RICE UNIVERSITY

A Data-Driven Information Theoretic Approach for Neural Network Connectivity Inference

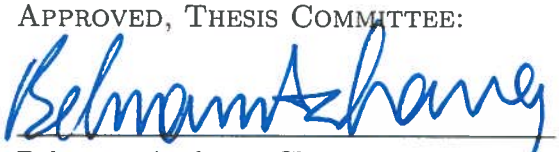
by

Zhiting Cai

A Thesis Submitted
in Partial Fulfillment of the
Requirements for the Degree

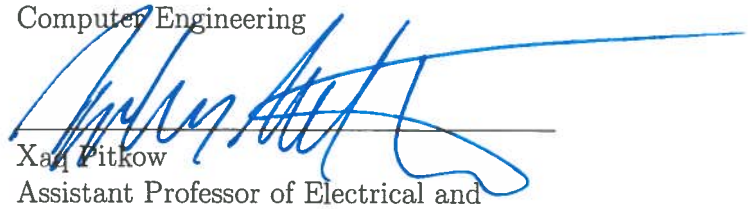
Master of Science

Approved, Thesis Committee:



Behnaam Aazhang, Chair
J.S. Abercrombie Professor of Electrical
and Computer Engineering

Caleb Kemere
Assistant Professor of Electrical and
Computer Engineering



Xaq Pitkow
Assistant Professor of Electrical and
Computer Engineering

Houston, Texas

April, 2017

ABSTRACT

A Data-Driven Information Theoretic Approach for Neural Network Connectivity Inference

by

Zhiting Cai

A major challenge in neuroscience is to develop effective tools that infer the circuit connectivity from large-scale recordings of neuronal activity patterns, such that we can study how the structures of neural networks enable brain functioning. To tackle this challenge, we used context tree maximizing (CTM) to estimate directed information (DI), which measures causal influences among neural spike trains in order to infer synaptic connections. In contrast to existing methods, our method is data-driven and can readily identify both linear and nonlinear relations between neurons. This CTM-DI method reliably identified circuit structures underlying simulations of realistic conductance-based networks. It detected direct connections, eliminated indirect connections, quantified the amount of information flow, reliably distinguished synaptic excitation from inhibition and inferred the time-course of the synaptic influence. From voltage-sensitive dye recordings of the buccal ganglion of Aplysia, our method detected many putative motifs and patterns. This method can be applied to other large-scale recordings as well. It offers a systematic tool to map network connectivity and to track changes in network structure such as synaptic strengths as well as the degrees of connectivity of individual neurons, which in turn could provide insights into how modifications produced by learning are distributed in a neural

network. Furthermore, this information theoretic approach can be extended to the analysis of other recordings that can be modeled as point processes, such as internet traffic, disease outbreak, and seismic activity.

Contents

Abstract i

List of Illustrations iv

List of Tables v

1 Introduction 1

2 Directed Information 5

2.1 Definition 5

2.2 Directed Information Rate 6

2.3 Convergence 7

2.4 Eliminating Indirect Connections 7

2.5 Calculating Final DI Values 10

3 Context Tree Estimation 12

3.1 Tree Structure 12

3.2 Estimation of Leaf Parameters 15

3.3 Maximum A Posteriori Tree Model with Penalties 16

3.4 Joint Probability Estimation for Multiple Sequences 17

3.5 Reconstructing the Synaptic Profile from the Tree 18

4 Validation and Results 21

4.1 Sparse Poisson Spiking Model 21

4.2 Simulated Neural Networks 24

4.3 Mapping Connectivity of Recorded Neurons 30

5 Conclusions **35**

Bibliography **38**

A VSD Recording Technique **46**

B Proofs **48**

B.1 Sequential KT Estimator 48

B.2 Incorporating Penalties into Recursive Model Finding 49

B.3 Property of the Profile Estimator 50

B.4 Eliminating Indirect Connections 52

Illustrations

2.1 Two fundamental types of indirect connections. 8

2.2 Receiver operating characteristic (ROC) curves plotted as true/false positive rates against threshold. 11

3.1 A binary context tree structures with maximum depth 3. 14

3.2 Synaptic profiles illustrating the time course of the synaptic action and distinguishing excitatory vs. inhibitory synaptic actions. 18

4.1 Trends of normalized DI tested under the sparse Poisson spiking model. 23

4.2 DI correctly inferred the connectivity in three simple networks. 25

4.3 DI tested on a network with synaptic plasticity. 27

4.4 Testing a conductance-based model of the central pattern generator (CPG) in the buccal ganglion of *Aplysia*. 28

4.5 Analyzing VSD recording data using DI. 31

4.6 Patterns of connectivity of the preparation in Fig. 4.5. 32

Tables

4.1 DI Performance on Simulated CPG Network 30

Chapter 1

Introduction

Understanding how the organization of neurons into neural circuits enables the different functions of the brain is one of the core goals of neuroscience and is a prerequisite for studying how the structures of these networks are modified by learning. Major advances have been made in the methods and techniques for simultaneously recording activity in large numbers of neurons [1–3]. With the ability to collect a large volume of data, the next step is to reverse-engineer the neural signals and to delineate the underlying circuits that have generated the activity. Integrating the techniques of large-scale recordings with analytical tools would contribute tremendously to delineating and deciphering functional connectomes. Functional connectivity provides greater insights than anatomical connectivity because it captures the active functional structure of the circuit, the strengths of different neural pathways, and the relevance of various neurons in the network. The main focus of this project is to develop and test a tool that can be used to reliably detect functionally relevant connections using a scalar metric based on the information provided in the neuronal recordings alone.

Several statistical or information theoretic tools have been used to infer the directed functional connectivity of a neural circuit [4]. In general, two different types of signals are analyzed by dedicated methods: methods that focus on inferring connectivities using local field potentials (LFPs) [5–8], and methods that focus on spiking activities from neuronal-level recordings [9–15]. There are also methods that can be applied to both signal types by analyzing the interactions among predefined states

of the recordings [16, 17]. Among methods that analyze spiking signals, one of the most commonly used tools is cross-correlation histogram (cross-correlogram), which deploys spike-triggered histograms to find the causal relationship between two neurons [9, 10]. [11] and [12] employ a point process-generalized linear model framework together with Granger Causality (GC), whereas [13] and [18] calculate directed information (DI) based on the same framework to detect pairwise causal influences; similarly, [14] uses the coupling strengths obtained from the spline coefficients fitted with the generalized linear model to reconstruct functional connectivity.

Among all the above-mentioned techniques, directed information has many advantages. Cross-correlation and cross-correlogram are defined on two processes and cannot be extended to analyze larger structures. Granger Causality assumes that the past samples in the recording have a linear influence on the future sample and that noise in the signal is modeled as a Gaussian distribution [19]. DI, on the contrary, is itself model-free, because its calculation is based simply on entropy [20]. The burden of reducing the estimation error of DI is hence shifted to the estimation of entropy for neural recordings. In order to fully exploit the advantages of DI, an accurate and data-driven entropy estimator should be used in conjunction with DI.

Neural spike trains are commonly used to represent spiking signals and are generated by segmenting the continuous time spiking data using a predetermined unit time called bin width into discrete time binary sequences, where a 1 bit indicates a spike and a 0 indicates no activity. Generally, the statistical properties of neural spike trains are estimated through a parametric approach, the generalized linear model (GLM) [11–13, 18]. GLM assumes that the likelihood of a future spike is modeled as the exponent of the linear combination of past activity of the spike trains. Because spiking activity may not always fit the GLM assumption, a data-driven approach to

extract statistical properties without assuming a linear relationship among spikes in one or more neural recordings is more desirable. Context tree weighting (CTW), a universal entropy estimators developed for data compression [21], is a tool that does not assume a linear relationship among data in neural spike trains and thus may serve as a good technique to infer functional connectomes. Yet the depth of the tree used to determine the length of the memory needs to be set externally, usually arbitrarily or by past experience. Also, CTW assumes that all firing patterns are equally likely, and this method of assigning all patterns equal weight is very data-intensive and does not provide insight into the data structure. In this thesis, we implement a different method that utilizes the context tree framework – context tree maximizing (CTM), which, in addition to being data-driven, also automatically finds the appropriate tree depth as well as the best tree structure that fits the data in the *a posteriori* sense, and prevents overfitting [22, 23], in which case the model tends to fit the noise other than the underlying patterns and relationships of the signals. Because it is data-driven, CTM is not constrained by model types and is able to detect both linear and nonlinear relationships between spike train sequences.

In this thesis, we begin by reviewing the theoretical basis for directed information for discrete time series. We then estimate DI for spike train data using CTM and construct a synaptic profile from the tree model that allows us to view the synaptic influence with different kinetics and time-courses as well as to differentiate excitatory and inhibitory connections. Next, we use a heuristic to determine direct vs. indirect connections. To demonstrate the robustness of our approach, we test it using a sparse Poisson spiking model and several small realistic conductance-based neuronal circuits. The performance is further tested using a larger network that resembles the central pattern generator (CPG) network of the *Aplysia* feeding circuit. And finally, we apply

the technique to actual data obtained from voltage-sensitive dye (VSD) recordings of a buccal ganglion from *Aplysia* and identify some promising putative connections.

Our method of constructing connectivity diagrams from large-scale recordings is an automated, comprehensive framework for inferring neuronal network structures. It is robust against synaptic plasticity such as facilitation and depression. It is able to exclude indirect connections, differentiate excitatory from inhibitory synapses and infer the time course of the synaptic responses. This method is generally applicable to analyzing neural network structures and could be used to track functional changes due to neuromodulation or learning. Preliminary results of this work were reported in abstract form [24, 25] as well as in a conference paper [15].

Chapter 2

Directed Information

Directed information will be used to quantify information flow from one neuron to another. It is an entropy-based measure that bears much resemblance to mutual information and can be applied to both discrete and analog random processes. We will focus on the first case, because we segmented spikes in time using fixed bin widths into neural spike trains and the random process used to represent these discretized neuronal activities is then a discrete time and binary alphabet random process.

Throughout the thesis, X_a^b for $b > a$ is a shorthand for the vector $[X_a X_{a+1} ... X_{b-1} X_b]$, whereas X^b is simply the string of random variables X_1^b from the beginning up to index b. An uppercase letter denotes a random variable, whereas a lowercase denotes one realization of that random variable. We denote a string $s = x_{b-k}^b$, and then $s' = x_{b-k'}^b$ where $k' \leq k$, is a suffix of s, denoted by $s \succeq s'$; yet if $k' < k$, s' is called a proper suffix of s, denoted by $s \succ s'$.

2.1 Definition

Directed information was originally formulated by H. Marko and formally defined by J. Massey [20]. It measures the amount of single-directional information flow from random sequence X to sequence Y. It is defined as

$$I(X^n \to Y^n) = H(Y^n) - H(Y^n || X^n), \tag{2.1}$$

where

$$H(Y^n) = \sum_{i=1}^{n} H(Y_i|Y^{i-1}) \tag{2.2}$$

is the chain rule of entropy quantifying the entropy of Y itself, and

$$H(Y^n||X^n) = \sum_{i=1}^{n} H(Y_i|Y^{i-1}, X^i) \tag{2.3}$$

is the causally conditioned entropy. Causally conditioned entropy is the entropy of Y conditioned on the causal part of X in addition to the history of itself. In the formulation of mutual information, the conditional entropy term in Eq. 2.3 is $H(Y_i|Y^{i-1}, X^n)$ instead. DI quantifies the reduction in entropy given the causal part of X in addition to the history of Y.

2.2 Directed Information Rate

In practice, directed information rate, which is defined as

$$\bar{I}(X \to Y) = \lim_{n\to\infty} \frac{1}{n} I(X \to Y), \tag{2.4}$$

is most commonly used, because it is bounded by the largest entropy a random variable can achieve. In binary spike trains, the bound is 1. With a slight abuse of terminology we are going to refer to DI rate as DI for the rest of the thesis. Both the entropy rate for Y alone, as well as the causally conditioned entropy rate, can be estimated using entropy estimators that directly provide the $H(Y_i|Y^{i-1}, X^i)$ terms. Such approach is demonstrated by estimator 1 and 3 in [21], where the asymptotic equipartition property (AEP) is evoked and entropy rate or divergence rate is estimated directly. It is also possible to estimate the entropy rates using plug-in estimators that approximate the probability distribution $P(Y_i|Y^{i-1}, X^i)$, with which entropy rates can later be calculated, such as in estimator 2 and 4 in [21]. In this thesis, we focused

on estimator 1, for its faster convergence rate based on our simulations. We used a context tree based algorithm to estimate causally conditioned entropy. Causally conditioned entropy rate obtained via AEP is shown to converge in the *almost sure* sense [26] as well as in the L_1 sense [21]. Almost sure sense convergence is when the probability of the estimate and the true distribution being the same goes to 1. L_1 convergence is that the expected value of the absolute error goes to 0.

2.3 Convergence

In actual implementations, convergence of DI implemented with context tree maximizing is very data dependent. For neurons with a relatively uniform firing pattern (i.e. the presynaptic neuron is self-actuating and homogeneous) the convergence is faster. The oscillation is $< 0.1\%$ with data length in the order of 10^3. In this favorable case, if the bin width is 10 ms, typically less than 1 min of data is needed. For neurons with vastly differently phases (i.e. bursting), the DI rate curve stabilizes slower. Based on simulations, the number of data points needed is in the order of 10^4, which corresponds to 2-10 min of data depending how diverse the spiking patters are.

