{i+1}|x^i,y^i\right)\right);$$

. Please note Q denotes a universal probability assignment.

Although almost all convergence properties of four estimators have been proved in [5], I will still do some compensation of a few implicit proofs related to the first estimator $\widehat{I_1}(x^n \to y^n)$.

4.2.2 Theorem 1 of [5]

Let Q be a universal probability assignment and $(\boldsymbol{X}, \boldsymbol{Y})$ be a pair of finite-alphabet stationary ergodic discrete-time stochastic processes, then

$$\lim_{n\to\infty} \widehat{I_1}(x^n \to y^n) = \bar{I}(\boldsymbol{X} \to \boldsymbol{Y}), \text{ almost everywhere and in } \boldsymbol{L_1}$$

, where

$$\widehat{I_1}(x^n \to y^n) = \widehat{H}_1(y^n) - \widehat{H}_1(y^n||x^n)$$

$$= -\frac{1}{n}\sum_{i=1}^{n} \log Q\left(y_i|y^{i-1}\right) - \frac{1}{n}\sum_{i=1}^{n} \log Q\left(y_i|y^{i-1}, x^i\right)$$

.

Proof:

Since $\boldsymbol{X}, \boldsymbol{Y}$ are finite-alphabet stationary ergodic discrete-time processes, directed information between $\boldsymbol{X}, \boldsymbol{Y}$ exists and directed information rate converges to it.

The equation has two parts $-\frac{1}{n}\sum_{i=1}^{n} \log Q\left(y_i|y^{i-1}\right)$ and $-\frac{1}{n}\sum_{i=1}^{n} \log Q\left(y_i|y^{i-1}, x^i\right)$.

1. Try to show $-\frac{1}{n}\sum_{i=1}^{n} \log Q\left(y_i|y^{i-1}\right) \to H(\boldsymbol{Y})$, almost everywhere (a.e.);
2. Try to show $-\frac{1}{n}\sum_{i=1}^{n} \log Q\left(y_i|y^{i-1}, x^i\right) \to H(\boldsymbol{Y}||\boldsymbol{X})$, almost everywhere;

For 1,

Since $\boldsymbol{Y}$ is finite-alphabet stationary ergodic discrete-time processes, by Shannon-McMillan-Breiman Theorem

$$-\frac{n+1}{n}\frac{1}{n+1}\sum_{i=1}^{n}\log p\left(y_i|y^{i-1}\right) = -\frac{1}{n+1}\log p(Y_0, Y_1, \dots, Y_n) \to H \text{ a.e.}$$

$$\Rightarrow \quad -\frac{1}{n}\sum_{i=1}^{n}\log p\left(y_i|y^{i}\right) = -\frac{1}{n}\log p(y^n) \to H(\boldsymbol{Y}) \text{ a.e. as } n \to \infty$$

, by convergence of relative entropy implies convergence in distribution [15] and since $\boldsymbol{Y}$ is finite-alphabet, which imply the $\mathcal{Y}^n$ is at most countable by $n \to \infty$ and both $p(y^n)$ and $Q(y^n)$ are pmfs,

$$\Rightarrow \quad p(y^n) \to Q(y^n), n \to \infty$$

Then $-\frac{1}{n}\log Q(y^n) \to H(\boldsymbol{Y})$ a.e.

For 2, first introduce lemma 1 of [5] as

Lemma 1

Let $(\boldsymbol{X}, \boldsymbol{Y})$ be a jointly stationary ergodic finite-alphabet process, then,

$$\lim_{n\to\infty} -\frac{1}{n}\log p(y^n||x^n) = H(\boldsymbol{Y}||\boldsymbol{X}) \text{ a.e. and in } L_1$$

In addition, if $(\boldsymbol{X}, \boldsymbol{Y})$ is irreducible aperiodic Markov, then

$$E\left|-\frac{1}{n}\log p(y^n||x^n) - H(\boldsymbol{Y}||\boldsymbol{X})\right| = O(n^{-\frac{1}{2}}\log n)$$

and for every $\epsilon > 0$,

$$-\frac{1}{n}\log p(Y^n||X^n) - H(\boldsymbol{Y}||\boldsymbol{X}) = o(n^{-\frac{1}{2}}(\log n)^{\frac{5}{2}+\epsilon}) \text{ a.e}$$

.

Proof

Generally will follow paper [5]. However, definition of causal condition entropy in [15] is a little different from 2.2.8 definition. Therefore I will reproof it.

Some definitions,

$\bar{H}^{\infty}(Y||X) \triangleq E[-\log p(Y|Y_{-1}, Y_{-2}, \dots, Y_{-\infty}, Y_0, Y_{-1}, \dots, Y_{-\infty})]$

$= E[-\log p(Y_0|Y_{-k}^{-1}, X_{-k}^{0})]$

$H^k \triangleq E[-\log p(Y_0|Y_{-k}^{-1}, X_{-k}^{0})]$

$$\underset{p}{\overset{k}{\rightarrow}}(Y^N|X^N) \triangleq \underset{p}{\rightarrow}(Y^k|X^k) \prod_{i=k+1}^{N} p\left(Y_i|Y_{i-k}^{-1}, X_{i-k}^{i}\right)$$

$$\underset{p}{\rightarrow}(Y^N|X_{-\infty}^{N}, Y_{-\infty}^{0}) \triangleq \prod_{i=1}^{N} p\left(Y_i|Y_{i-k}^{-1}, X_{-\infty}^{i}\right)$$

.

Lemma 1.1

$$-\frac{1}{N}\log\underset{p}{\rightarrow}^{k}(Y^N|X^N) \to H^k, N \to +\infty$$

$$-\frac{1}{N}\log\underset{p}{\rightarrow}(Y^N|X_{-\infty}^{N}, Y_{-\infty}^{N}) \to H^k, N \to +\infty$$

Proof:

$$-\frac{1}{N}\log\underset{p}{\rightarrow}^{k}(Y^N|X^N)$$

$$= -\frac{1}{N} \underset{p}{\to} (Y^k|X^k) - \frac{1}{N}\sum_{i=k+1}^{N} \log p\left(Y^k|Y_{i-k}^{i-1}, X_{i-k}^{i}\right)$$

$\to 0 + H^k$(ergodic theorem),

$$-\frac{1}{N}\log \underset{p}{\to} (Y^N|X_{-\infty}^{N}, Y_{-\infty}^{0})$$

$= -\frac{1}{N}\sum_{i=1}^{N} \log p\left(Y_i|Y_{-\infty}^{i-1}, X_{-\infty}^{i}\right) \to \bar{H}^{\infty}(Y||X)$ (Ergodic theorem) q. e. d.

Lemma 1.2

$H^k \to \bar{H}^{\infty}(Y||X)$, $\bar{H}(Y||X) = \bar{H}^{\infty}(Y||X)$ where $\bar{H}(Y||X) \triangleq \lim_{n\to\infty} \bar{H}(Y_N|X^N, Y^{N-1})$

Proof:

$\bar{H}(Y||X) = \bar{H}^{\infty}(Y||X) = E[-\log p(Y_0|Y_{-\infty}^{-1}, X_{-\infty}^{0})]$

Since $\{H_k\}$ is non-increasing and both $\boldsymbol{Y}$ and $\boldsymbol{X}$ are stationary, by monotone convergence theorem,

⇨ $H^k \to \bar{H}^{\infty}(Y||X)$;

$p(Y_0|\, Y_{-k}^{-1}, X_{-k}^{0}) \to p(Y|Y_{-\infty}^{-1}, X_{-\infty}^{0})$ by martingale convergence theorem, $Y \in \mathcal{Y}$ where $|\mathcal{Y}|$ finite,

⇨ will have $\lim_{k\to\infty} H^k = \lim_{n\to\infty} E\left[-\sum_{y_0\in\mathcal{Y}} p\,(Y_0|Y_{-k}^{-1}, X_{-k}^{0}) \log p(Y_0|Y_{-k}^{-1}, X_{-k}^{0})\right]$

, since $p(Y_0|Y_{-k}^{-1}, X_{-k}^{0})$ measurable function, then $\log p(Y_0|Y_{-k}^{-1}, X_{-k}^{0})$ and

$p(Y_0|Y_{-k}^{-1}, X_{-k}^{0}) \log p(Y_0|Y_{-k}^{-1}, X_{-k}^{0})$ are measurable as well,

⇨ $\lim_{k\to\infty} H^k = \lim_{n\to\infty} E\left[-\sum_{y_0\in\mathcal{Y}} p\,(Y_0|Y_{-k}^{-1}, X_{-k}^{0}) \log p(Y_0|Y_{-k}^{-1}, X_{-k}^{0})\right]$

$$= E\left[\lim_{n\to\infty} -\sum_{y_0\in\mathcal{Y}} p\,(Y_0|Y_{-k}^{-1}, X_{-k}^{0}) \log p(Y_0|Y_{-k}^{-1}, X_{-k}^{0})\right]$$

$= \bar{H}^{\infty}(Y||X)$, Q. e. d.