2.4 Eliminating Indirect Connections

A positive directed information value between two neurons does not guarantee an actual direct link between these two neurons. In fact, the information can flow through an indirect route. [13] describes two fundamental structures where indirect connections could be incorrectly identified as direct connections: the cascade structure and the proxy structure (Fig. 2.1). In the proxy configuration, the path of information flow is from X via Z to Y, but a false connection from X to Y could be detected. In the cascade configuration, neuron Z drives neuron X and neuron Y through two dif-

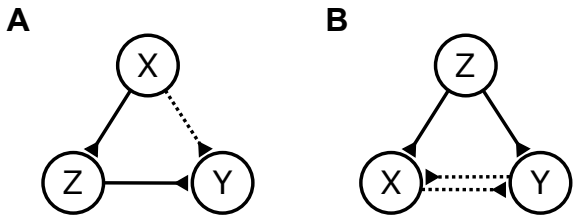


Figure 2.1 : **Two fundamental types of indirect connections.** ***A:*** Proxy configuration. The path of information flow is from X to Z to Y, yet a false connection from X to Y can be detected. ***B:*** Cascade configuration. Neuron Z is driving neurons X and Y through two different paths, and a false connection can be detected between X and Y.

ferent paths, but a false connection could be detected between X and Y. To address this issue, [13] employs Kramer's concept of causally conditioned directed information (CCDI) [27], which is defined as

$$I(X^n \to Y^n||Z^n) \triangleq H(Y^n||Z^n) - H(Y^n||X^n, Z^n) \tag{2.5}$$

The interpretation of this measure is very intuitive. If a connection between X and Y is suggested by DI and yet Z is the agent that actually directly influences Y, then the entropy estimate of Y knowing Z alone can account for the external information Y receives, and knowing X additionally would not yield any more information, and therefore, would not reduce the entropy further. On the other hand, if $I(X^n \to Y^n||Z^n) > 0$, the connection between X and Y is direct and should be kept in the graph. Any context tree estimation method can easily estimate the $H(Y^n||X^n, Z^n)$ term by joining the bits of the spike trains from X, Y and Z to form an alphabet of size 8, an example of which is $W = X + 2Y + 4Z$.

Strictly speaking, to identify whether a connection is direct, it is necessary to calculate CCDI simultaneously conditioned on all other neurons, which, on one hand, creates a forbiddingly large number of states, and on the other, demands a large

amount of data to estimate those states. However, a heuristic is employed here analyzing small triangular structures. Evoking data processing inequality (DPI) for DI (see Theorem 2), it is sufficient to perform CCDI analysis on all groups of three neurons that form the structures identified in Fig. 2.1 to eliminate all single-path indirect connections.

Definition 1 Random sequences U^n, V^n *and* W^n *form a directed causal chain if* $I(U^n \to V^n) > 0$, $I(V^{n-1} \to U^n) = 0$, $I(V^n \to W^n) > 0$, $I(W^{n-1} \to U^n) = 0$, *and* $I(U^n \to W^n || V^n) = 0$. *With a slight abuse of notation, we denote this causal chain as* $U^n \to V^n \to W^n$. *Using the notation of Markov chains, this relationship is expressed as* $U^i \to V_i$, $V^i \to W_i$ *for* $i = 1, 2, ..., n$.

Theorem 1 (DPI for DI) If $U^n \to V^n \to W^n$ *causally, then* $I(V^n \to W^n) \geq I(U^n \to W^n)$ *and* $I(U^n \to V^n) \geq I(U^n \to W^n)$.

The proof of Theorem 1 can be found in Appendix B.4.

We define a "1-relay" link as an indirect link with one intermediate relay neuron, a "2-relay" link as an indirect connection with two intermediate neurons, and so on. A "1-relay" link conveys more information than a "2-relay" link in signal transmission with noise, according to the data processing inequality. Therefore, by calculating CCDI on all groups of three and deleting links with 0 CCDI values, we can eliminate all single-path indirect connections. The limitation of this heuristic is that if in case multi-path indirect connections arise, the indirect link might not be eliminated, and conditioning on more neurons is required.

2.5 Calculating Final DI Values

For every circuit analyzed, continuous time spiking signals were converted into discrete time spike trains using bin widths ranging from 2 ms to 30 ms with an increment of 1 ms. The DI algorithm was executed on all these spike trains. This range of bin widths was chosen in order to survey a sufficient amount of history to capture causal influences. From this, we plotted the values of DI vs. bin widths, where true connections had DI curves that plateaued after an initial rise (Fig. 4.3C). If the DI curve for a connection had at least four consecutive values larger than the 0.01 threshold, all non-zero entries were averaged to produce the final DI value. In cases where four consecutive values > 0.01 did not occur, DI was set to 0.

The threshold 0.01 was chosen based on the Receiver operating characteristic (ROC) curves generated from simulated neuronal spike trains. A typical excitatory and an inhibitory synapse were constructed and tested. There was a unidirectional forward connection from neuron X to neuron Y, and therefore the connection from Y to X served as the null hypothesis. Each simulation consisted of 1000 trials of 12000 (120 s with a 10 ms bin width) data points. True positives as well as false positives were plotted against the threshold (Fig. 2.2). The threshold should be chosen where the difference between true positive and false positive was the greatest, which was 0.01. Further more, this threshold yielded higher accuracy experimentally on the conductance based models.

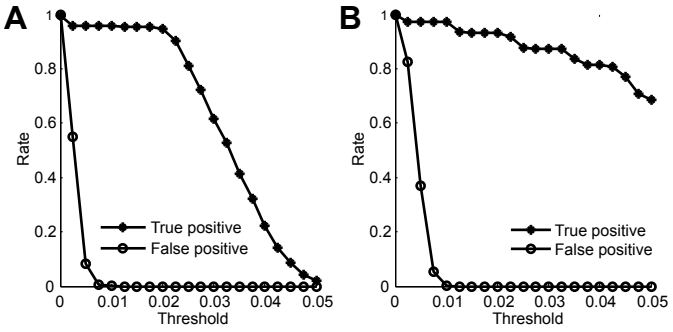


Figure 2.2 : **Receiver operating characteristic (ROC) curves plotted as true/false positive rates against threshold. The threshold whose true positive rate nad false positive rate have the largest margin was chosen.** ***A:*** The ROC curve of a typical excitatory synapse. The largest margin was acheived at Threshold=0.01. ***B:*** The ROC curve of a typical inhibitory synapse. The largest margin was also acheived at Threshold=0.01.

Chapter 3

Context Tree Estimation

To calculate DI, the conditional entropy terms in Eq. 2.1 need to be estimated. Multiple entropy or plug-in probability estimators are available, which aim to detect patterns in data sequences and to minimize entropy, in the area of data compression. Examples include Lempel-Ziv [28], Burrows-Wheeler transform (BWT) [29] and prediction by partial matching (PPM) [30]. However, context tree based algorithms show a faster convergence [31]. In this section, we provide a brief description of how a tree model generates a sequence. If we know that tree model entirely, the likelihood of the observed sequence as well as its conditional probability can be calculated. Yet if the model is unknown, it is necessary to first use the sequence to estimate a tree model that most likely (in the *a posteriori* sense) has generated the given sequence, and with this estimated model, we can then obtain the probability measure we need to calculate directed information.

3.1 Tree Structure

Tree structures are commonly used to model finite-alphabet, finite memory, stationary and ergodic sources that have generated the observed sequences. We denote a unique tree structure by $\mathcal{T}$. Assume that a fictive tree model $\mathcal{T}$ has depth D. Its symbols are drawn from a finite alphabet $A = \{0, ..., |A| - 1\}$, and the cardinality (also known as the size) of the alphabet is $|A|$. For a complete tree model, it then has D levels, and

on each level, every node splits up into $|A|$ branches, and therefore, there are $|A|^D$ leaf nodes (see Fig. 3.1A for a simple example). Note that this complete tree corresponds to a Markov chain model of order D. Each leaf defines the complete path from the root to a leaf, with the segments closer to the root being more recent and the ones closer to the leaves older. It represents a context that is unique and is independent of all others. A context is denoted by s and represents the history of spike activity with a duration of depth D times bin width. The tree $\mathcal{T}$ is formed by all its contexts. Associated with each leaf there is a parameter vector $\boldsymbol{\theta}_s$, an $|A|$-dimensional simplex $[\theta_s(0), ..., \theta_s(|A|-1)]$, with each entry dictating the probability of the next symbol being a, where $a \in A$. Let y^n be a sequence generated by this tree $\mathcal{T}$. If all the digits y_i in y^n that follow the same context s are grouped into a new sequence y_s, then the subsequent $y_s = \{y_i | y_i \in y^n, y_{i-D}^{i-1} = s\}$ emitted by the same leaf is modeled as an independently identically distributed (i.i.d.) process. Furthermore, $y_{s'}$ is independent of y_s if $s' \neq s$, for $s, s' \in \mathcal{T}$. Because each leaf is modeled as an i.i.d. source, the probability of an outcome y^n using the known tree model $\mathcal{T}$ is

$$P_{\mathcal{T}}(y^n) = \prod_{s \in \mathcal{T}} P(y_s) \tag{3.1}$$

$$= \prod_{s \in \mathcal{T}} \prod_{a=0}^{|A|-1} P(a|s)^{c_s(n,a)} \tag{3.2}$$

where $c_s(n, a)$, $a \in A$, is the count of all occurrences of symbol a that directly follow the context s.

However, a tree structure does not have to be complete, and this is one advantage over the Markov chain model. If the lengths of the branches are allowed to vary, letting some longer contexts to be merged into one shorter context, the number of contexts can be markedly reduced (Fig. 3.1B) and therefore the model's complexity is also substantially reduced. The tree model can be simplified as long as the tree is

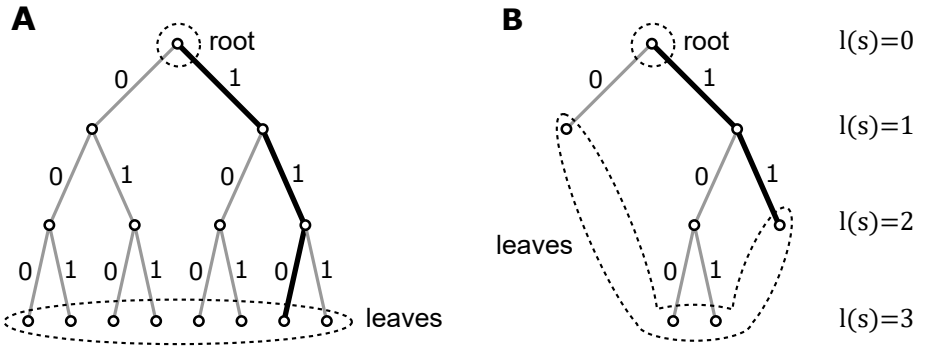


Figure 3.1 : **A binary context tree structures with maximum depth 3.** ***A:*** A complete tree with leaves associated with all possible $2^3 = 8$ contexts. All contexts are equally long with 3 digits. The highlighted path refers to $(Y_{n-3} = 0, Y_{n-2} = 1, Y_{n-1} = 1)$, which also shows that the branch segment closer to the root corresponds to a newer sample (y_{n-1}) and the segment closer to the leaf corresponds to an older one (y_{n-3}). ***B:*** A trimmed tree with only 4 contexts. Note that any trimmed branch cannot be a suffix of another branch. In this case, both context ...011 as well as context ...111 are mapped to the same branch – the highlighted branch ...11

irreducible, which means, no branch can be a suffix of another. The entire class of irreducible tree models is denoted by $\mathcal{I}$. This requirement is automatically guaranteed by the tree structure itself.

In many scientific applications such as neural spike trains, the real model that has generated the observation is never known. An intermediate step is to assume that we know the tree structure $\mathcal{T}$ but do not know the leaf parameters. In this scenario, the leaves could be used to partition the observation y^n into individual y_s. Because the subsequence y_s corresponds to each context s is i.i.d., an appropriate estimator for a memoryless source can be used to find its parameters.

3.2 Estimation of Leaf Parameters

In this work, leaf parameter estimation was accomplished through the Krichevskii-Trofimov (KT) estimator [32]. KT estimator is a Bayesian estimator and it also has the capability to be implemented sequentially for potential real time applications. Suppose a sequence X^n is i.i.d and each variable X_i can only take on values from a finite-sized alphabet A; therefore, x^n is generated by a multinomial with parameter $[\theta(0), ..., \theta(|A|-1)]$. Denote a symbol in the alphabet by a, for $a \in \{0, 1, ..., |A|-1\}$. Let the count of each symbol a observed prior to index n be $c(n, a)$, but for simplicity, we denote it as $c(a)$. KT estimator produces an estimate of the probability of an entire realization x^n of a stationary memoryless string using Bayesian statistics (See Appendix B.1 for details). Let us denote the estimate of such an i.i.d. sequence by $\hat{P}(x^n)$. It can be computed sequentially as:

$$\hat{P}(X_n = a, x^{n-1}) = \frac{c(a) + \gamma}{c(0) + c(1) + ... + c(|A|-1) + \gamma|A|} \hat{P}(x^{n-1}) \tag{3.3}$$

$$= \frac{c(a) + \gamma}{(n-1) + \gamma|A|} \hat{P}(x^{n-1}) \tag{3.4}$$

starting with $\hat{P}(\phi) = 1$, which means an empty string starts with probability 1 (see Appendix B.1). Here γ is the "add-something" parameter of the sequential estimator [33]. Most often $\gamma = \frac{1}{2}$ such that the error of the estimated log likelihood is uniformly bounded [32].

In the framework of context tree estimation, for a Markov chain process Y^n with order D, we simply have for each context s

$$\hat{P}(Y_n = a | Y_{n-D}^{n-1} = s) = \hat{P}(a|s) = \frac{c_s(a) + \frac{1}{2}}{(n_s - 1) + \frac{|A|}{2}} \tag{3.5}$$

which is the conditional probability needed for Eq. 3.2.

3.3 Maximum A Posteriori Tree Model with Penalties

After estimating the leaf parameters $\boldsymbol{\theta}_s$ of a tree structure, we next searched from the entire set of all irreducible tree models $\mathcal{I}$ the tree structure that described the observed data best in the *a posteriori* sense. Because KT estimator is used to produce an estimate $\hat{P}(a|s)$, maximizing $\hat{P}_{\mathcal{T}}(y^n)$ in Eq. 3.2 among all assumed model $\mathcal{T}$'s yields the maximum a posteriori (MAP) estimator, which is equivalent to minimizing the negative log likelihood $-\log \hat{P}_{\mathcal{T}}(y^n)$. Negative log likelihood of a sequence is the number of bits needed to encode that sequence in the field of compression. Intuitively, the lower the number, the better fit the model has. However, in an effort to control complexity and prevent overfitting, we want to penalize models with higher orders and find a model with limited tree depth D_0, $D_0 \leq D$. We can introduce a minimum description length (MDL) criterion that takes into account the *cost* of the model: the number of bits needed to describe the tree model itself including both the parameters as well as the tree structure. Define the cost of the model to be

$$\Gamma_{\mathcal{T}} = (|\mathcal{T}| + |u : u \prec s|) \cdot \log |A| \tag{3.6}$$

where $|\mathcal{T}|$ is the number of contexts (i.e. the number of leaves); $|u : u \prec s|$ is the total number of inner nodes. A tree with the same number of leaves but a higher order requires more bits to detail all the layers of the longer branches. This condition often arises in compression where an optimal trade-off between the code length of the sequence and the cost of the model is desired [22]. With this penalty term, we construct our objective function as a trade off between the model complexity and finding a tree model that maximizes the *a posteriori* probability:

$$\hat{\mathcal{T}}(y^n) = \arg\min_{\mathcal{T}\in\mathcal{I}}\{-\log \hat{P}_{\mathcal{T}}(y^n) + \Gamma_{\mathcal{T}}\}, \tag{3.7}$$

minimized among set $\mathcal{I}$. This objective function can be solved recursively and the penalty term can be readily broken down and incorporated into the recursive optimization process [22]. We define the maximized probability $\hat{P}_s^*$ at node s as:

$$\hat{P}_s^*(y^n) = \begin{cases} \max\{\frac{1}{|A|}\hat{P}_s(y^n), \frac{1}{|A|}\prod_{a=0}^{|A|-1}\hat{P}_{s_a}^*(y^n)\}, & 0 \leqslant l(s) < D \\ \frac{1}{|A|}\hat{P}_s(y^n), & l(s) = D \end{cases} \tag{3.8}$$

Here s_a is a child node of s, which represents a string with symbol a appended to the end of the string represented by s. $\hat{P}^*$ at the root level is the maximized probability for this sequence. The recursive process defined by Eq. 3.8 is equivalent to solving the optimization problem Eq. 3.7, which is expanded in details in Appendix B.2. We can interpret the recursive maximizing process this way: if $\hat{P}_s(y^n) \geqslant \prod_{i=0}^{|A|-1}\hat{P}_{s_a}^*(y^n)$, branches below s are trimmed as in Fig. 3.1B. Assume the depth D of the model we start with is deeper than that of the actual source. It is worth noticing, however, that the magnitude of D is limited by the amount of data points and is actually conveniently bounded by a function of the length of the data n, which is $D(n) = o(\log n)$ [23].

This method is the so-called Context Tree Maximizing algorithm [34]. Although CTM is not a consistent estimator in general, it has a very low computational complexity as well as a low memory requirement, and, by penalizing complex models, it mitigates the problem of overfitting.

3.4 Joint Probability Estimation for Multiple Sequences

Because neural spike trains are discrete time binary sequences where $|A| = 2$, a binary context tree is used to estimate the probabilities of individual neurons. However, in order to calculate the causally conditioned entropy term in directed information, the individual terms $P(y_i|y^{i-1}, x^i)$ is needed to calculate $H(Y_i|Y^{i-1}, X^i)$ for Eq. 2.3.

The most common way to estimate joint probability is to augment X with Y. Let $Z = X + 2Y$ and conduct context tree estimation algorithm on Z, whose alphabet size is then 4. Context tree estimation is executed on this new sequence Z to find $\hat{P}(z_i|z^{i-1})$. In fact, $\hat{P}(z_i|z^{i-1}) = \hat{P}(x_i, y_i|x^{i-1}, y^{i-1})$. To obtain the $\hat{P}(y_i|x^{i-1}, y^{i-1})$ term needed for directed information we simply take the marginal about X. This way, we can calculate the DI, which indicates the strength of the influence, from one neuron to another.

3.5 Reconstructing the Synaptic Profile from the Tree

In addition to estimating the strength of information flow through a synaptic connection quantified by directed information, it is also essential to distinguish excitation

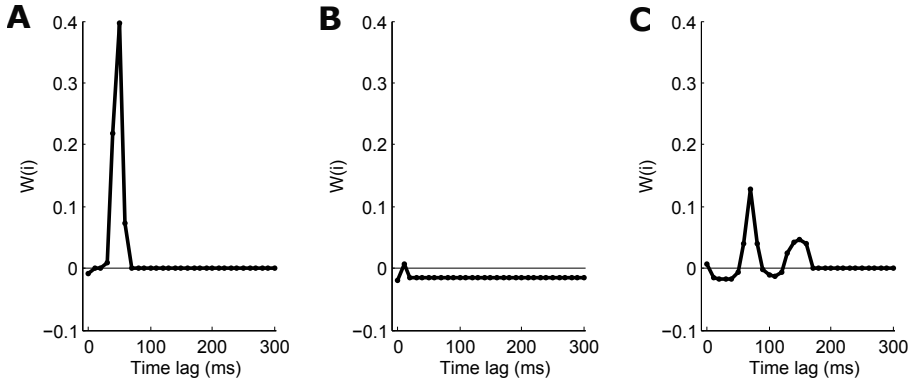


Figure 3.2 : **Synaptic profiles illustrating the time course of the synaptic action and distinguishing excitatory vs. inhibitory synaptic actions.** ***A:*** The profile of synapse AB from the network illustrated in Fig. 4.3A. The bin width used in this example was 10 ms. The influence from A to B is fast and strong. The values are positive, which indicate that the synapse is excitatory. ***B:*** The profile of synapse BD from Fig. 6A. The influence from B to D is negative suggesting an inhibitory connection. It is weak yet has a longer duration as compared to AB. ***C:*** The profile of synapse B4-B51 from Fig. 4.4. B51 is initially inhibited by B4 and yet exhibits post-inhibitory rebound (PIR) afterwards.

from inhibition as well as to infer the synaptic profile. We define synaptic profile as the time course of the synaptic action. It is a set of parameters that depict the relative impact of spikes with respect to time lags in neuron X on the likelihood of observing a spike in neuron Y. We specifically examined whether a 1 bit (spike) in X leads to a higher probability of Y having a 1 (excitation) or lower probability of a 1 bit (inhibition) compared to the average firing rate of Y. This problem of extracting the synaptic parameters only pertains to binary neural spike trains.

From the tree structure estimated using the joint sequence $Z = X + 2Y$ discussed in Section 3.4, we can obtain how likely Y will be a 1 seeing certain contexts, which is defined by $P(Y_n = 1|X_{n-D_0}^{n-1}, Y_{n-D_0}^{n-1})$, where D_0 is the depth of the longest branch of the truncated tree. To describe the influence from the context digit i, we need to find $P(Y_n = 1|X_{n-i} = 1)$ for each $i \in \{0, ..., D_0\}$. $P(Y_n = 1|X_{n-i} = 1)$ can be obtained by taking the marginal of the target context index i:

$$\hat{P}(Y_n = 1|X_{n-i}) = \sum_{\substack{\forall y_{n-k},\ k\in\{1,\ldots,D_0\} \\ \forall x_{n-k},\ k\neq i}} \hat{P}(Y_n = 1|X_{n-D_0}^{n-1}, Y_{n-D_0}^{n-1}) \times \hat{P}(X_{n-D_0}^{n-1} \backslash X_{n-i}, Y_{n-D_0}^{n-1}|X_{n-i}) \tag{3.9}$$

where

$$\hat{P}(X_{n-D_0}^{n-1} \setminus X_{n-i}, Y_{n-D_0}^{n-1}|X_{n-i}) = \frac{\{\text{count of context } X_{n-D_0}^{n-1}, Y_{n-D_0}^{n-1}\}}{\{\text{count of context } X_{n-i}\}} \tag{3.10}$$

Theorem 2 Two binary sequences X^n and Y^n are both Markov chains, of order D_X and D_Y respectively. Then, for $i \leq \max\{D_X, D_Y\}$,

$$\lim_{n\to\infty} \hat{P}(Y_n|X_{n-i}) - P(Y_n|X_{n-i}) = 0 \tag{3.11}$$

The proof of Theorem 2 can be found in Appendix B.3.

Then, simply subtracting away the average firing rate of neuron Y would produce the synaptic profile:

$$W(i) = \hat{P}(Y_n = 1 | X_{n-i} = 1) - \hat{P}(Y_n = 1) \tag{3.12}$$

The value and shape of $W(i)$ not only convey the sign of the synaptic action, with positive values signaling synaptic excitation and negative values synaptic inhibition; they also depict the time course of the effect of a presynaptic spike in X on the firing probability of Y, which can help classify synapses as fast vs. slow. In Fig. 3.2, the synaptic profile differentiates a fast excitatory synapse from a much slower inhibitory one, as well as provides information on the time course of the synaptic influence.

Chapter 4

Validation and Results

4.1 Sparse Poisson Spiking Model

As a first step to validate the method, a simple model of two neurons was used to examine the effect of various conditions and parameters on DI. This model is a variation of the Sparse Poisson Spiking Model [35]. When the firing pattern of a neuron is sparse, a homogeneous Poisson process with a fixed rate can be used to model its spiking activity.

Neuron X was designed to be the "master neuron." Spiking activity in X was generated by a Poisson model with a total length of T seconds and a rate of λ_X. For a fixed bin width Δ, the digitized sequence had $n = T/\Delta$ samples and on average $k = \lambda_X T$ spikes. λ_X was chosen such that $k \ll n$ and that signal X was sparse. Spikes in neuron Y were generated directly based on the spikes in X, and therefore, this synapse was excitatory. d represented the delay between a spike in X and a spike in Y, and $P(Y_i = 1|x_{i-d} = 1)$ the probability of one spike in X eliciting one spike in Y, which quantified the strength of the synapse. Some "jitter" was also introduced in Y's spikes and this temporal variation was defined by a Gaussian random variable $w \sim \mathcal{N}(0, \sigma^2)$. Some baseline level activity was added to Y, which was defined by a Poisson process with rate λ_Y in addition to the spikes induced by X. MATLAB implementation of this model can also be found in our online repository.

In actual experiments, normalized directed information $\tilde{I}(X \to Y) = \bar{I}(X \to$

$Y)/\bar{H}(Y)$ is preferred because it bounds the DI value between 0 and 1 as well as normalizes the information Y receives with respect to the level of information in itself. Normalized directed information was used in the following results.

In the first example, we examined the effect of synaptic strength on $\tilde{I}(X \to Y)$ values. In this case, $T = 600$ s, $\Delta = 0.01$ s, $\lambda_X = 0.01$, $d = 0.05$ s and $\sigma = \frac{1}{\sqrt{2}}\Delta = 0.007$. Background activity of Y was suppressed by setting $\lambda_Y \approx 0$. $P(Y_i = 1|x_{i-d} = 1)$ was varied from 0 to 1 with an increment of 0.05. As expected, DI value increased with increased synaptic strength as the baseline activity level in the postsynaptic neuron was kept constant (Fig. 4.1A). Also notice that normalized DI was plotted and the jump at 0 was caused by thresholding.

In the second example, we examined the effect of background activity level in Y on $\tilde{I}(X \to Y)$ values. λ_Y was varied from 0.001 to 10 while the synaptic strength $P(Y_i = 1|x_{i-d} = 1)$ was held constant at 0.8. Although the synaptic strength was kept constant, DI value decreased as the baseline activity level of Y increased (Fig. 4.1B), illustrating that DI is not solely determined by synaptic strength. Indeed, DI quantifies the amount of information flow from one neuron to another. DI quantifies how much the information present in neuron Y can be accounted for by the information in X. Therefore, if X accounts for only a small portion of the spikes in Y, then DI would be relatively small.