Lemma 1.3

$$\lim_{N\to\infty} sup \frac{1}{N} \log \frac{\overrightarrow{p^k}(Y^N|X^N)}{\overrightarrow{p}(Y^N|X^N)} \leq 0$$

$$\lim_{N\to\infty} sup \frac{1}{N} \log \frac{\overrightarrow{p^k}(Y^N|X^N)}{\overrightarrow{P}(Y^N|X_{-\infty}^{N}, Y_{-\infty}^{0})} \leq 0$$

Where $\underset{p}{\rightarrow}(X^N|X_{-\infty}^{N}, Y_{-\infty}^{0}) \triangleq \prod_{i=1}^{N} p(Y_i|X_{-\infty}^{i}, Y_{-\infty}^{i-1})$

Proof:

$$\mathrm{E}\left[\frac{\overrightarrow{p^k}(Y^N|X^N)}{\overrightarrow{p}(Y^N|X^N)}\right] = \sum_{y^N,x^N} p(x^N, y^N) \frac{\overrightarrow{p^k}(y^N|x^N)}{\overrightarrow{p}(y^N|x^N)}$$

$$\underset{p^k}{\rightarrow}(y^N|x^N) = \underset{p^k}{\rightarrow}(y^k|x^k) \prod_{i=k+1}^{N} p(y_i|x_{i-k}^{i}, y_{i-k}^{i-1})$$

$$\underset{p}{\rightarrow}(y^N|x^N) = \prod_{i=1}^{N} p(y_i|y^{i-1}, x^i)$$

$$\Rightarrow \frac{\overrightarrow{p^k}(Y^N|X^N)}{\overrightarrow{p}(Y^N|X^N)} = \frac{\prod_{i=1}^{k} p(y_i|y^{i-1}, x^i) \prod_{i=k+1}^{N} p(y_i|x_{i-k}^{i}, y_{i-k}^{i-1})}{\prod_{i=1}^{N} p(y_i|y^{i-1}, x^i)}$$

$$= \prod_{i=k+1}^{N} \frac{p(y_i|x_{i-k}^{i}, y_{i-k}^{i-1})}{p(y_i|y^{i-1}, x^i)} \quad *$$

Since $p(y_i|y^{i-1},x^i)=\frac{p(y_i,y^{i-1},x^i)}{p(y^{i-1},x^i)}=\frac{p(y^i,x^i)}{p(y^{i-1},x^i)}$ replace *, then will have

*=$\prod_{i=k+1}^{N}\frac{p(y_i|x_{i-k}^{i},y_{i-k}^{i-1})}{p(x^i,y^i)}\ p(y^{i-1},x^i)$

=> $\mathrm{E}\left[\frac{\overrightarrow{p^k}(Y^N|X^N)}{\overrightarrow{p}(Y^N|X^N)}\right]=\sum_{y^N,x^N}p(x^N,y^N)\prod_{i=k+1}^{N}\frac{p(y_i|x_{i-k}^{i},y_{i-k}^{i-1}}{p(x^i,y^i)}\ p(y^{i-1},x^i)$

$$=\sum_{y^N,x^N}p(x^N,y^N)\prod_{i=k+1}^{N}\frac{p(y^{i-1},x^i)}{p(y^i,x^i)}p(y_i|x_{i-k}^{i},y_{i-k}^{i-1})$$

, shift left by one bit, most left shifted to most right, will have

$$=\sum_{y^N,x^N}p(x^{k+1},y^k)\prod_{i=k+1}^{N-1}\frac{p(y^i,x^{i+1})}{p(y^i,x^i)}p(y_i|x_{i-k}^{i},y_{i-k}^{i-1})\ p(y^N,x^N)p(y_N|x_{N-k}^{N},y_{N-k}^{N-1})$$

=

$\sum_{y^{N-1},x^N}p(x^{k+1},y^k)\prod_{i=k+1}^{N-1}\frac{p(y^i,x^{i+1})}{p(y^i,x^i)}\ p(y_i|x_{i-k}^{i},y_{i-k}^{i})\sum_{y_N\in\mathcal{Y}}\ p(y^N,x^N)p(y_N|x_{N-k}^{N},y_{N-k}^{N-1})$

**

Since

$$\sum_{y_N\in\mathcal{Y}}p(y^N,x^N)p(y_N|x_{N-k}^{N},y_{N-k}^{N-1})\le\sum_{y_N\in\mathcal{Y}}p(y_N|y_{N-k}^{N-1},x_{N-k}^{N})=1$$

⇨ **$\le\sum_{y^{N-1},x^{N-1}}p(x^{k+1},y^k)\prod_{i=k+1}^{N-1}\frac{p(y^i,x^{i+1})}{p(y^i,x^i)}p(y_i|x_{i-k}^{i},y_{i-k}^{i-1})$

$$=\sum_{y^{N-2},x^{N-1}}p(x^{k+1},y^k)\prod_{\mathrm{i=k+1}}^{\mathrm{N-2}}\frac{p(y^i,x^{i+1})}{p(y^i,x^i)}\ p(y_i|x_{i-k}^{i},y_{i-k}^{i-1})$$

$$\times\sum_{\substack{y_{N-1}\in\mathcal{Y}\\x_N\in\mathcal{X}}}\left[p(x_{i+1}|y^i,x^i)p(y_i|y_{i-k}^{i-1},x_{i-k}^{i}\right]|_{i=N-1}$$

. Since

$$\sum_{\substack{y_{N-1}\in\mathcal{Y}\\ x_N\in\mathcal{X}}} p(x_N|y^{N-1},x^{N-1})\ p(y_{N-1}|y_{N-k-1}^{N-2},x_{N-k-1}^{N-1})$$

$$= \sum_{y_N\in\mathcal{Y}} p(y_{N-1}|y_{N-k-1}^{N-2},x_{N-k-1}^{N-1}) \sum_{x_N\in\mathcal{X}} p(x_N|y^{N-1},x^{N-1}) = 1$$

⇨ Follow the decomposition above, **≤ 1

$$0 \leq E\left[\frac{\overrightarrow{p^k}(Y^N|X^N)}{\overrightarrow{p}(Y^N|X^N)}\right] \leq 1$$

⇨ Markov inequality in [15] is still hold. The proof left is similar to [15]. Q. e. d.

By those lemmas above, AEP in [15] holds, then Lemma 1 holds.

⇨ $\lim_{n\to\infty} -\frac{1}{n}\log p(y^n||x^n) = H(\boldsymbol{Y}||\boldsymbol{X})$ a.e.

Rest of proofs are same with [5], Q. e. d.

For proofs related to other lemmas and theorems, please refer to [5, Ⅳ].

4.3 Implementation of algorithm

Paper [5] developed a set of codes. I will use it and analyze the model of Fig 3.1.

Basically, algorithm [5] was developed based on the complex reduced CTW in [16].

Complex reduced CTW introduced an updating tree method that each time update two inputs and two outputs for the next stage. Mutual information is estimated by 2.3.4 theorem.

5. APPLICATIONS OF ESTIMATION OF DIRECTED INFORMATION

Generally, I will illustrate three examples of inferring information flow or causality. One is based a schematic logic circuit and the other two are biomedical related.

5.1 Estimation of directed information of an AND gate

Fig 3.1 illustrate theoretical analysis of directed information on a usual two inputs AND gate controlled by D trigger with $N = 3$.

Suppose let $N \to \infty$. Obviously, inputs $X(n)$, $\xi(n)$ and output $Y(n)$ become stationary stochastic processes. By definition 2.2.10, directed information rate of $\boldsymbol{X}$ to $\boldsymbol{Y}$ is

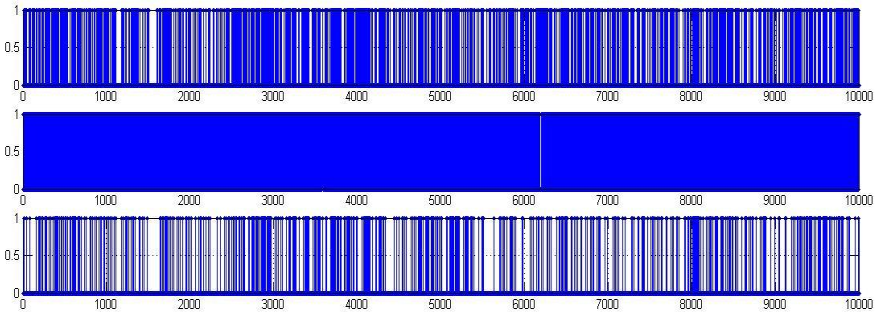
$$I(\boldsymbol{X} \to \boldsymbol{Y}) = \lim_{N\to\infty} \frac{1}{N} I(X^N \to Y^N)$$

$$= \lim_{N\to\infty} \frac{0.2123(N-1)}{N} = 0.2123$$

. That is asymptotically, its directed information rate from $\boldsymbol{X} \to \boldsymbol{Y}$ will converge to 0.2123 while directed information rate from $\boldsymbol{Y} \to \boldsymbol{X}$ is always 0.