In the third example, we examined the effect of the *variation* in the time course of the synaptic response on normalized directed information. Let $P(Y_i = 1|x_{i-d} = 1) = 0.8$ and $\lambda_Y \approx 0$. σ was increased from 0 to 0.02. It makes intuitive sense that more variance in the distance between a pre- and a postsynaptic spike made the pattern more unpredictable, and hence the lower the DI value (Fig. 4.1C).

In the final example, we examined the effect of varying bin width on $\tilde{I}(X \to Y)$ in

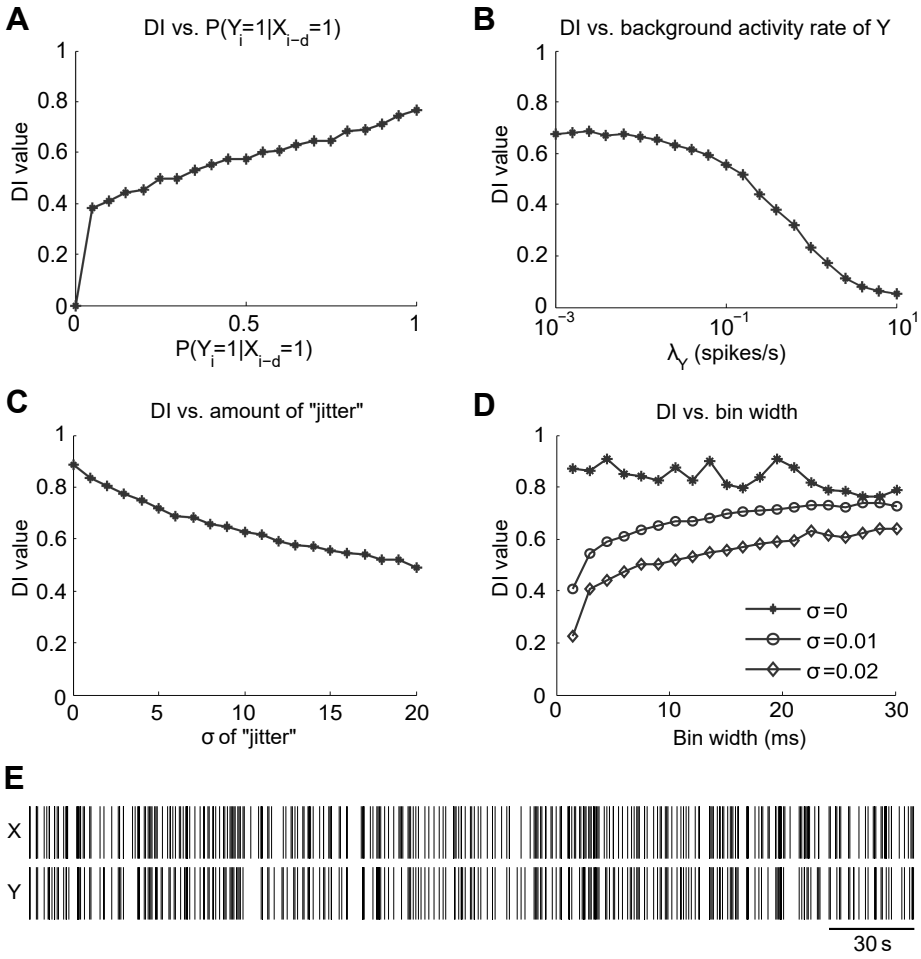


Figure 4.1 : **Trends of normalized DI tested under the sparse Poisson spiking model. Parameters such as the synaptic strength, background activity rate of the postsynaptic neuron, the amount of "jitter" for postsynaptic spikes and bin width were examined.** ***A:*** Relationship between DI and varying levels of synaptic strength. Synaptic strength was varied by changing the probability of a postsynaptic spike being elicited following a presynaptic spike. As predicted, DI value increases with a stronger synapse when the baseline activity level in Y is kept constant. ***B:*** Relationship between baseline activity level λ_Y of Y and DI. The value of DI decreases as the baseline activity in Y increases. ***C:*** Relationship between the variance of the time course of the synaptic action and DI. The value of DI is inversely related to the variance of the time course. ***D:*** The relationship between bin width and DI for different levels of "jitter". A large drop in DI can be observed for $\sigma \geq 0.01$ for small bin widths (≤ 3 ms). However, relatively small changes in DI can be seen for bin widths ≥ 10 ms. At $\sigma = 0$, DI remains high regardless of the size of the bin width. ***E:*** Sample spike trains generated by the Poisson spiking model. $\lambda_Y = 0.1$ and other parameters are the same as the model in Panel B.

the presence of noise. Bin widths between 1.5 ms and 30 ms were examined. When the synaptic delay had 0 variance, different bin widths should not have any influence on the normalized DI values, because each time a presynaptic spike occurred, CTM was sure to find a postsynaptic spike exactly d/Δ bins away. We demonstrated this by setting $P(Y_i = 1|x_{i-d} = 1) = 0.8$ and $\sigma = 0$ (Fig. 4.1D). Notice that the fluctuation in DI was caused by the artifact of binning an already discretized signal. Then, we set $\sigma = 0.01$ and 0.02. In the presence of synaptic time course variation, however, as we used smaller bin width Δ, normalized DI became smaller as well. Essentially, $(d + w)/\Delta$ landed in more different patterns as Δ decreased. This result illustrates that a small bin width is not always desirable in order to detect a slow connection, whose slower dynamics entails a greater range of variation in the time course of its synaptic response.

4.2 Simulated Neural Networks

Next, we used the realistic networks generated in the neurosimulator SNNAP (Simulator for Neural Networks and Action Potentials) to further validate our toolbox. SNNAP has the ability to simulate each neuron with a set of Hodgkin-Huxley type conductance-based equations and different types of chemical and electrical synapses with or without plasticity [36–39]. This toolbox also has the ability to introduce random noise into various components of the mathematical formulation such as the membrane leakage current and the synaptic current. SNNAP has been used to model the central pattern generator in the buccal ganglion of *Aplysia* [40, 41], and therefore it is a useful tool to check the performance of our method. It is also worth noting that, unlike the previous example which is based on a linear spiking model, SNNAP can simulate realistic neural connections and their activity. The Java-based SNNAP

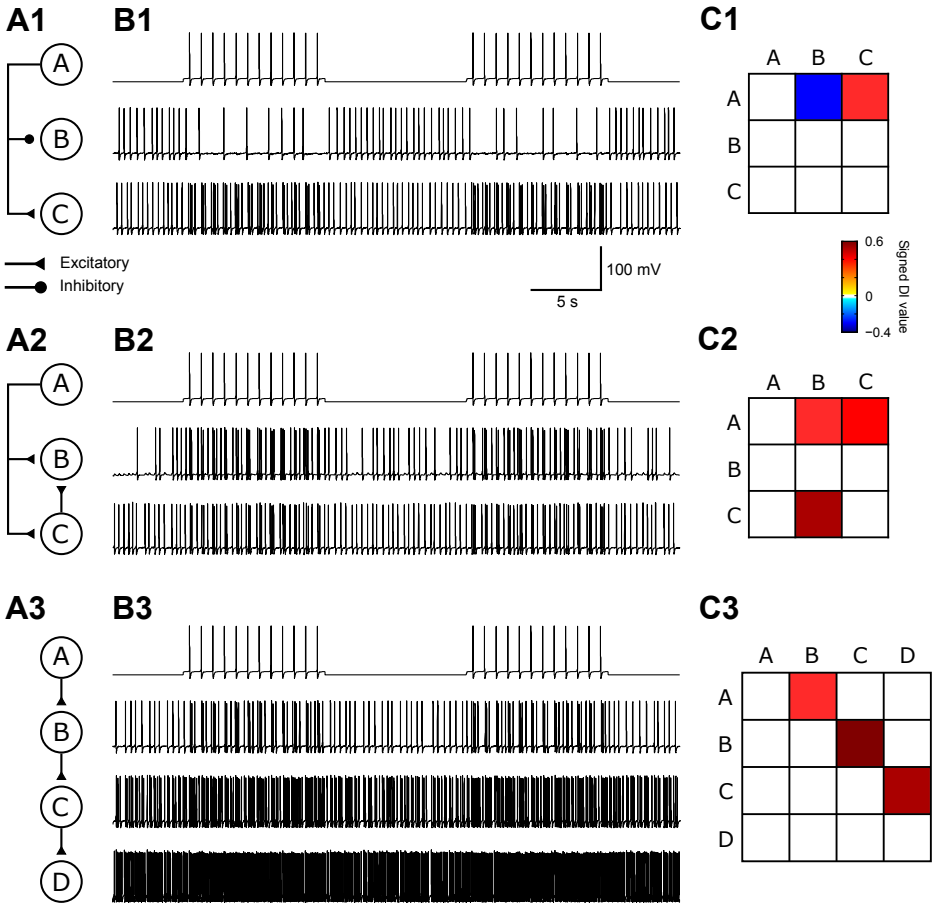


Figure 4.2 : **DI correctly inferred the connectivity in three simple networks.** ***A1:*** Neuron A inhibits neuron B and excites neuron C. ***A2:*** A network with convergence of a direct connection and a disynaptic connection, where neuron A excites neurons B directly as well as sends information through neuron C and again to neuron B. ***A3:*** A chain of feedforward excitation from neuron A to neuron D. ***B:*** Simulated membrane potential for each neuron within the corresponding network in Panel A. Each trace corresponds to the adjacent neuron in Panel A. ***C:*** DI values inferred from the spike activity. A row is a presynaptic neuron whereas a column is a postsynaptic neuron. Warm colors represent excitatory connections, whereas cool colors represent inhibitory connections. Note that DI values are always positive. The sign determined using the method introduced in Section 3.5 are attached to the DI values for ease of visualization.

software, as well as all the networks used in this section, can be found in the online repository.

We began by testing our method using three simple circuits without synaptic plasticity. For all three circuits, simulations were 200 s in duration and neuron A was activated by a depolarizing current (0.57 mA, 10 s) added at an interstimulus interval of 10 s. The connectivity matrix detected by DI was represented as a heat map. In this heat map, rows represent presynaptic cells and columns represent postsynaptic cells. Therefore, each entry of the matrix represents a connection from the cell of the corresponding row to the cell of the corresponding column. Warmer colors indicate excitatory connections, whereas cooler ones indicate inhibitory connections (Fig. 4.2C). In the network of Fig. 4.2A1, DI correctly distinguished the excitatory connection from the inhibitory connection. The network of Fig. 4.2A2 contained a convergence of a direct and an indirect path, and the method was able to identify the correct connections without incorrectly eliminating the disynaptic link. In the network of Fig. 4.2A3, the four neurons formed a feedforward chain. The method successfully predicted the appropriate connections and eliminated all indirect connections that could possibly arise from the long chain.

We next tested the effect of synaptic plasticity on DI. Three different conditions were simulated where the plasticity was manipulated for synapse B to D: no plasticity, with facilitation, and with depression. No other synaptic connections within this network had plasticity. Sample signal traces from the simulations are shown in Fig. 4.3B. Estimates of DI were made on spike trains with different time resolutions (Fig. 4.3C). This example shows that the method is able to correctly infer the network even in the presence of synaptic plasticity (Fig. 4.3D). Notice that introducing depression reduced the DI value but did not eliminate it entirely in this connection.

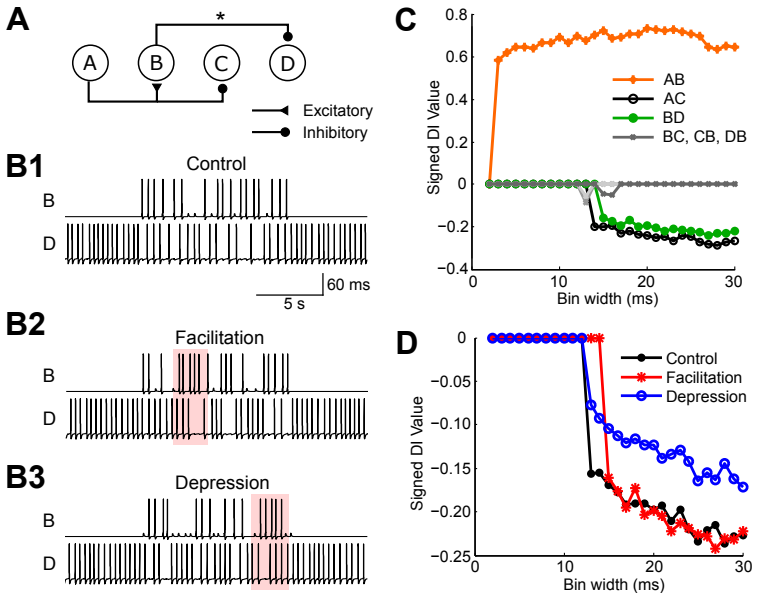


Figure 4.3 : **DI tested on a network with synaptic plasticity.** ***A:*** A neural circuit in which neuron A excites neuron B and inhibits neuron C. Neuron B inhibits neuron D. Facilitation or depression was added to synapse BD (marked with an asterisk). ***B:*** Sample traces generated by the circuit in three conditions: control (***B1***), facilitated (***B2***), and depressed (***B3***). The area of interest where the effects of plasticity can be observed is marked by pink. In Panel B2, the strength of the synapse increases where spikes come in quick succession, whereas in Panel B3, the strength of the synapse decreases where spikes closely follow each other. ***C:*** Signed DI values plotted against bin widths for condition B2 with facilitation. ***D:*** Signed DI values for synapse BD against different bin widths for all three different conditions.

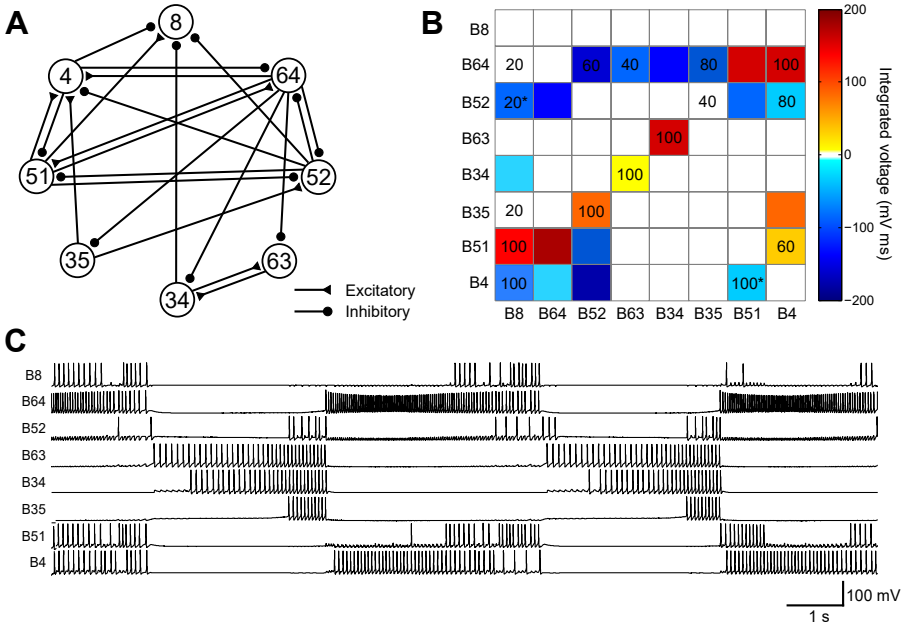


Figure 4.4 : **Testing a conductance-based model of the central pattern generator (CPG) in the buccal ganglion of *Aplysia*. *A:*** Connections of the CPG model. ***B:*** The same model network represented in a matrix format. Values represented as colors in the matrix were obtained from integrated voltages of the PSPs from the simulation. The numbers inside the matrix indicate the percentage of time a given synapse was detected by DI. A number on a colored background is a true positive. A number on a white background is a false positive. A number marked by an asterisk indicates an incorrect sign, which occurred at B52-B8 and B4-B51. Zeros values were not included ***C:*** Simulated spiking activity of the CPG.

We next tested the algorithm on a model of components of a central pattern generator (CPG) circuit of *Aplysia* [40], which simulates some of the neuronal activity underlying feeding behavior. The model was slightly modified to simulate ingestion buccal motor patterns (iBMP). The model included eight known neurons: B4, B8, B31, B34, B35, B51, B52, B63, and B64 (Fig. 4.4). This network contains excitatory and inhibitory synaptic connections, many of which exhibit facilitation or depression. Lastly, some of these neurons contain regenerative properties which elicit recurrent spike activity that outlasts the excitatory input. All of these features are present in the feeding CPG of *Aplysia*, and therefore this model provides a comprehensive test for DI. Five trials of 2 min each were generated. The algorithm correctly identified 9.2 $\pm$ 0.7 key connections (Fig. 4.4B and Table 4.1), and together with true negatives, DI correctly located or rejected 41.4 out of all 56 possible connections (diagonals excluded). The number of false positives was on average less than one synapse per trial. The number of false negatives, however, was 12.4 $\pm$ 0.2 synapses per trial. However, it is worth noting that seven of the undetected synapses were weak connections: B64-B34, B52-B64, B52-B51, B34-B8, B25-B4, B51-B52, and B4-B52. When these connections were trimmed from the simulator, the quality of the feeding pattern was not affected. This possible simplification illustrates the importance of a functional connectome, which identifies active information pathways that are in a subset of the anatomical connectome. Three essential connections, B64-B51, B51-B64 and B4-B64, went undetected throughout all five trials. The electrical coupling between B64 and B51 might have been overlooked by DI because the connection did not produce any clear spike-to-spike relationship, which is a limitation of this method. The inhibitory synapse B4-B64 was not detected by DI presumably because the B4-B64 synaptic connection did not have adequate strength to overcome the strong regenerative prop-

Table 4.1 : DI Performance on Simulated CPG Network

	True (+)		True (-)	Incorrect Sign		False (+)		False (-)	
Trial	E-E	I-I	N-N	E-I	I-E	E-N	I-N	N-E	N-I
1	6	4	32	1	0	0	1	3	9
2	5	4	31	1	0	2	0	4	9
3	5	5	32	1	0	1	0	4	8
4	6	2	33	2	0	0	0	3	10
5	6	3	33	2	0	0	0	3	9
mean	5.6	3.6	32.2	1.4	0	0.6	0.2	3.4	9
stderr	0.2	0.5	0.4	0.2	0	0.4	0.2	0.2	0.3

erties of B64. There were on average 1.4 incorrect signs, all of which were inhibitory synapses inferred to be excitatory. The B4-B51 synaptic connection was a biphasic synapse with an early excitatory and later inhibitory component. The excitatory component seemed to override the inhibitory component leading DI to infer an excitatory rather than inhibitory connection.

4.3 Mapping Connectivity of Recorded Neurons

After testing DI on the simulated CPG network, DI was applied to VSD recordings (Fig. 4.5). Neurons with fewer than 10 spikes were excluded from the analysis. Bin widths ranging from 2 ms to 30 ms were used to generate spike trains, on which DI was applied. The DI algorithm detected many putative connections, their signs (i.e., excitatory or inhibitory) and their relative strengths of influence (Fig. 4.5C). Some of the detected connections were consistent with visual inspection, such 2-1, 14-15

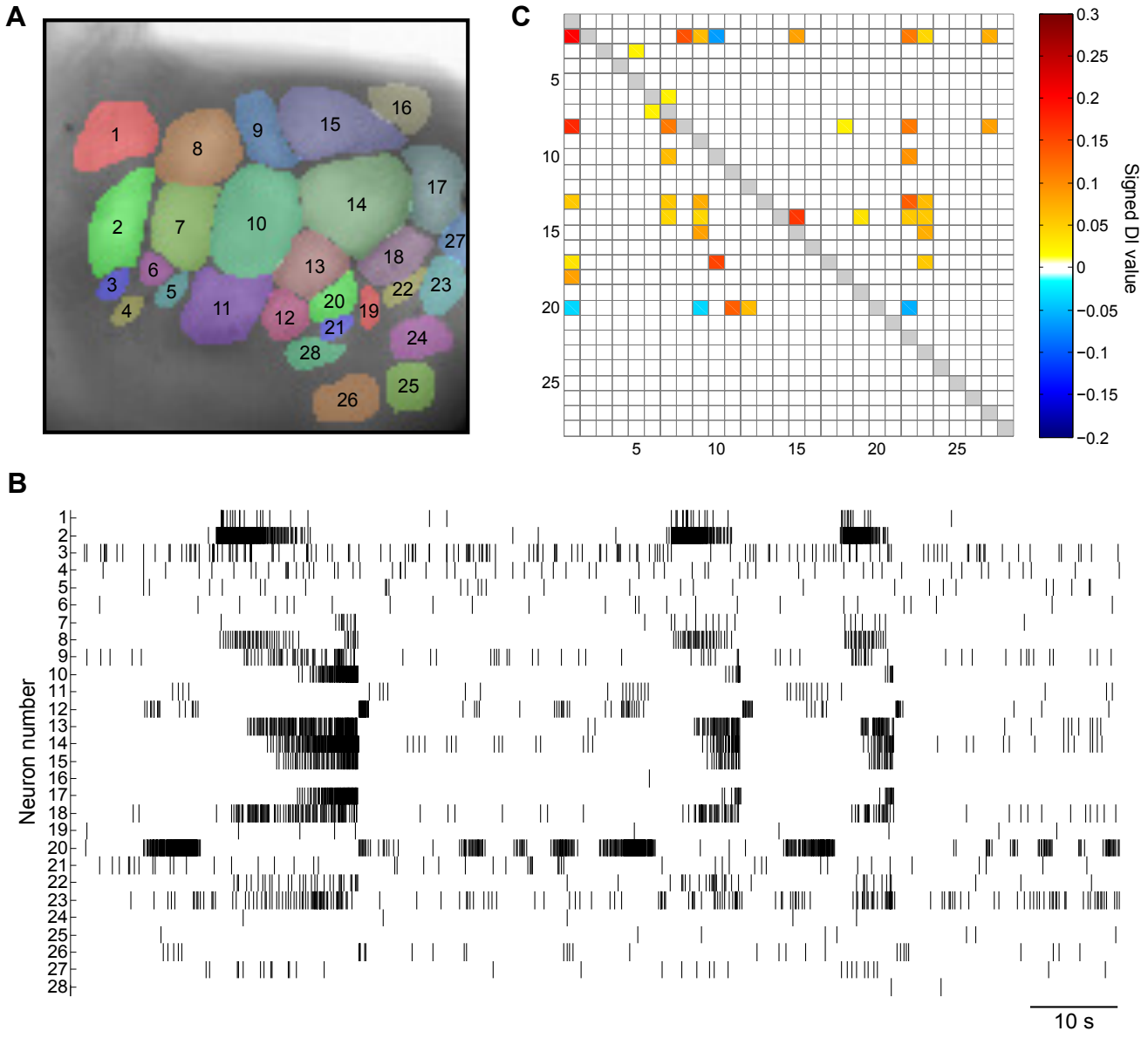


Figure 4.5 : **Analyzing VSD recording data using DI.** ***A:*** The VSD imaging surface of the caudal surface of the left buccal hemiganglion and the kernel markup of the recording surface. ***B:*** Raster plot of a 2 min VSD recording from the ganglion. ***C:*** The adjacency matrix of the network obtained from DI analysis. Many putative connections were detected.

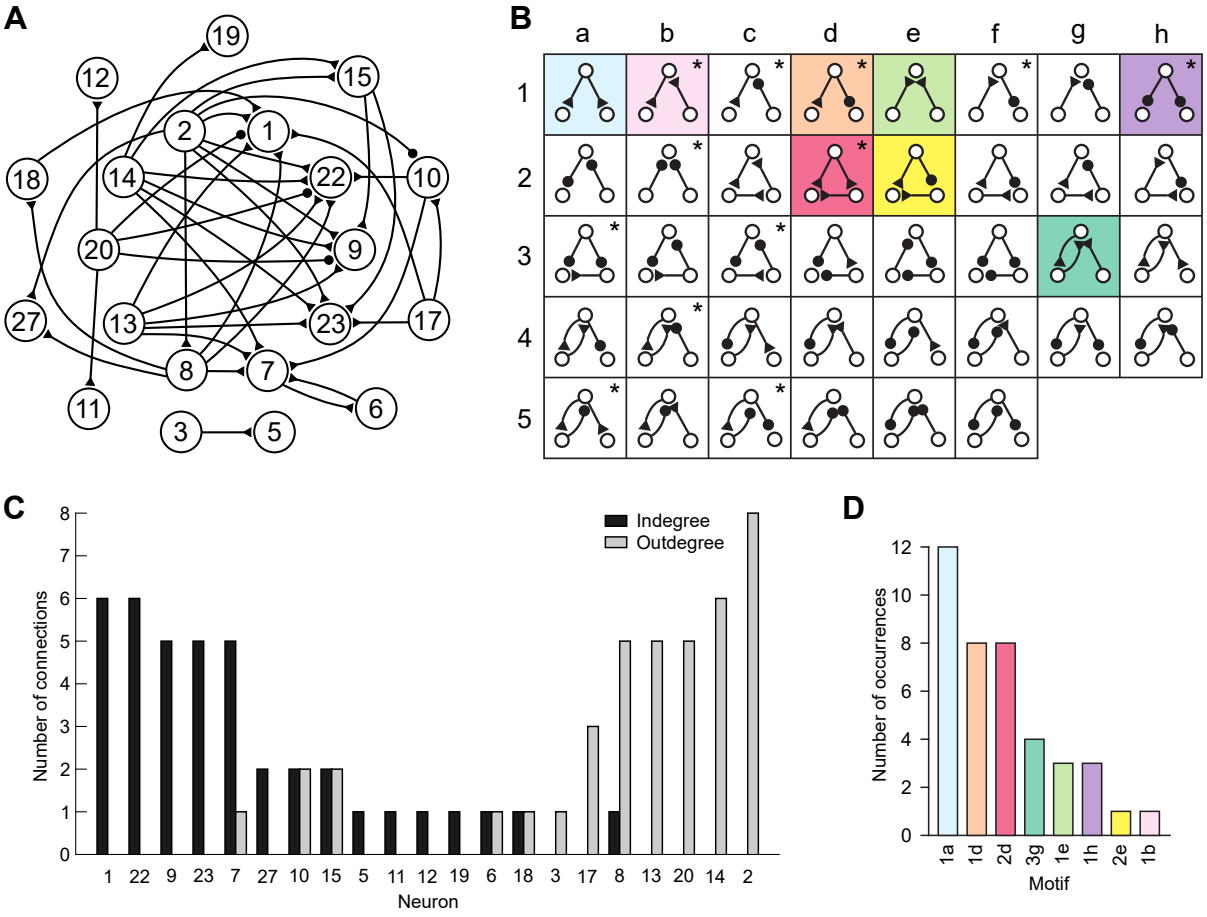


Figure 4.6 : **Patterns of connectivity of the preparation in Fig. 4.5.** ***A:*** The inferred connectivity diagram. ***B:*** All possible three-cell motifs containing 2 – 3 connections. Motifs that were present in the examples in Fig. 4.2-4.4 and were correctly identified are marked by asterisks. Motifs that were detected in the VSD recording by DI are indicated with colored backgrounds. ***C:*** Indegrees and outdegrees of neurons. Neurons without any connections are not shown. This graph shows neurons that primarily receive connections (left) and those that send out connections (right). ***D:*** Number of occurrences for each motif. The labels correspond to the indice of the motifs in Panel B. Motifs that were not detected by DI are not shown.

and 17-10, in which activity of one neuron seemed to follow the activity of the other. The algorithm also revealed some connections that would otherwise be difficult to observe, such as 2-10, 3-5 and 8-27.