Simulate Fig 3.1 by 10^4 samples.

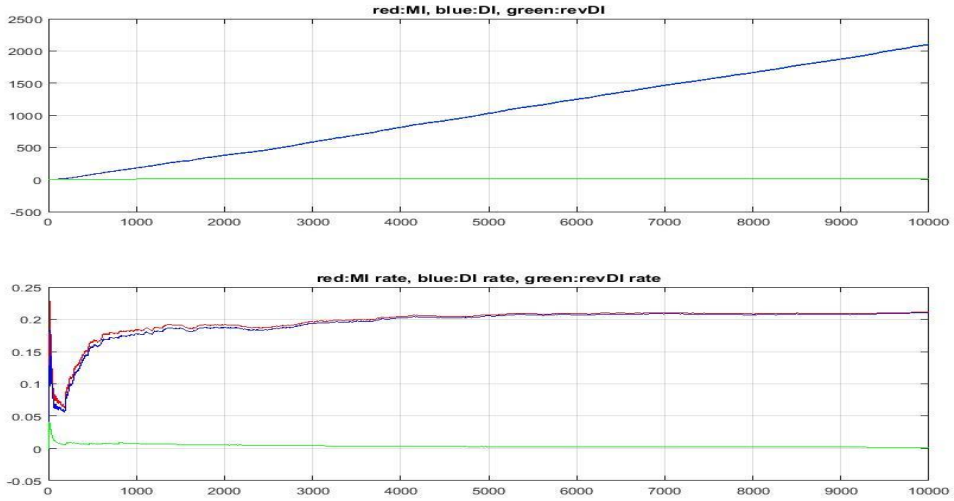
Fig 5.1 Simulation results based on schematic of Fig 3.1



Plots from top to bottom represents values of input $\xi(n)$, input $X(n)$ and output $Y(n)$, respectively.

Then run algorithm from chapter 4.3, results are below.

Fig 5.2 Performance of inferring causality by directed information rate on Fig 3.1



Red, blue and green curves denotes mutual information rate (calculated by 2.3.4 theorem), directed information rate $\boldsymbol{I}(\boldsymbol{X} \to \boldsymbol{Y})$, and reversed directed information rate $\boldsymbol{I}(\boldsymbol{Y} \to \boldsymbol{X})$, respectively.

The blue curve converges to theoretical 0.2123. Green curve goes to zero representing for trend of reversed directed information rate that ($\boldsymbol{I}(\boldsymbol{Y} \to \boldsymbol{X})$) turns out to be zero. In conclusion, it verifies the theoretical analyses in part 3 that information flow only from $\boldsymbol{X} \to \boldsymbol{Y}$, or in other words, $\boldsymbol{X}$ causes $\boldsymbol{Y}$, not vice versa.

5.2 Estimation of directed information in single parent neural network

Neural network is a set of neurons that communicate by either transmitting electrical signal (electrical potential) or neurotransmitter. We focus on the former. Electrical signals have been modeled as a set of Poisson processes of spike train in [7].

Spiking activity of a neuron $i \in \Psi$ can be completely described by a Poisson process with *conditional intensity function* (CIF) that is $\lambda_i(t|H_i(t))$ in $N(t)$, satisfying following properties

1) $N_i(0) = 0$;

2) $\{N_i(t), t \geq 0\}$ has independent increments;

3) $P(\{N_i(t + \Delta t) - N_i(t) \geq 2\}|H(t)) = o(\Delta t)$;

4) $P(\{N(t + \Delta t) - N(t) = 1\}|H(t)) = \lambda\big(t|H_i(t)\big)\Delta t + o(\Delta t)$

. $N_i(t)$ denotes number of arrivals of spikes of the neuron i till time point $t \in (0, T]$ where $(0, T]$ denotes an observation interval; From (4), $\lambda_i(t|H_i(t))$ can be defined as

$$\lambda_i\big(t|H(t)\big) = \lim_{\Delta t \to 0} \frac{P(\{N_i(t+\Delta t) - N_i(t) = 1|\, H(t)\})}{\Delta t}$$

, where $H(t) = \{n_h(t)|h \in \Psi\}$ is history of spikes of the ensemble network before time t and $n_h(t)$ is history of neuron h before time t. With a sufficiently small Δt, actually, the probability of neuron i spikes once at interval $[t, t+\Delta t)$ can be approximated by $\lambda_i\big(t|H(t)\big)\Delta t$.

Generally, spiking activity of a neuron comes from two kinds of effects: one is from the history of itself and the other is from activities of other neurons in the network. In fact, based on Poisson regression in generalized linear model (GLM) [17], $\lambda_i\big(t|H(t)\big)$ is decided by

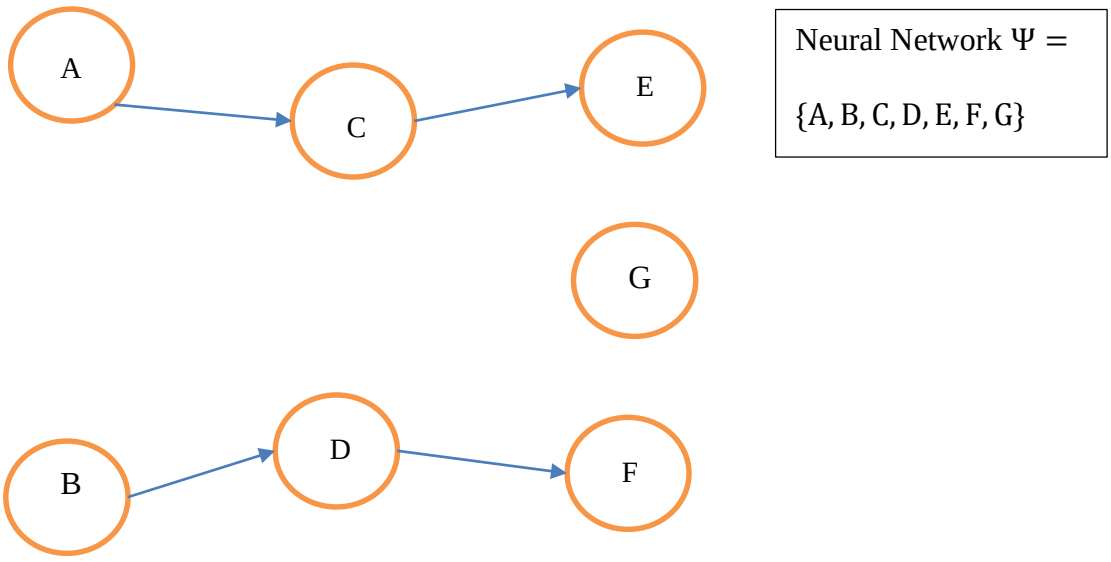
$$\log \lambda_i\big(t|H(t)\big) = \alpha_0 + \sum_{j=1}^{J} \alpha_j n_i(t-j) + \sum_{h \in (\Psi \backslash \mathrm{i})} \sum_{k_h=1}^{K_h} \beta_{k_h} n_h\big(t-(k_h-1)\big), i \in \Psi$$

. Here J or K_h is condition depth of history of the neuron i itself or other neuron h, respectively. $[\alpha_0\ \alpha_1\ \ldots\ \alpha_J]$ and $\{[\beta_{k_1}\ \beta_{k_2}\ \ldots\ \beta_{K_h}]|h \in (\Psi \backslash \mathrm{i})\}$ are parameter vectors. First two parts, $\alpha_0 + \sum_{j=1}^{J} \alpha_j n_i(t-j)$, history of the neuron i itself, represent intrinsic effect while last part $\sum_{h \in (\Psi \backslash \mathrm{i})} \sum_{k_h=1}^{K_h} \beta_{k_h} n_h\big(t-(k_h-1)\big)$ is named extrinsic effect. Both intrinsic and extrinsic effects have two typical types of potentials--one is inhibitory and the other is excitatory. Inhibitory potential makes a postsynaptic neuron less likely to generate an action potential, or more concisely, the neuron becomes less like to spike.

On the contrary, excitatory potential increases the probability of an action potential occurring in a postsynaptic neuron. Neurons usually spike under both two effects. For example, a neuron A's behavior is influenced by both its own inhibition called self-inhibition and the other neuron's excitatory potential. More specifically, from the perspective of parameters, $[\alpha_0\ \alpha_1\ \ldots\ \alpha_J]$ is all negative like $[-0.5\ -2\ -3.2]$ (depth is 2) and $[\beta_1\ \beta_2\ \ldots\ \beta_7]$ is all positive such as [1.2 0.9 8.6 4.8] (depth is 4).