We next examined to what extent neurons within the buccal ganglia could be categorized as either sources or sinks. We compared the indegree and outdegree for each neuron. In graph theory, indegree is the number of incoming connections to a node, and outdegree is the number of outgoing connections from a node [42]. Nodes with positive indegrees and 0 outdegrees are sinks, whereas those with positive outdegrees and 0 indegrees are sources. Neurons either primarily sent outgoing (e.g., 2, 8, 13, 14, 17 and 20) connections or primarily received connections (e.g., 1, 9, 22, and 23) suggesting that these may be either pre-motor or motor neurons, respectively. The degrees of connectivity of neurons will aid in identifying cells in the buccal ganglion network (Fig. 4.6C). For example, neuron 20 makes many connections and is located in a region of the ganglia where many pattern initiator neurons are found.

We next searched for microcircuit motifs within the network. Previously, motifs have been used to infer computational processes within neuronal and molecular networks [43–45]. Figure 4.6B illustrates possible motifs emerging from networks containing three neurons. It is worth noting that motifs that are highlighted by asterisks were embedded in the testing networks (Fig. 4.2-4.4) and were successfully identified by DI. We found that in the VSD recording, motif 1a, 1d, and 2d occurred most frequently. Motif 1a diverges from a single cell without any other connections. If the feeding CPG network was randomly connected then we would expect that 1a, 1b, and 1e to occur with the same frequency, but as Fig. 4.6D indicates, motif 1b and 1e were relatively infrequent. Motif 1d has both an excitatory and an inhibitory synaptic connection and occur frequently. There are several neurons within the feeding CPG

with the capability of projecting both excitatory and inhibitory synaptic connections (e.g., B4 and B71) [46, 47]. Again, we would expect motifs 1c-d and 1f-g to occur with the same frequency if the detected connections occurred by chance. However, motifs 1c, 1f and 1g were not detected. Motif 2d includes a neuron with a direct connection and a feedforward excitatory connection to a single neuron (network in Fig. 4.2A2). In a random network, we would expect 2c and 2d to occur with the same frequency however the 2c motif was not present. These results indicate that there may be a preferred pattern of connectivity for the feeding CPG. These data indicate that the method will be fruitful for analyzing the general features of the functional connectivity of neurons in the buccal ganglion.

Chapter 5

Conclusions

Our method of exploiting the context tree maximizing entropy estimator together with directed information can infer functional connectivity in small realistic simulated neural networks. The CTM based estimator has the advantage of low computational complexity, fast convergence, being non-parametric, as well as being able to mitigate overfitting. We have shown that our implementation of CTM can identify direct connections, eliminate indirect connections, reliably distinguish excitatory from inhibitory synaptic actions and quantify the amount of information flow from one neuron to another (Fig. 4.2). Furthermore, this inference technique based on DI is robust against signal nonlinearities, which linear methods such as Granger Causality or estimates based on the generalized linear model might not be able to capture. For example, it is able to detect connections with facilitation or depression (Fig. 4.3), which are common throughout invertebrate and vertebrate nervous systems.

The CTM-DI based method has its own limitations. A challenge for the DI-CTM approach is weak connections between sparse signals. Weak connections are not necessarily unessential connections in the overall network. This is one of the general disadvantages of using discrete time spike trains that other analyses using point process-GLM based GC and DI encounter as well. Bin widths that are used to segment a spiking signal into a binary spike train are small compared to the inter-event intervals of spikes. Therefore, 0's predominate the sequence. Maximum entropy is achieved when 0's and 1's are equally likely. With predominately 0's, the

entropy of the sequence is already low, and then any further drop in entropy due to conditioning on another sequence will be negligible. This problem could potentially be mitigated by dynamically setting a baseline firing rate. Another limitation of our method, as well as other methods analyzing binary spike trains, is the difficulty in detecting inhibitory synapses especially when the post synaptic cell is not active in the same phase or is completely suppressed. Such type of synapses, however, could have a significant influence on the network due to their strength. Strong inhibition and weak excitation are challenging for statistical methods because they are based on spike trains that do not reflect information on subthreshold EPSPs or IPSPs, which are all mapped to the value 0. Further developments of the DI method might include a combination of spike train analysis and analysis of the analog signals.

Despite some limitations, the CTM-DI based method has practical advantages. It naturally turns its focus onto active neurons that are generating information and playing an important role in the network. It captures the salient, active communication pathways of a neural network. The method produced promising results on the realistic *Aplysia* buccal CPG network. DI correctly identified many connections in the CPG model circuit with a relatively small false positive rate. We applied the technique to the VSD recordings of the *Aplysia* buccal ganglion and discovered some interesting putative functional connectome structures. In a single recording this method identified 40 putative synaptic connections, a feat that would be virtually impossible using pairwise intracellular electrodes. We detected several frequently occurring network motifs. However, this success should be interpreted cautiously because we only detected 40% of the connections of the CPG model by DI. In addition, there are many neurons outside of the field of view, so the actual network motifs are likely to be more complicated. It is important to note that these simple motifs will be embedded

in those more complicated motifs. The overrepresentation of specific network motifs may indicate an underlying pattern of connectivity of the feeding CPG or it may be a result of a differing level of sensitivity for DI in detecting connections within particular motifs. Separating out these two possibilities and investigating the roles of these particular motifs during behavior would be an important area of investigation for future research. It will be interesting to apply the technique to ganglia before and after different forms of learning such as operant and classical conditioning [48–51]. DI has the potential to identify distributed sites of plasticity and the ways in which the circuit is reconfigured by learning. The results of this study indicate that this method is highly versatile and correctly infers the connectivity of networks containing many different features in complex circuits. This versatility indicates that this technique can also be applied to more complex systems such as the vertebrate central nervous system.

Bibliography

[1] I. H. Stevenson and K. P. Kording, "How advances in neural recording affect data analysis," *Nature Neuroscience*, vol. 14, no. 2, pp. 139–142, 2011.

[2] D. S. Peterka, H. Takahashi, and R. Yuste, "Imaging voltage in neurons," *Neuron*, vol. 69, no. 1, pp. 9–21, 2011.

[3] A. P. Alivisatos, M. Chun, G. M. Church, R. J. Greenspan, M. L. Roukes, and R. Yuste, "The brain activity map project and the challenge of functional connectomics," *Neuron*, vol. 74, no. 6, pp. 970–974, 2012.

[4] E. N. Brown, R. E. Kass, and P. P. Mitra, "Multiple neural spike train data analysis: state-of-the-art and future challenges," *Nature Neuroscience*, vol. 7, no. 5, pp. 456–461, 2004.

[5] R. Malladi, G. P. Kalamangalam, N. Tandon, and B. Aazhang, "Inferring causal connectivity in epileptogenic zone using directed information," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 822–826, IEEE, 2015.

[6] R. Malladi, G. Kalamangalam, N. Tandon, and B. Aazhang, "Identifying seizure onset zone from the causal connectivity inferred using directed information," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 7, pp. 1267–1283, 2016.

[7] M. Dhamala, G. Rangarajan, and M. Ding, "Analyzing information flow in brain networks with nonparametric granger causality," *Neuroimage*, vol. 41, no. 2, pp. 354–362, 2008.

[8] A. J. Cadotte, T. B. DeMarse, P. He, and M. Ding, "Causal measures of structure and plasticity in simulated and living neural networks," *PLoS One*, vol. 3, no. 10, p. e3355, 2008.

[9] D. H. Perkel, G. L. Gerstein, and G. P. Moore, "Neuronal spike trains and stochastic point processes: Ii. simultaneous spike trains," *Biophysical journal*, vol. 7, no. 4, pp. 419–440, 1967.

[10] L. Nowak and J. Bullier, "Cross correlograms for neuronal spike trains: Different types of temporal correlation in neocortex, their origin and significance," *Time and the Brain*, vol. 3, pp. 53–96, 2000.

[11] W. Truccolo, U. T. Eden, M. R. Fellows, J. P. Donoghue, and E. N. Brown, "A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects," *Journal of Neurophysiology*, vol. 93, no. 2, pp. 1074–1089, 2005.

[12] S. Kim, D. Putrino, S. Ghosh, and E. N. Brown, "A granger causality measure for point process models of ensemble neural spiking activity," *PLoS Comput Biol*, vol. 7, no. 3, p. e1001110, 2011.

[13] C. J. Quinn, T. P. Coleman, N. Kiyavash, and N. G. Hatsopoulos, "Estimating the directed information to infer causal relationships in ensemble neural spike train recordings," *Journal of Computational Neuroscience*, vol. 30, no. 1, pp. 17–44, 2011.

[14] F. Gerhard, T. Kispersky, G. J. Gutierrez, E. Marder, M. Kramer, and U. Eden, "Successful reconstruction of a physiological circuit with known connectivity from spiking activity alone," *PLoS Comput Biol*, vol. 9, no. 7, p. e1003138, 2013.

[15] Z. Cai, C. L. Neveu, J. H. Byrne, and B. Aazhang, "On inferring functional connectivity with directed information in neuronal networks," in *2016 50th Asilomar Conf on Signals, Syst and Comput, Asilomar, CA*, pp. 356–360, IEEE, 2016.

[16] I. Gat, N. Tishby, and M. Abeles, "Hidden markov modelling of simultaneously recorded cells in the associative cortex of behaving monkeys," *Network: Computation in Neural Systems*, vol. 8, no. 3, pp. 297–322, 1997.

[17] A. Friedman, J. F. Slocum, D. Tyulmankov, L. G. Gibb, A. Altshuler, S. Ruangwises, Q. Shi, S. E. Toro-Arana, D. W. Beck, J. E. Sholes, and A. M. Graybiel, "Analysis of complex neural circuits with nonlinear multidimensional hidden state models," *Proceedings of the National Academy of Sciences*, p. 201606280, 2016.

[18] K. So, A. C. Koralek, K. Ganguly, M. C. Gastpar, and J. M. Carmena, "Assessing functional connectivity of neural ensembles using directed information," *Journal of Neural Engineering*, vol. 9, no. 2, p. 026004, 2012.

[19] L. Barnett and A. K. Seth, "The MVGC multivariate granger causality toolbox: a new approach to granger-causal inference," *Journal of Neuroscience Methods*, vol. 223, pp. 50–68, 2014.

[20] J. Massey, "Causality, feedback and directed information," in *Proc Int Symp Inf Theory Applic (ISITA-90), Honolulu, USA*, pp. 303–305, CiteSeer, 1990.

[21] J. Jiao, H. H. Permuter, L. Zhao, Y.-H. Kim, and T. Weissman, "Universal estimation of directed information," *IEEE Transactions on Information Theory*, vol. 59, no. 10, pp. 6220–6242, 2013.

[22] P. A. Volf and F. M. Willems, "A study of the context tree maximizing method," in *Proc 16th Benelux Symp Inf Theory, Nieuwerkerk Ijsel, Netherlands*, pp. 3–9, CiteSeer, 1995.

[23] I. Csiszár and Z. Talata, "Context tree estimation for not necessarily finite memory processes, via BIC and MDL," *IEEE Transactions on Information Theory*, vol. 52, no. 3, pp. 1007–1016, 2006.

[24] Z. Cai, J. H. Byrne, and B. Aazhang, "Inferring functional connectivity of neural circuits using information theoretic causality measures," in *Neurosci 2015 Abstracts, Chicago, IL*, Society for Neuroscience, 2015. 629.08.

[25] Z. Cai, C. L. Neveu, D. A. Baxter, J. H. Byrne, and B. Aazhang, "Inferring functional connectivity of neural circuits using information theoretic causality measures," in *Neurosci 2016 Abstracts, San Diego, CA*, Society for Neuroscience, 2016. 642.14.

[26] R. Venkataramanan and S. S. 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, pp. 2154–2179, 2007.

[27] G. Kramer, *Directed information for channels with feedback*. PhD thesis, ETH Zürich, Switzerland, 1998.

[28] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate

coding," *IEEE Transactions on Information Theory*, vol. 24, no. 5, pp. 530–536, 1978.

[29] M. Burrows and D. J. Wheeler, "A block-sorting lossless data compression algorithm," *SRC Research Report*, no. 124, 1994.

[30] J. Cleary and I. Witten, "Data compression using adaptive coding and partial string matching," *IEEE Transactions on Communications*, vol. 32, no. 4, pp. 396–402, 1984.

[31] Y. Gao, I. Kontoyiannis, and E. Bienenstock, "Estimating the entropy of binary time series: Methodology, some theory and a simulation study," *Entropy*, vol. 10, no. 2, pp. 71–99, 2008.

[32] R. E. Krichevsky and V. K. Trofimov, "The performance of universal encoding," *IEEE Transactions on Information Theory*, vol. 27, no. 2, pp. 199–207, 1981.

[33] R. E. Krichevskii, "Minimum description length prediction of next symbol," in *Information Theory. 1997. Proceedings., 1997 IEEE International Symposium on*, p. 314, Jun 1997.

[34] F. M. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "Context-tree maximizing," in *Proc 2000 Conf Inf Sci and Syst, Princeton, USA*, pp. 7–12, 2000.

[35] C. S. Cutts and S. J. Eglen, "Detecting pairwise correlations in spike trains: an objective comparison of methods and application to the study of retinal waves," *Journal of Neuroscience*, vol. 34, no. 43, pp. 14288–14303, 2014.

[36] E. Av-Ron, J. H. Byrne, and D. A. Baxter, "Teaching basic principles of neuroscience with computer simulations," *Journal of Undergraduate Neuroscience*

Education, vol. 4, no. 2, pp. A40–A52, 2006.

[37] E. Av-Ron, M. J. Byrne, J. H. Byrne, and D. A. Baxter, "SNNAP: A tool for teaching neuroscience," *Brains, Minds, and Media*, vol. 3, 2008.

[38] D. A. Baxter and J. H. Byrne, "Short-term plasticity in a computational model of the tail-withdrawal circuit in *Aplysia*," *Neurocomputing*, vol. 70, no. 10, pp. 1993–1999, 2007.

[39] I. Ziv, D. A. Baxter, and J. H. Byrne, "Simulator for neural networks and action potentials: description and application," *Journal of Neurophysiology*, vol. 71, no. 1, pp. 294–308, 1994.

[40] E. Cataldo, J. H. Byrne, and D. A. Baxter, "Computational model of a central pattern generator," in *Int Conf on Comput Methods in Syst Biol, Trento, Italy*, pp. 242–256, Springer, 2006.

[41] A. J. Susswein, I. Hurwitz, R. Thorne, J. H. Byrne, and D. A. Baxter, "Mechanisms underlying fictive feeding in *Aplysia*: Coupling between a large neuron with plateau potentials activity and a spiking neuron," *Journal of Neurophysiology*, vol. 87, no. 5, pp. 2307–2323, 2002.

[42] G. Chartrand and P. Zhang, *A first course in graph theory*. Courier Corporation, 2012.

[43] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network motifs: simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824–827, 2002.

[44] Y. Zhang, P. Smolen, D. A. Baxter, and J. H. Byrne, "Computational analyses of synergism in small molecular network motifs," *PLoS Comput Biol*, vol. 10, no. 3, p. e1003524, 2014.

[45] C. L. Rees, D. W. Wheeler, D. J. Hamilton, C. M. White, A. O. Komendantov, and G. A. Ascoli, "Graph theoretic and motif analyses of the hippocampal neuron type potential connectome," *eNeuro*, vol. 3, no. 6, pp. ENEURO–0205, 2016.

[46] D. Gardner, "Interconnections of identified multiaction interneurons in buccal ganglia of *Aplysia*," *Journal of Neurophysiology*, vol. 40, no. 2, pp. 349–361, 1977.

[47] K. Sasaki, E. C. Cropper, K. R. Weiss, and J. Jing, "Functional differentiation of a population of electrically coupled heterogeneous elements in a microcircuit," *Journal of Neuroscience*, vol. 33, no. 1, pp. 93–105, 2013.

[48] R. Nargeot, D. A. Baxter, and J. H. Byrne, "In vitro analog of operant conditioning in *Aplysia*. I. Contingent reinforcement modifies the functional dynamics of an identified neuron," *Journal of Neuroscience*, vol. 19, no. 6, pp. 2247–2260, 1999.

[49] H. A. Lechner, D. A. Baxter, and J. H. Byrne, "Classical conditioning of feeding in *Aplysia*," *Journal of Neuroscience*, vol. 20, no. 9, pp. 3369–3386, 2000.

[50] B. Brembs, "Operant reward learning in *Aplysia*," *Current Directions in Psychological Science*, vol. 12, no. 6, pp. 218–221, 2003.

[51] F. D. Lorenzetti, R. Mozzachiodi, D. A. Baxter, and J. H. Byrne, "Classical and operant conditioning differentially modify the intrinsic properties of an identified neuron," *Nature Neuroscience*, vol. 9, no. 1, pp. 17–19, 2006.

[52] E. S. Hill, S. K. Vasireddi, J. Wang, A. M. Bruno, and W. N. Frost, "Memory formation in tritonia via recruitment of variably committed neurons," *Current Biology*, vol. 25, no. 22, pp. 2879–2888, 2015.

[53] E. A. Kabotyanski, D. A. Baxter, S. J. Cushman, and J. H. Byrne, "Modulation of fictive feeding by dopamine and serotonin in *Aplysia*," *Journal of Neurophysiology*, vol. 83, no. 1, pp. 374–392, 2000.

[54] I. Csiszár, "Large-scale typicality of markov sample paths and consistency of MDL order estimators," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1616–1628, 2002.

Appendix A

VSD Recording Technique

Aplysia californica (20 – 45 g) were obtained from the University of Miami NIH National Resource for *Aplysia*. Animals were anesthetized by injection of isotonic $MgCl_2$ (0.5 ml/g). The buccal ganglia were isolated and pinned down to Sylgard-lined imaging chamber containing artificial sea water (ASW) with a high concentration of divalent ions (NaCl 330, KCl 10, $MgCl_2(6H_2O)$ 90, $MgSO_4$ 20, $CaCl_2(2H_2O)$ 30, HEPES 10) with a pH adjusted to 7.5. The ganglion was stained with the VSD, RH-155 (0.25 mg/ml, AnaSpec^{TM}), for 7 min and imaged for 120 s in normal ASW (NaCl 450, KCl 10, $MgCl_2(6H_2O)$ 30, $MgSO_4$ 20, $CaCl_2(2H_2O)$ 10, HEPES 10) with a pH adjusted to 7.5 and containing 10x dilution of RH-155 similar to [52]. The bath solution was maintained at 23°C room temperature. An Olympus BX50WI upright microscope was equipped with a 20x 0.95 NA water immersion objective. Light from a 150 watt halogen bulb was passed through a 710/40 bandpass filter (BrightLine®), a 0.8 NA Olympus condenser, through the ganglia and projected to a 128x128 CMOS camera (Redshirt Imaging^{TM}) recording at 2.5 kHz. The neurons were recorded for two minutes which was preceded by a 15 s nerve stimulation (10 Hz, 100 V, 0.5 ms) and application of 40 μM L-DOPA (Tocris^{TM}) to facilitate the induction of buccal motor programs [53]. 28 Cells were marked and signals from pixels overlaying each cell were averaged to obtain the recording of a given neuron (Fig. 4.5A). Each VSD signal was bandpass filtered in MATLAB (Butterworth, Fpass1 = 15 Hz, Fstop1 = 0.1 Hz, Fpass2 = 1 kHz, Apass = 0.1, Astop1 = 60, Astop2 = 60). Spikes were

detected in the VSD signal for each cell and converted to spike trains (Fig. 4.5B).

Appendix B

Proofs

B.1 Sequential KT Estimator

Likelihood function $P(y^n|\boldsymbol{\theta})$ is modeled as a sequence of multivariate Bernoulli variables with parameters $\boldsymbol{\theta} = (\theta_0, ..., \theta_{|A|-1})$, where $\theta_i > 0$ and $\sum_{i=0}^{M-1} \theta_i = 1$. In a realization y^n, the count for symbol $a_i \in A$ is denoted by c_i for simplicity in notation, and $\sum_{i=0}^{|A|-1} = n$. The probability of this specific string y^n being generated is

$$P(y^n|\boldsymbol{\theta}) = \prod_{i=0}^{|A|-1} (\theta_i)^{c_i}. \tag{B.1}$$

A Dirichlet $(\gamma, ..., \gamma)$ distribution is the prior, denoted by $P(\boldsymbol{\theta}|\gamma)$ where γ is the hyperparameter:

$$P(\boldsymbol{\theta}|\gamma) = \frac{\Gamma(\gamma|A|)}{\Gamma(\gamma)^{|A|}} \prod_{i=0}^{|A|-1} \theta_i^{\gamma-1} \tag{B.2}$$

where $\Gamma(\cdot)$ is the gamma function. Let us denote the estimated probability generated specifically by the KT estimator by $\hat{P}$. Then the the probability of the sequence is

$$\hat{P}(y^n) = P(y^n|\gamma) = \int_{\boldsymbol{\theta}} P(\boldsymbol{\theta}|\gamma) P(y^n|\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{B.3}$$

$$= \frac{\Gamma(\gamma|A|)}{\Gamma(\gamma)^{|A|}} \cdot \int_0^1 \int_0^{1-\theta_0} \cdots \int_0^{1-\theta_0 \cdots - \theta_{|A|-2}} \prod_{i=0}^{|A|-2} \theta_i^{c_i+\gamma-1} (1 - \theta_0 \cdots - \theta_{|A|-2})^{c_{|A|-1}+\gamma-1} d\theta_{|A|-2} \cdots d\theta_0 \tag{B.4}$$

$$= \frac{\Gamma(\gamma|A|)}{\Gamma(\gamma)^{|A|}} \frac{\prod_{i=0}^{|A|-1} \Gamma(c_i + \gamma)}{\Gamma(n + \gamma|A|)} \tag{B.5}$$

We express $\theta_{|A|-1} = 1 - \theta_0 \cdots - \theta_{|A|-2}$. $\boldsymbol{\theta}$ here is a $|A|$-dimensional simplex. The integral in (B.4) is the multivariate beta integral. When $\gamma = 1/2$, this estimated probability evaluates to

$$\hat{P}(y^n) = \frac{\Gamma(\frac{|A|}{2}) \prod_{i=0}^{|A|-1} \Gamma(c_i + \frac{1}{2})}{\pi^{\frac{|A|}{2}} \Gamma(n + \frac{|A|}{2})} \tag{B.6}$$

Using the properties of the gamma function $\Gamma(x) = (x-1)\Gamma(x-1)$ as well as $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, it is easily shown that

$$\hat{P}(y^n) = \frac{\Gamma(\frac{|A|}{2}) \prod_{i=0}^{|A|-1} (c_i - 1 + \frac{1}{2})(c_i - 2 + \frac{1}{2}) \cdots (\frac{1}{2})\Gamma(\frac{1}{2})}{\pi^{\frac{|A|}{2}} (n - 1 + \frac{|A|}{2})(n - 2 + \frac{|A|}{2}) \cdots (\frac{|A|}{2})\Gamma(\frac{|A|}{2})} \tag{B.7}$$

which is exactly (3.3). Equation (B.6) also shows that $\hat{P}(\phi) = 1$ (in this case n and c_i are 0).

B.2 Incorporating Penalties into Recursive Model Finding

For the objective function defined by (3.7) in terms of code length with total number of nodes as a trade-off

$$\hat{\mathcal{T}}(y^n) = \arg\min_{\mathcal{T} \in \mathcal{I}} - \log \hat{P}_{\mathcal{T}}(y^n) + (|\mathcal{S}| + |u : u \prec s|) \log |A| \tag{B.8}$$

Take $\exp\{-(\cdot)\}$ on both sides and we have,

$$\exp\{-\hat{\mathcal{T}}(y^n)\} = \hat{P}_{\mathcal{T}}(y^n) \exp\{-|\mathcal{S}| \log |A| - |u : u \prec s| \log |A|\} \tag{B.9}$$

$$= (\frac{1}{|A|})^{|u:u \prec s|} \cdot \prod_{s \in |\mathcal{S}|} \hat{P}(y^n|s) \prod_{s \in |\mathcal{S}|} |A|^{-|\mathcal{S}|} \tag{B.10}$$

$$= (\frac{1}{|A|})^{|u:u \prec s|} \cdot \prod_{s \in |\mathcal{S}|} \frac{1}{|A|} \hat{P}(y^n|s) \tag{B.11}$$

Therefore, it is clearly shown that a penalty factor of $\frac{1}{|A|}$ is applied to the estimate of each leaf. The term $(\frac{1}{|A|})^{|u:u \prec s|}$ corresponds to the penalty generated by all the

inner nodes. While this term cannot be factored into individual contexts, it can be understood this way: if instead of the parent node, the child nodes are kept as contexts, the parent node still needs to be kept and hence 1 more node in addition to the child nodes is added to the structure. Therefore, a factor of $\frac{1}{|A|}$ is added to the product term in (3.8).