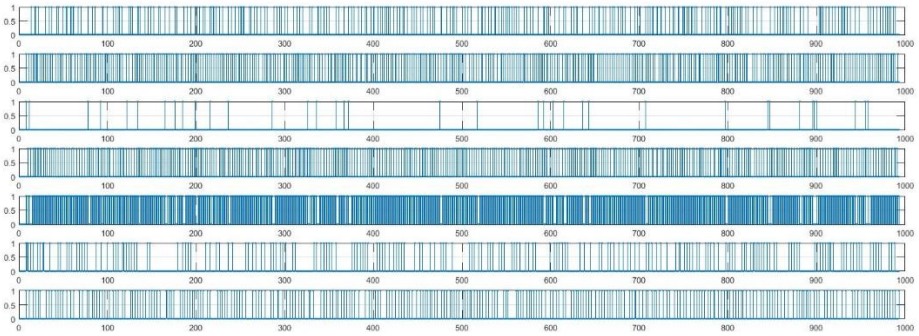
Fig 5.3 demonstrates an ensemble of neural network Ψ containing seven neurons and each neuron has at most one presynaptic neuron (*single parent neural network*: each node in a neural network has at most one parent.). Its GLM is described as there is only one $h \in (\Psi \backslash i)$ such that $[\beta_{k_1}\ \beta_{k_2}\ \ldots\ \beta_{K_h}] \neq \mathbf{0}$. An arrow describes directivity of causal influence from one neuron to the other.

Fig 5.3 An ensemble of single parent neural network containing 7 neurons



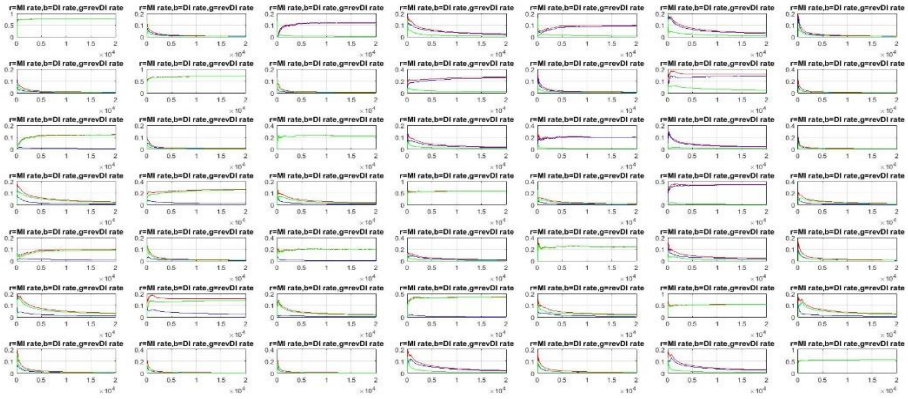
Simulation generated total 20k samples of neuron spike train based on Fig 5.3. First 1000 samples representing for one-second data are showed below (Top to bottom are neurons A, B, C, D, E, F, G).

Fig 5.4 Simulation results based on schematic of Fig 5.3



Results of estimates of three kinds of information rates are below. Permutation of seven neurons is 49 and columns are seven inputs and rows are seven outputs. Top to bottom and left to right are both A, B, C, D, E, F, G. Diagonals are self-input and self-output. Similar to the example in chapter 5.1, red curve is mutual information rate (MI_rate), blue curve is directed information rate (DI_rate) and green curve represents reverse directed information rate (revDI_rate).

Fig 5.5 Performance of inferring causality by DI rate



Since each node in network has at most one parent, ignore diagonal seven figures, there is at most one neuron showing causal influence in each of the seven columns (list below) and others' DI_rates tend to zero. So it is more reasonable to pick up the figure that has the maximum value of DI rate (blue) in each of the seven columns except diagonal.

1st column: None;

2nd column: None;

3rd column: $A \rightarrow C$, MI rate= 0.1209, DI rate= 0.1181;

4th column: $B \rightarrow D$, MI rate= 0.2686, DI rate= 0.2563;

$F \rightarrow D$, MI rate= 0.4306, DI rate= 0.0074;

5th column: $C \rightarrow E$, MI rate= 0.2001, DI rate= 0.1972;

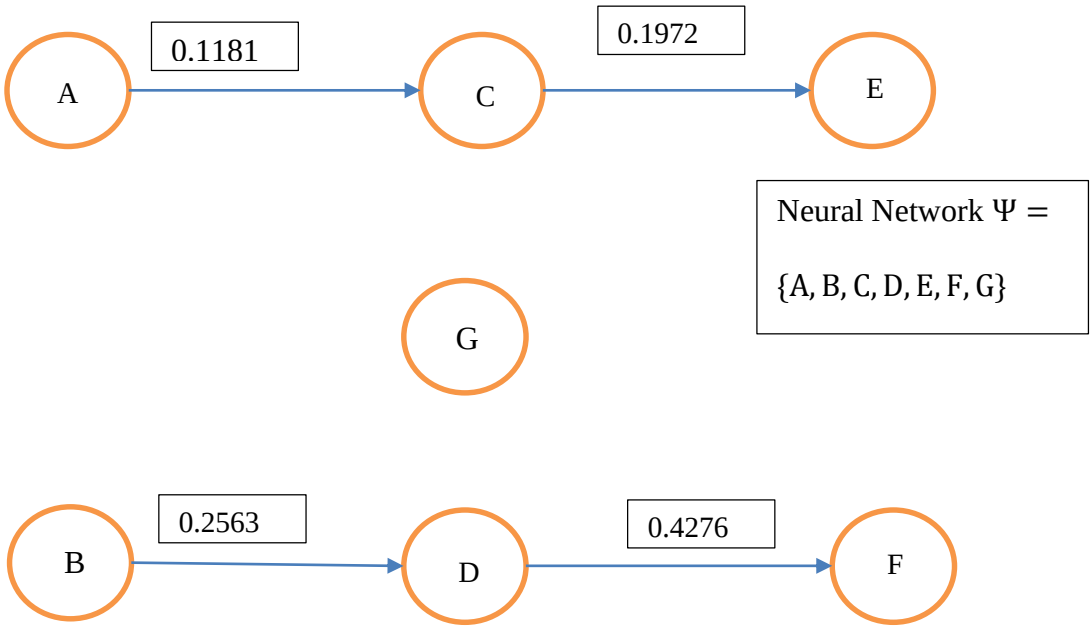
$A \rightarrow E$, MI rate= 0.0985, DI rate= 0.0934;

6th column: $D \rightarrow F$, MI rate= 0.4346, DI rate= 0.4276;

$B \rightarrow F$, MI rate= 0.1588, DI rate= 0.1434;

In the last one (7^{th} column), all first six rows are going to zero except the last one, which means that G is autarkic and independent from other neurons in the network. Based on DI_rates and single parent stochastic network, I rebuild schematic diagram in Fig 5.6. All causal inference have been dug out successfully.

Fig 5.6 Demonstration of causality inferred based on single parent network



5.3 Estimation of directed information in multiple-parent neural network

A multiple-parent neural network denotes each neuron in the neural network is not restricted to one parent at most and can actually have multiple ones.

Fig 5.7 depicts a topology of multi-parent neural network. An arrow describes directivity of causal influence from one neuron to the other.

Fig 5.7 An ensemble of multiple parent neural network

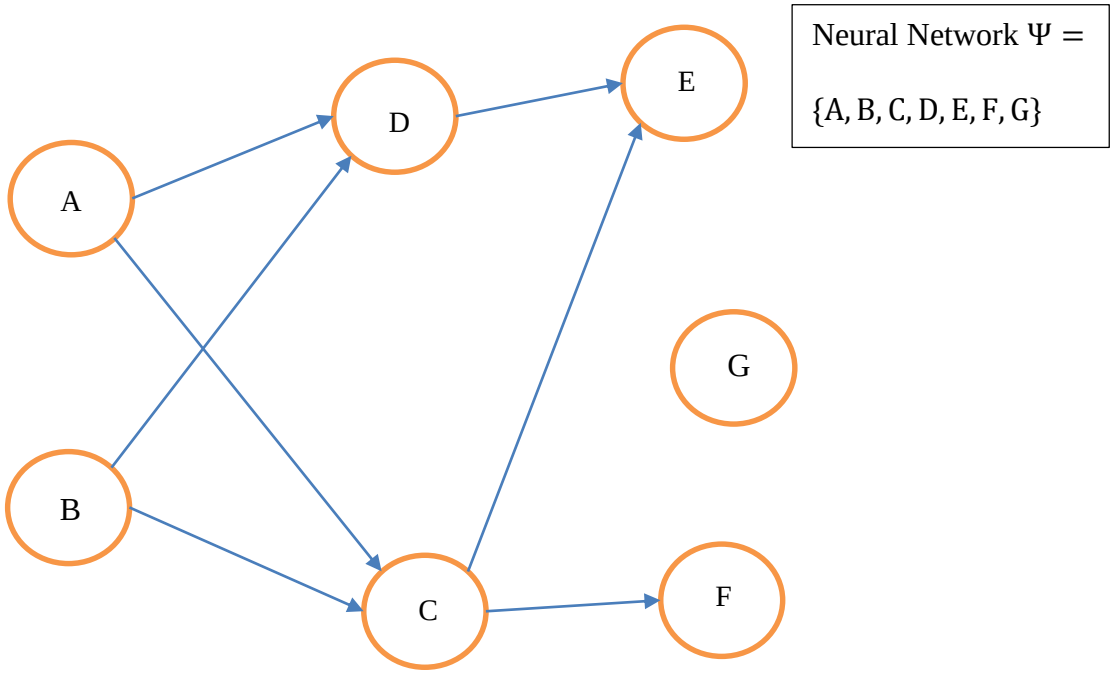
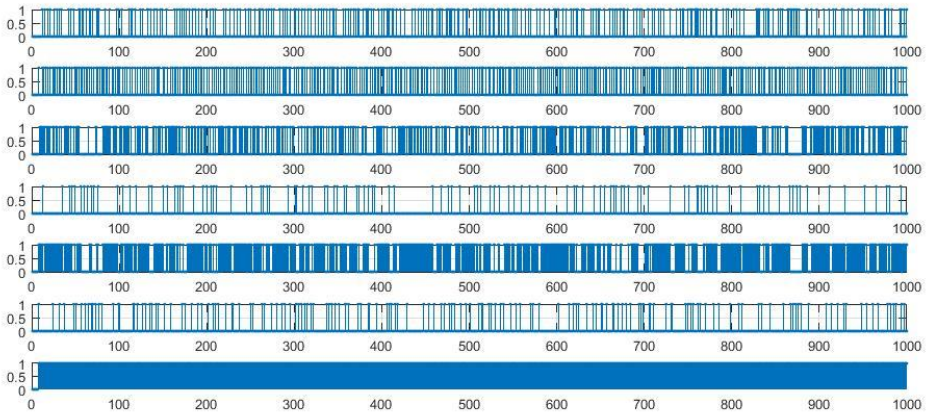


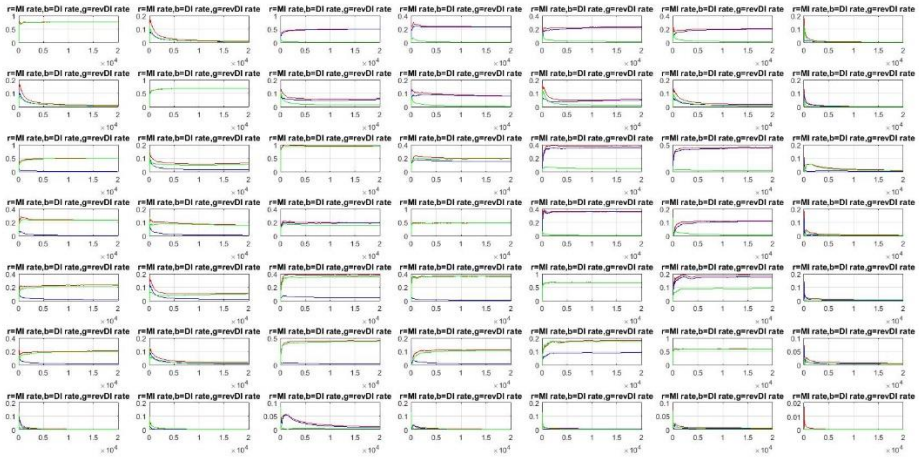
Fig 5.8 Simulation results based on Fig 5.7



Simulation ran for 20k samples and first 1000 samples were plotted out above (top to bottom are neurons A, B, C, D, E, F, G).

Results of information rates are below. Similar to chapter 5.2, 49 figures represent permutation of seven inputs and seven outputs and both top to bottom and left to right are A, B, C, D, E, F, G, either.

Fig 5.9 Performance of inferring causality by DI rate



Ignoring those DI values approaching zeros with horizontal axis increase and screening for significant DI values (significant blue curve), will have

1st: $A \rightarrow C$, MI rate= 0.5046, DI rate= 0.4977;

$A \rightarrow D$, MI rate= 0.2326, DI rate= 0.2290;

$A \rightarrow E$, MI rate= 0.2298, DI rate= 0.2208;

$A \rightarrow F$, MI rate= 0.2048, DI rate= 0.1978;

2nd: $B \rightarrow C$, MI rate= 0.0617, DI rate= 0.0539;

$B \rightarrow D$, MI rate= 0.0807, DI rate= 0.0781;

$B \rightarrow E$, MI rate= 0.0548, DI rate= 0.0492;

3rd: $C \rightarrow D$, MI rate= 0.1959, DI rate= 0.1516;

$C \rightarrow E$, MI rate= 0.3751, DI rate= 0.3556;

$C \rightarrow F$, MI rate= 0.4496, DI rate= 0.4385;

4th: $D \rightarrow C$, MI rate= 0.1941, DI rate= 0.1881;

$D \rightarrow E$, MI rate= 0.3619, DI rate= 0.3585;

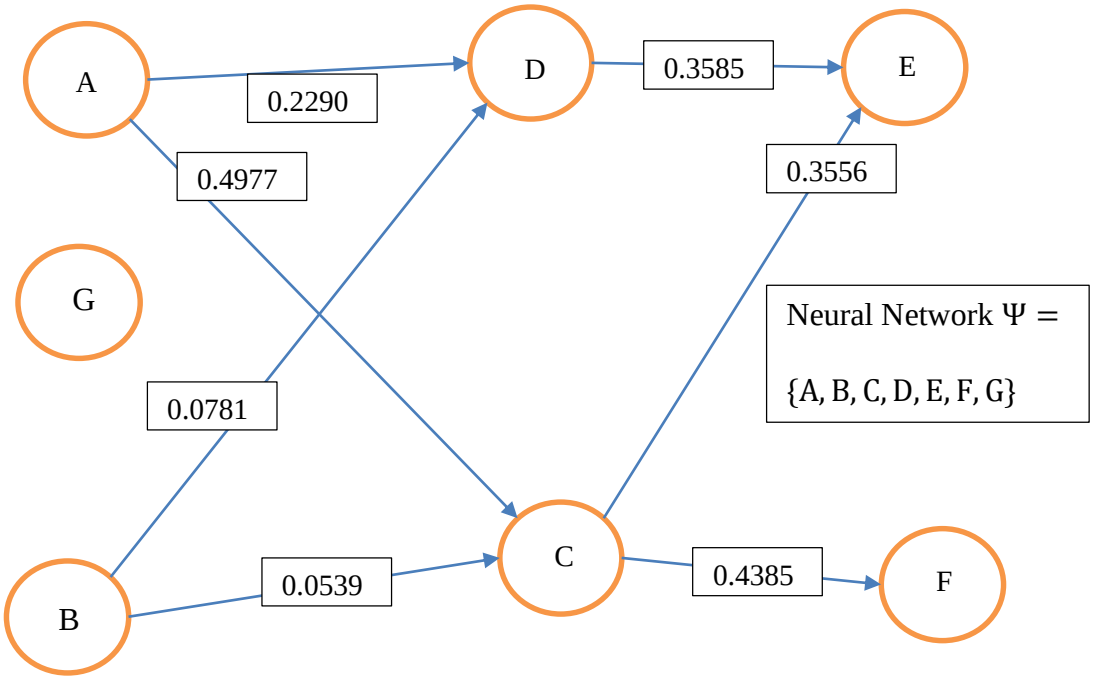
$D \rightarrow F$, MI rate= 0.1122, DI rate= 0.1076;

5th: $E \rightarrow F$, MI rate= 0.1857, DI rate= 0.1730;

6th: $F \rightarrow E$, MI rate= 0.1823, DI rate= 0.0901;

Based on DI_rates, rebuild schematic diagram below.

Fig 5.10 Demonstration of causality inferred by directed information rate



All causal relationships in Fig 5.10 have been figured out (DI_rates colored blue). However, there are still significant positive DI_rates (colored purple), which are indirect causality mistaken for direct causality. Besides, $C - D$ and $E - F$ are informed to have causal influence (DI_rates colored yellow) while actually not. This is because $C - D$ and $E - F$ have identical parents respectively.

6. SUMMARY AND FUTURE WORK

In conclusion, directed information builds up a bridge between statistical causality and direction of information flow, quantifying uncertainty of causality into bits. Estimation of directed information is computationally efficient to connect theory of directed information with practice, promoting its ideal candidate for causality inferring.

Although promising results has been achieved in single parent neural network, currently there is a flaw that the algorithm cannot distinguish direct causality from indirect causality when it works in multiple-parent neural network. I will keep working and hope to be able to make breakthrough on decouple indirect causality from direct causality.