B.3 Property of the Profile Estimator

Proof of Theorem 2: Let Z be a new random variable with an alphabet of size 4 obtained by $Z = 2X+Y$, then $P(Z_n|Z^{n-1}) = P(Z_n|Z_{n-D_0}^{n-1})$, where $D_0 = \max\{D_X, D_Y\}$. Let $P(Z_n|Z^{n-1})$ denote the true conditional probability of Z. Denote the KT estimate of the conditional probability by $\hat{P}_{KT}(Z_n|Z^{n-1})$ and the ML estimate by $\hat{P}_{ML}(Z_n|Z^{n-1})$. We know that $\forall k \in \{1, ..., D_0\}$

$$\hat{P}(Y_n = 1|X_{n-i}) = \sum_{\substack{\forall x_{n-k},\, k\neq i \\ \forall y_{n-k}}} \hat{P}_{KT}(\underbrace{Y_n = 1|X_{n-D_0}^{n-1}, Y_{n-D_0}^{n-1}}_{A}) \times \hat{P}(\underbrace{X_{n-D_0}^{n-1} \setminus X_{n-i}, Y_{n-D_0}^{n-1}|X_{n-i}}_{B}) \tag{B.12}$$

First notice that $\hat{P}_{KT}(Y_n = 1|X_{n-D_0}^{n-1}, Y_{n-D_0}^{n-1})$ is obtained by taking the marginal of $\hat{P}_{KT}(Z_n|Z_{n-D_0}^{n-1})$, where for each $a \in Z$

$$\hat{P}_{KT}(Z_n = a|Z_{n-D_0}^{n-1}) - \hat{P}_{ML}(Z_n = a|Z_{n-D_0}^{n-1}) \tag{B.13}$$

$$= \frac{c(a) + \frac{1}{2}}{n + \frac{|A|}{2}} - \frac{c(a)}{n} \tag{B.14}$$

$$= \frac{\frac{1}{2}(n - c(a) \cdot |A|)}{n^2 + \frac{|A|}{2}n} \tag{B.15}$$

As $n \to \infty$, (B.15) $\to 0$. Since MLE is asymptotically consistent, we have

$$\hat{P}_{KT}(Z_n = a|Z_{n-D_0}^{n-1}) - P(Z_n = a|Z_{n-D_0}^{n-1}) = \epsilon_1 \xrightarrow{n\to\infty} 0 \tag{B.16}$$

The second term for (definition equation) is calculated by

$$\hat{P}(X_{n-D_0}^{n-1} \setminus X_{n-i}, Y_{n-D_0}^{n-1} | X_{n-i}) = \frac{c(X_{n-D_0}^{n-1}, Y_{n-D_0}^{n-1})}{c(X_{n-i})} \tag{B.17}$$

where $c(\cdot)$ is the count function. If Z^n is Markovian, $Z^{n-1} \cap \{X_{n-i} = 1\}$ is also Markovian. We then can use the typicality theorem for stationary, irreducible Markov chains [54]. For a fixed order D_0, the empirical frequency of a length D_0 string tends to its true probability:

$$\left| \frac{\hat{P}(Z_{n-D_0}^{n-1} \setminus X_{n-i} | X_{n-i})}{P(Z_{n-D_0}^{n-1} \setminus X_{n-i} | X_{n-i})} \right| < C\sqrt{\frac{\log\log n}{n}} \tag{B.18}$$

$$1 - C\sqrt{\frac{\log\log n}{n}} < \frac{\hat{P}(Z_{n-D_0}^{n-1} \setminus X_{n-i} | X_{n-i})}{P(Z_{n-D_0}^{n-1} \setminus X_{n-i} | X_{n-i})} < 1 + C\sqrt{\frac{\log\log n}{n}} \tag{B.19}$$

$$\begin{aligned} &\left| \hat{P}(Z_{n-D_0}^{n-1} \setminus X_{n-i} | X_{n-i}) - P(Z_{n-D_0}^{n-1} \setminus X_{n-i} | X_{n-i}) \right| \\ &\qquad < C \cdot P(Z_{n-D_0}^{n-1} \setminus X_{n-i} | X_{n-i}) \sqrt{\frac{\log\log n}{n}} \end{aligned} \tag{B.20}$$

Since $P(Z_{n-D_0}^{n-1} \setminus X_{n-i} | X_{n-i})$ is bounded by 1,

$$\left| \hat{P}(Z_{n-D_0}^{n-1} \setminus X_{n-i} | X_{n-i}) - P(Z_{n-D_0}^{n-1} \setminus X_{n-i} | X_{n-i}) \right| = \epsilon_2 \xrightarrow{n\to\infty} 0 \tag{B.21}$$

Putting everything together we have

$$\hat{P}(Y_n = 1 | X_{n-i}) = \sum_{\substack{\forall x_{n-k},\, k \neq i \\ \forall y_{n-k}}} \hat{P}_{KT}(A) \times \hat{P}(B) \tag{B.22}$$

$$= \sum_{\substack{\forall x_{n-k},\, k \neq i \\ \forall y_{n-k}}} [P(A) + \epsilon_1]\,[P(B) + \epsilon_2] \tag{B.23}$$

$$= \sum_{\substack{\forall x_{n-k},\, k \neq i \\ \forall y_{n-k}}} [P(A)P(B) + \epsilon_1 P(B) + \epsilon_2 P(A) + \epsilon_1\epsilon_2] \tag{B.24}$$

$$= P(Y_n = 1 | X_{n-i}) + \sum_{\substack{\forall x_{n-k},\, k \neq i \\ \forall y_{n-k}}} [\epsilon_1 P(B) + \epsilon_2 P(A) + \epsilon_1\epsilon_2] \tag{B.25}$$

And then

$$\begin{aligned}\hat{P}(Y_n = 1|X_{n-i}) - P(Y_n = 1|X_{n-i}) &= \sum_{\substack{\forall x_{n-k},\, k\neq i \\ \forall y_{n-k}}} [\epsilon_1 P(B) + \epsilon_2 P(A) + \epsilon_1\epsilon_2] \\ &\leq 2^{(2D_0-1)}(\epsilon_1 + \epsilon_2 + \epsilon_1\epsilon_2) \qquad \text{(B.26)}\end{aligned}$$

In most neural science applications, D_0 is finite and does not grow with n; therefore $2^{(2D_0-1)} < \infty$. Because $\epsilon_1 P(B)$, $\epsilon_2 P(A)$ and $\epsilon_1\epsilon_2$ tend to 0 as $n \to \infty$, the statement is proved. □

B.4 Eliminating Indirect Connections

Proof of Theorem 1: The first inequality can be easily shown, because we have W^n as a common receptor.

$$\begin{aligned}I(U^n \to W^n) &= H(W^n) - H(W^n||U^n) \qquad \text{(B.27)} \\ &= H(W^n) - \sum_{i=1}^{n} H(W_i|W^{i-1}, U^i) \\ &= H(W^n) - \left[\sum_{i=1}^{n} H(W_i|W^{i-1}, U^i, V^i) + \sum_{i=1}^{n} I(W_i; V^i|W^{i-1}, U^i)\right] \\ &= H(W^n) - H(W^n||U^n, V^n) - \sum_{i=1}^{n} I(W_i; V^i|W^{i-1}, U^i) \qquad \text{(B.28)}\end{aligned}$$

We can expand $I(V^n \to W^n)$ similarly:

$$I(V^n \to W^n) = H(W^n) - H(W^n||U^n, V^n) - \sum_{i=1}^{n} I(W_i; U^i|W^{i-1}, V^i) \qquad \text{(B.29)}$$

Then

$$I(V^n \to W^n) - I(U^n \to W^n) = \sum_{i=1}^{n} \left[I(W_i; V^i | W^{i-1}, U^i) - \underbrace{I(U^i; W_i | W^{i-1}, V^i)}_{=0} \right] \tag{B.30}$$

$$= \sum_{i=1}^{n} I(W_i; V^i | W^{i-1}, U^i) \tag{B.31}$$

$$\geq 0 \tag{B.32}$$

The term $I(W_i; U^i | W^{i-1}, V^i) = 0$ because $I(U^n \to U^n || V^n) = 0$.

To show the second inequality, we express DI as a cumulative sum of mutual information:

$$I(U^n \to V^n) = \sum_{i=1}^{n} I(U^i; V_i | V^{i-1}) \tag{B.33}$$

$$= \underbrace{\sum_{i=1}^{n} I(U^i; V^i)}_{A} - \underbrace{\sum_{i=1}^{n} I(U^i; V^{i-1})}_{B} \tag{B.34}$$

(B.35)

and

$$I(U^n \to W^n) = \underbrace{\sum_{i=1}^{n} I(U^i; W^i)}_{C} - \underbrace{\sum_{i=1}^{n} I(U^i; W^{i-1})}_{D} \tag{B.36}$$

(B.37)

$$A = \sum_{i=1}^{n} I(U^i; V^i) = \sum_{i=1}^{n} I(U^i; W^i, V^i) - \sum_{i=1}^{n} I(U^i; W^i | V^i) \tag{B.38}$$

$$= \sum_{i=1}^{n} I(U^i; W^i, V^i) - \sum_{i=1}^{n} \left[H(U^i | V^i) - H(U^i | V^i, W^i) \right]$$

$$= \sum_{i=1}^{n} I(U^i; W^i, V^i) - \sum_{i=1}^{n} \left[H(U^i | V^i) - H(U^i | V^i) \right]$$

$$= \sum_{i=1}^{n} I(U^i; W^i, V^i) \tag{B.39}$$

Because V^i contains all the information in W^i, $H(U^i|V^i,W^i) = H(U^i|V^i)$. On the other hand,

$$C = \sum_{i=1}^{n} I(U^i;W^i) = \sum_{i=1}^{n} I(U^i;V^i,W^i) - \sum_{i=1}^{n} I(U^i;V^i|W^i) \tag{B.40}$$

Therefore,

$$A - C = \sum_{i=1}^{n} I(U^i;V^i) - \sum_{i=1}^{n} I(U^i;W^i) \tag{B.41}$$

$$= \sum_{i=1}^{n} I(U^i;V^i|W^i) \tag{B.42}$$

Now, let's look the the remaining terms:

$$B = \sum_{i=1}^{n} I(U^i;V^{i-1}) = \sum_{i=1}^{n} I(U^{i-1},U_i;V^{i-1}) \tag{B.43}$$

$$= \underbrace{\sum_{i=1}^{n} I(V^{i-1};U_i|U^{i-1})}_{=0} + \sum_{i=1}^{n} I(V^{i-1};U^{i-1}) = \sum_{i=1}^{n} I(U^{i-1},V^{i-1}) \tag{B.44}$$

The reverse directed information $\sum_{i=1}^{n} I(V^{i-1};U_i|U^{i-1}) = I(V^{n-1} \to U^n) = 0$ per problem statement. Similarly,

$$D = \sum_{i=1}^{n} I(U^i;W^{i-1}) = \sum_{i=1}^{n} I(U^{i-1},W^{i-1}) \tag{B.45}$$

From equation (B.42) we know that

$$B - D = \sum_{i=1}^{n} I(U^{i-1};V^{i-1}) - \sum_{i=1}^{n} I(U^{i-1};W^{i-1}) \tag{B.46}$$

$$= \sum_{i=1}^{n} I(U^{i-1};V^{i-1}|W^{i-1}) \tag{B.47}$$

Then,

$$
\begin{aligned}
& I(V^n \to W^n) - I(U^n \to W^n) && \text{(B.48)}\\
= \;& A - C - (B - D) && \text{(B.49)}\\
= \;& \sum_{i=1}^{n} I(U^i; V^i | W^i) - \sum_{i=1}^{n} I(U^{i-1}; V^{i-1} | W^{i-1}) && \text{(B.50)}\\
= \;& \sum_{i=1}^{n} \left[I(U^i; V^i, W^i) - I(U^i; W^i) - I(U^{i-1}; V^{i-1}, W^{i-1}) + I(U^{i-1}; W^{i-1}) \right] \\
= \;& \sum_{i=1}^{n} [I(U^i; V_i, W_i | V^{i-1}, W^{i-1}) + I(U^i; V^{i-1}, W^{i-1}) - I(U^i; W_i | W^{i-1}) \\
& - I(U^i; W^{i-1}) - I(U^{i-1}; V^{i-1}, W^{i-1}) + I(U^{i-1}; W^{i-1})] && \text{(B.51)}\\
= \;& \sum_{i=1}^{n} I(U^i; V_i, W_i | V^{i-1}, W^{i-1}) - \sum_{i=1}^{n} I(U^i; W_i | W^{i-1}) && \text{(B.52)}\\
= \;& I(U^n \to V^n, W^n) - I(U^n \to W^n) && \text{(B.53)}\\
\geq \;& 0 && \text{(B.54)}
\end{aligned}
$$

Equation (B.52) follows from (B.51) because of (B.44). Therefore, $I(V^n \to W^n) \geq I(U^n \to W^n)$ and $I(U^n \to V^n) \geq I(U^n \to W^n)$. $\square$