REFERENCES

1. C. E. Shannon, *A Mathematical Theory of Communication, The Bell System Technical Journal*, Vol. 27, pp. 379-423, 623-656, July, October, 1948.
2. James L. Massey, *Causality, Feedback and Directed information, Proc. 1990 Intl. Symp. on Info. Th. and its Applications*, Waikiki, Hawaii, Nov. 27-30, 1990.
3. Judea Pearl, *Causality: Models, Reasoning, and Inference, 1st edition*, Cambridge University Press, March 12, 2000.
4. Lei Zhao, Haim Permuter, Young-Han Kim, and Tsachy Weissman, *Universal Estimation of Directed Information, in Proc. IEEE Int. Symp. Inf. Theory*, pp.230-234, 2010.
5. Jiantao Jiao, Haim H. Permuter, Young-Han Kim, Tsachy Weissman, *Universal estimation of Directed information*, arXiv: 2101.2334v4 [cs. IT], 30 May 2013.
6. Frans M. J. Willems, Yuri M. Shtarkov, and Tjalling J. Tjalkens, *The Context-Tree Weighting Method: Basic Properties, IEEE Trans. Inf. Theory*, vol. 41, no.3 pp. 653-664, 1955.
7. Peter Dayan, L. E. Abbott, *Theoretical Neuroscience Computational and Mathematical Modeling of Neural Systems*, The MIT Press, Cambridge, Massachusetts, London, England, 2001.
8. Thomas M. Cover, Joy A. Thomas, *Elements of Information Theory, 2nd*, Wiley, 2006.
9. Gerhard Kramer, *Directed Information for Channels with Feedback*, Konstanz: Hartung-Gorre Verlag, Dr. sc. Thchn. Dissertation, Swiss Federal Institute of Technology, Zurich, 1998.
10. Wikipedia, *causality, 12. 2015*, retrieve from https://en.wikipedia.org/wiki/Causality.
11. Raphail E. Krichevsky and Victor K. Trofimov, *The Performance of Universal Encoding, IEEE Transactions On Information Theory*, VOL, IT-27, No.2, March, 1981.

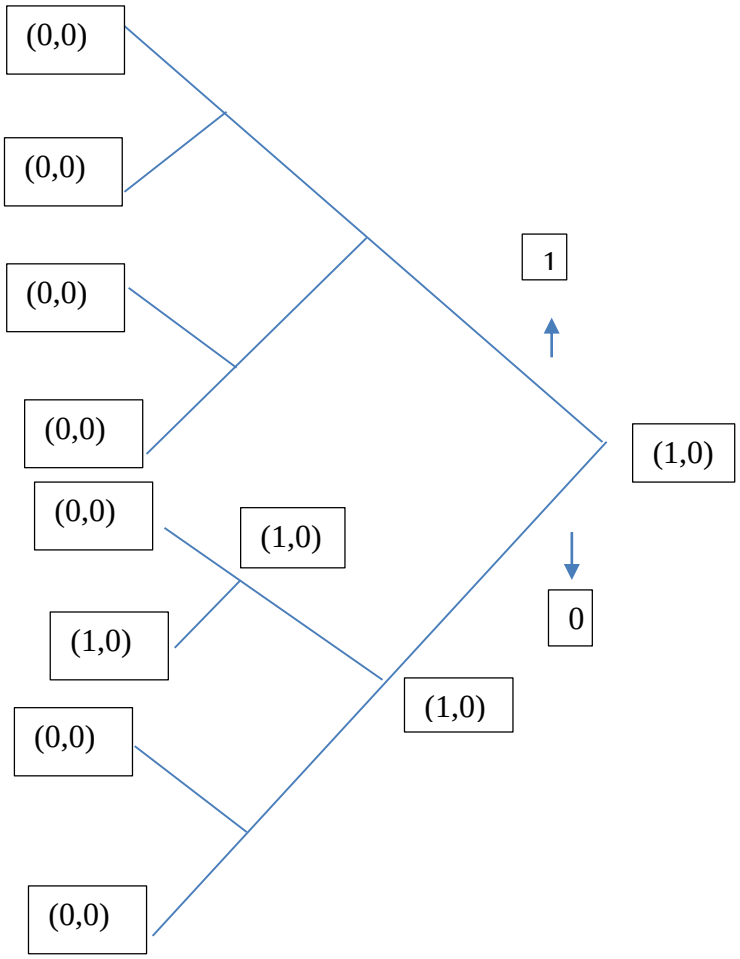
12. Tjalling J. Tjalkens, Yuri M. Shtarkov and Frans M. J. Willems, *Sequential Weighting Algorithms for Multi-Alphabet Sources, 6th Joint SwedishRussian Int. Workshop Inf. Theory,* pp. 230-234, 1993.
13. Paul H. Algoet and Thomas M. Cover, *A Sandwich Proof of the Shannon-McMillan-Breiman Theorem, The Annual Probability,* 16(2): pp 899-909, 1988.
14. Andrew R. Barron, *Entropy And The Central Limit Theorem, The Annual pf Probability,* Vol. 14, No. 1, 336-342, 1986.
15. Ramji Venkataramanan and S. Sandeep Pradhan, *Source Coding With Feed-Forward: Rate-Distortion Theorems and Error Exponents for a General Source, IEEE Transactions on Information Theory,* VOL. 53, No. 6, June 2007.
16. Frans M.J. Willems and Tjalling J. Tjalkens , *Complexity reduction of the context-tree weighting algorithm a study for KPN Research,* Tech. Rep. University of Eindhoven, the Netherlands, EIDMA RS.97.01, 1997.
17. Robert E. Kass, Uri T. Eden, Emery N. Brown, *Analysis of Neural Data, ISBN: 978-1-4614-9601-4, Springer Series in Statistics,* 2014.
18. Christopher J. Quinn, Todd P. Coleman, Negar Kiyavash, Nicholas G. Hatsopoulous, *Estimating the directed information to infer causal relationships in ensemble neural spike train recordings, Journal of Computational Neuroscience,* June 2010.
19. Christopher Quinn, Negar Kiyavash, Todd P. Coleman, *Directed Information Graphs,* arXiv: 1204.2003v1 [cs. IT], 9 April, 2012.

APPENDIX

Example of sequentially assigning probabilities to Fig 4.3 (for 4.1.3.5)

(1)

Fig A1.1



For $s = 010$ =>$P_e^{010}(a_{010}, b_{010}) = \frac{1}{2}$ => $P_w^{010} = p_e^{010}$

For $s = 110$ =>$P_e^{110}(a_{110}, b_{110}) = 1$ => $P_w^{110} = p_e^{110}$

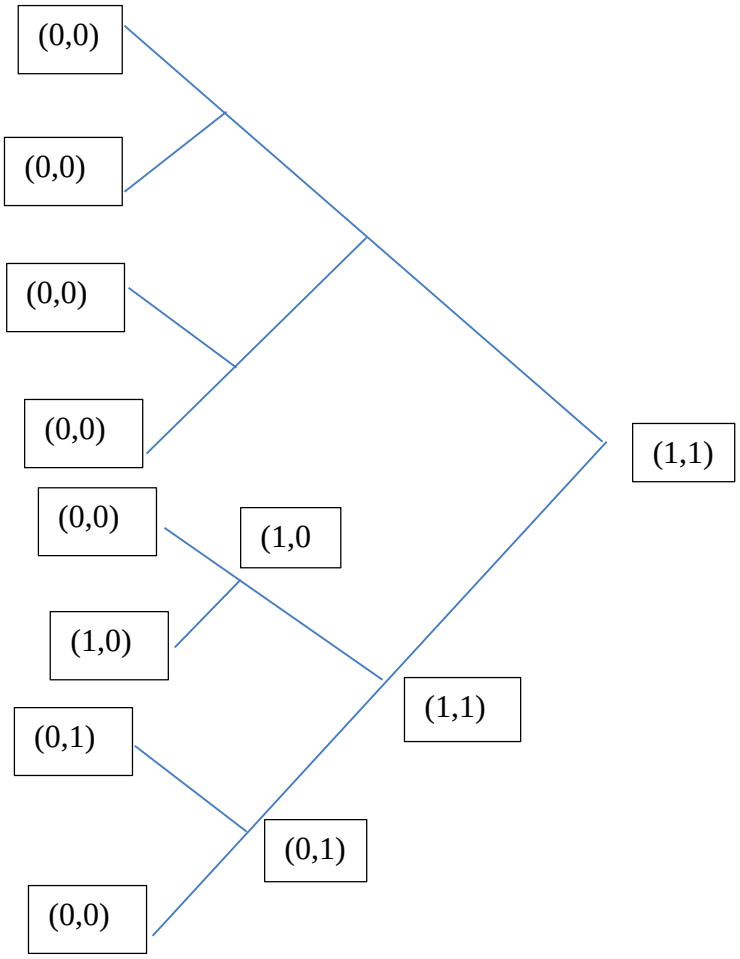
For $s = 10$ =>$P_w^{10} = \frac{1}{2}p_e^{10}(1,0) + \frac{1}{2}P_e^{010}P_e^{110} = \frac{1}{2}$

For $s = 0$ =>$P_w^0 = \frac{1}{2}$ => $P_w^\lambda = \frac{1}{2}$ => $P(x_1 = 0|x_{-2}^0 = 010) = \frac{1}{2}$

⇨ $P(x_1 = 1|x_{-2}^0 = 010) = \frac{1}{2}$;

(2)

Fig A1.2



For $s = 10$ => $P_w^{10} = \frac{1}{2}$

For $s = 00$ => $P_w^{00} = \frac{1}{2}$

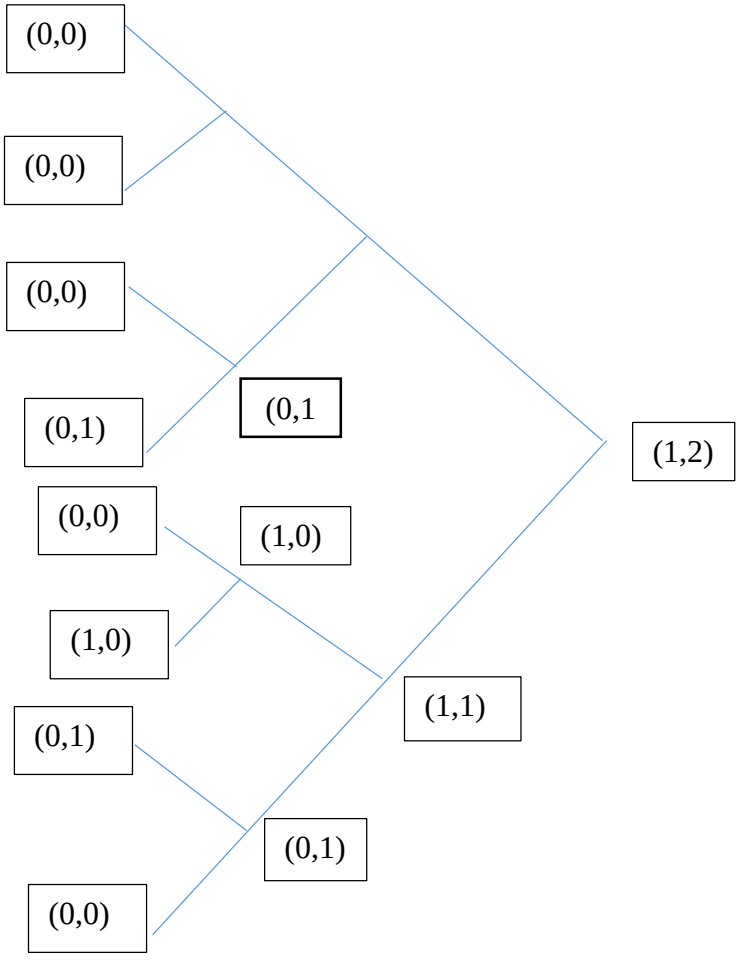
For $s = 0 => P_w^0 = \frac{3}{16} => P_w^\lambda = \frac{5}{32} => \frac{P(x_2^1=10|x_{-2}^0=010)}{P(x_1^1=0|x_{-2}^0=010)} = \frac{5}{16} =>$

$P\ (x_2 = 1|x_1^1 = 0, x_{-2}^0 = 010) = \frac{5}{16} =>$

$P(x_2 = 0|x_1^1 = 0, x_{-2}^0 = 010) = \frac{11}{16};$

(3)

Fig A1.3



For $s = 001 => P_e^{001}(a_{001}, b_{001}) = \frac{1}{2} => P_w^{001} = P_e^{001} = \frac{1}{2}$

For $s = 010$ =>$P_e^{010}(a_{010}, b_{010}) = \frac{1}{2}$ => $P_w^{010} = p_e^{010} = \frac{1}{2}$

For $s = 100$ =>$P_e^{100}(a_{100}, b_{100}) = \frac{1}{2}$ => $P_w^{100} = \frac{1}{2}$

For $s = 00$ => $P_{-w}^{00} = \frac{1}{2}$

For $s = 10$ => $P_w^{10} = \frac{1}{2}$

For $s = 0$ => $P_w^{0} = \frac{3}{16}$

For $s = 01$ => $P_w^{01} = \frac{1}{2}$

For $s = 1$ => $P_w^{1} = \frac{1}{2}$

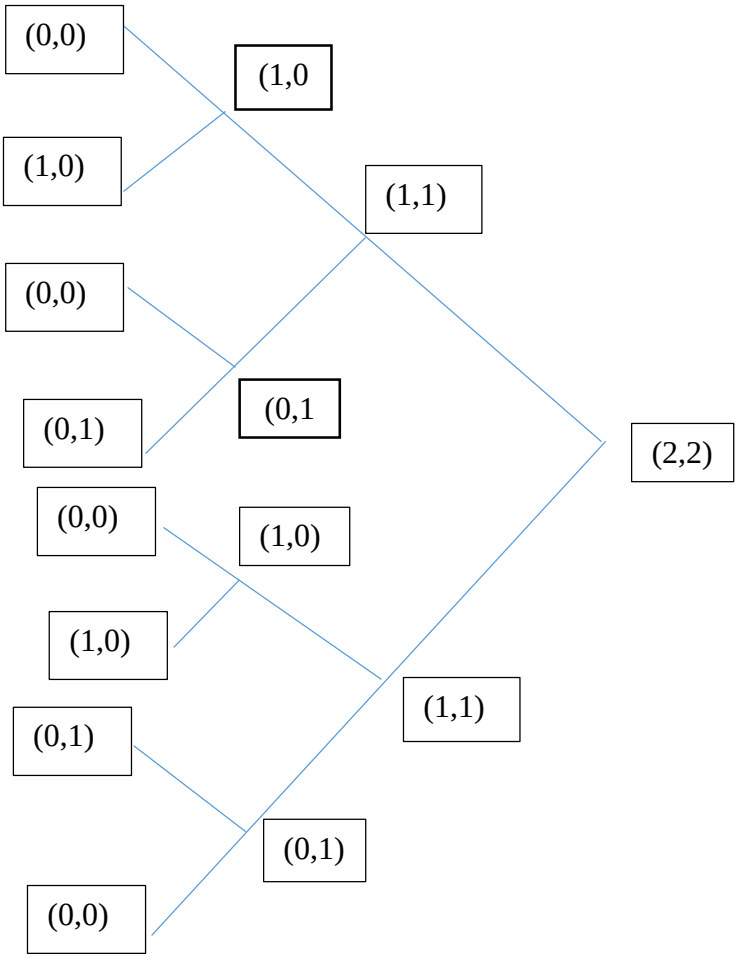
For $s = \lambda$ => $P_w^{\lambda} = \frac{5}{64}$

All above => $P_w^{\lambda}(011|010) = \frac{5}{64}$, $P_w^{\lambda}(01|010) = \frac{5}{32}$

=>$P_w^{\lambda}(x_3 = 1|x_1^2 = 01, x_{-2}^0 = 010) = \frac{1}{2}$;

(4)

Fig A1.4



For $s = 100$ =>$P_e^{100}(a_{100}, b_{100}) = \frac{1}{2}$ => $P_w^{100} = \frac{1}{2}$

For $s = 010$ =>$P_e^{010}(a_{010}, b_{010}) = \frac{1}{2}$ => $P_w^{010} = \frac{1}{2}$

For $s = 00$ =>$P_w^{00} = \frac{1}{2}$

For $s = 10$ =>$P_w^{10} = \frac{1}{2}$

For $s = 0$ =>$P_w^{0} = \frac{3}{16}$

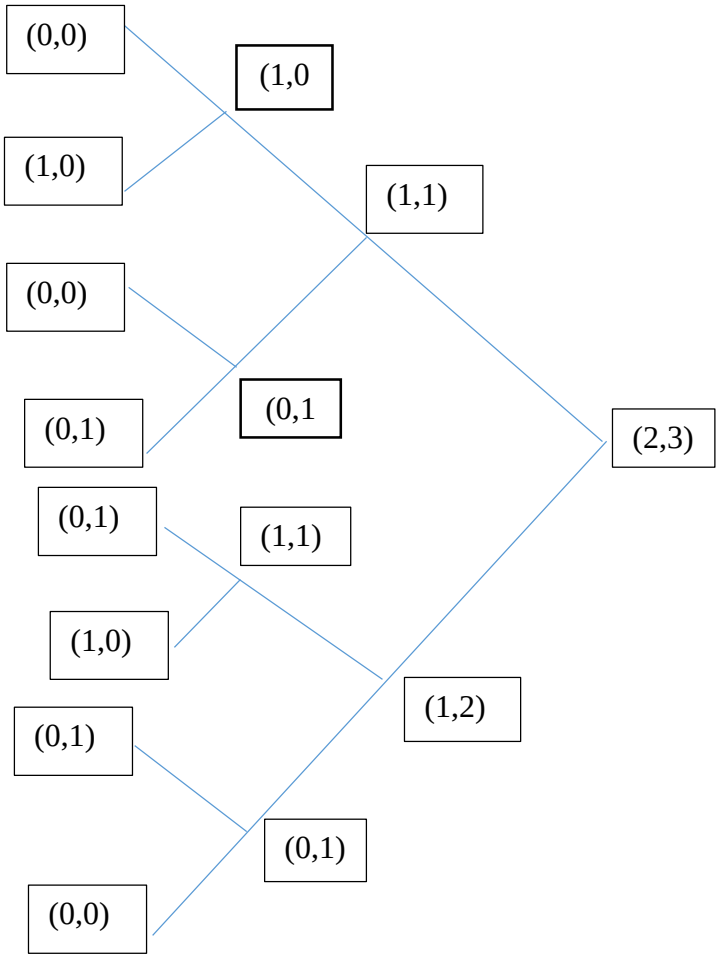
For $s = 1$ =>$P_w^{1} = \frac{3}{16}$

For $s = \lambda$ =>$P_w^{\lambda} = \frac{15}{512}$

=>$P_w^{\lambda}(x_4 = 0 | x_1^3 = 011, x_{-2}^0 = 010) = 0.375$;

(5)

Fig A1.5



For $s = 011$ => $P_e^{011}(1,0) = \frac{1}{2}$ => $P_w^{011} = \frac{1}{2}$

For $s = 11$ => $P_w^{11} = \frac{1}{2}$

For $s = 001 => P_w^{001} = \frac{1}{2}$

For $s = 01 => P_w^{01} = \frac{1}{2}$

For $s = 1 => P_w^{1} = \frac{3}{16}$

For $s = 10 => P_w^{10} = \frac{3}{16}$

For $s = 00 => P_w^{00} = \frac{1}{2}$

For $s = 0 => P_w^{0} = \frac{5}{64}$

For $s = \lambda$ $=> P_w^{\lambda} = \frac{1}{2} P_e^{\lambda}(2,3) + \frac{1}{2} P_w^0 P_w^1$

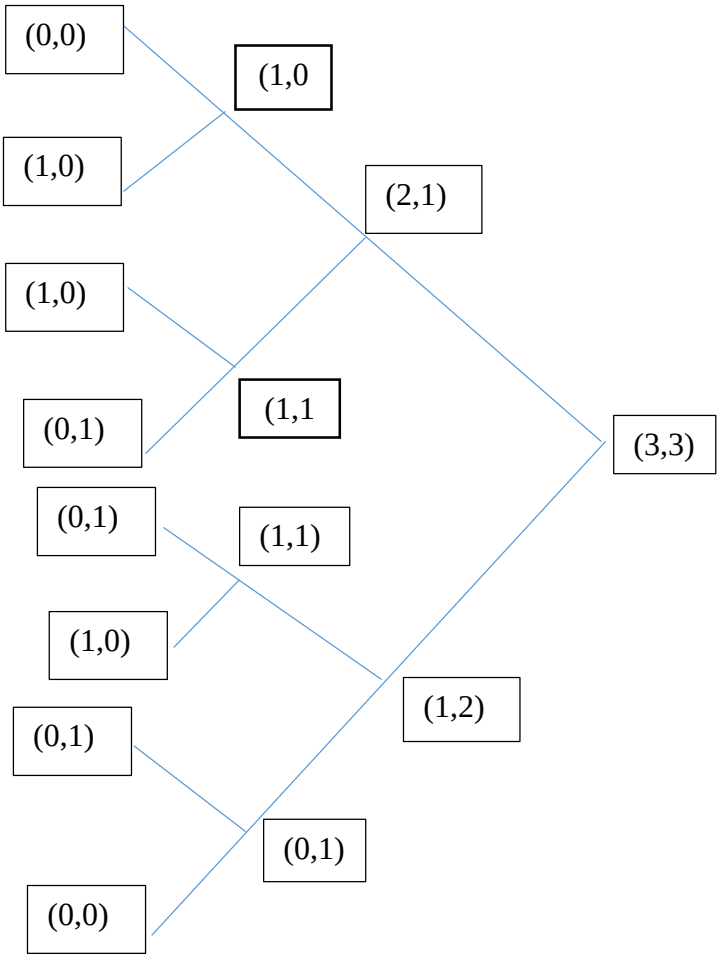
Since $P_e^{\lambda}(2,3) = \frac{3}{4 \times 64}$

⇨ $P_w^{\lambda} = \frac{27}{64 \times 32}$

⇨ $P_w^{\lambda}(x_5 = 1 | x_1^4 = 0110, x_{-2}^0 = 010) = \frac{9}{20}$;

(6)

Fig A1.6



For $s = 111 => P_e^{111}(a_{111}, b_{111}) = P_w^{111}(0,0) = 1$

For $s = 011 => P_e^{011}(a_{011}, b_{011}) = P_w^{011}(0,0) = \frac{1}{2}$

For $s = 11 => P_e^{11}(a_{11}, b_{11}) = \frac{1}{2} => P_w^{11} = \frac{1}{2}$

For $s = 101 => P_e^{101}(a_{101}, b_{101}) = P_e^{101}(1,0) = \frac{1}{2}$

For $s = 001 => P_e^{001}(a_{001}, b_{001}) = P_e^{001}(1,0) = \frac{1}{2}$

For $s = 01 => P_w^{01} = \frac{3}{16}$

For $s = 1 => P_w^1 = \frac{5}{64}$

For $s = 000 => P_w^{000} = 1$

For $s = 100 => P_w^{100} = \frac{1}{2}$

For $s = 00 => P_w^{00} = \frac{1}{2}$

For $s = 110 => P_w^{110} = \frac{1}{2}$

For $s = 010 => P_w^{010} = \frac{1}{2}$

For $s = 10 => P_w^{10} = \frac{3}{16}$

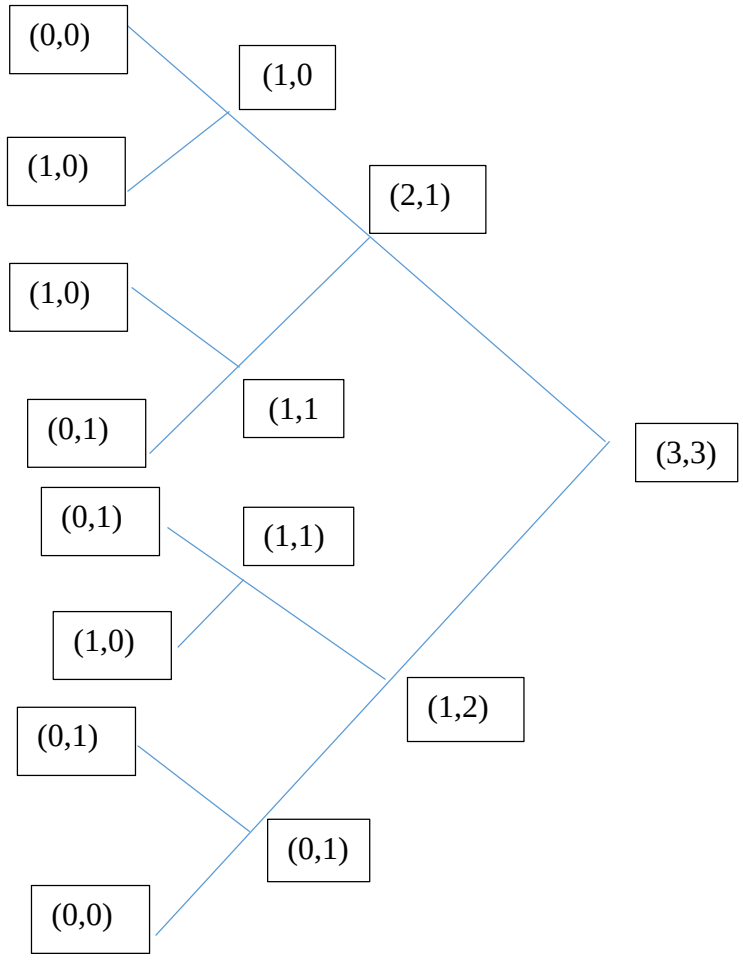
For $s = 0 => P_w^0 = \frac{5}{64}$

For $s = \lambda$ =>$P_w^\lambda = \frac{45}{64^2 \times 2}$

⇨ $P_w^\lambda(x_6 = 0 | x_1^5 = 01101, x_{-2}^0 = 010) = \frac{5}{12}$;

(7)

Fig A1.7



$$P_w^{\lambda} = \frac{95}{32768}$$

⇨ $P_w(x_7 = 0 | x_1^6 = 011010, x_{-2}^0 = 010) = \frac{19}{36}$.