AUTHORIZATION TO LEND AND REPRODUCE THE THESIS

As the sole author of this thesis, I authorize Brown University to lend it to other institutions or individuals for the purpose of scholarly research.

Date____________ ____

Andreas Hoenselaar, Author

I further authorize Brown University to reproduce this thesis by photocopying or other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Date____________ ____

Andreas Hoenselaar, Author

Mutual Information as a
Measure of Relevance in Neural Coding

By

Andreas Hoenselaar

Thesis

Submitted in partial fulfillment of the requirements for the Degree of Master of Science in the Department of Computer Science at Brown University

PROVIDENCE, RHODE ISLAND

MAY 2007

This thesis by Andreas Hoenselaar is accepted in its present form
by the Department of Computer Science as satisfying the
thesis requirements for the degree of Master of Science

Date_____________ ____
Gregory Shakhnarovich, PhD, Advisor

Approved by the Graduate Council

Date_____________ ____
Sheila Bonde, PhD, Dean of the Graduate School

Abstract

This work is concerned with the applicability of mutual information as a measure of relevance in neural coding. We give a detailed overview of common approaches to the application of information-theoretic measures in neuroscience. In order to create an environment that allows for the analysis of mutual information, we generate synthetic spike trains based on stimulus models that are easy to describe but sufficiently powerful to be relevant in practice.

Different methods to extract a random variable suitable for the estimation of mutual information from spike trains are discussed and their shortcomings illuminated. We go on and demonstrate the sensitivity of mutual information to an inappropriate choice of parameters in the extraction of the random variable and show how the temporal characteristics of the stimulus have to be factored in. Finally, we illustrate the detrimental effects on certain estimators of mutual information that occur if insufficient amounts of data are available.

Acknowledgments

I would like to thank Gregory Shakhnarovich for the time and patience he invested into this project and his much needed guidance on all aspects of this thesis.

Thank you, Andreas Tolias, for the igniting spark that aroused my interest in the field of computational neuroscience.

Thanks to Jadrian J Miles and all my friends who supported me in a stressful but yet exciting year.

I am deeply indebted to the Federal State of Bavaria, the German National Merit Foundation and the German Academic Exchange Service for their financial support.

Finally, I thank my parents for their unquestioned and ongoing support.

> *"Have no fear of perfection — you'll never reach it."*
>
> S. Dalí

Contents

List of Figures

1 Introduction

1.1 Motivation

Ever since the technology for electrophysiological single-cell and population recordings has become available, two major questions have been the center of attention in the research community. The first one is related to the problem what qualities of a given stimulus are encoded by neurons in a certain brain area or population. An answer to this question does not explain how these stimulus qualities are encoded by action potentials and this issue constitutes the second big problem.

To give an example for this distinction: There is significant evidence that neurons in motor cortex encode intended movement direction. Separate from this observation, the current opinion seems to tend towards a rate code (vs. temporal code). Unfortunately, it is extremely hard to devise an experimental setup which allows for a clear-cut separation of the encoded stimulus quality and the coding method. It seems as if certain assumptions have to be made about the latter aspect before the first one can be illuminated.

To circumvent this problem, we try to develop a concept of "relevance", i.e., the presence or absence of a functional relationship between a stimulus and a neural signal with a minimum of assumptions about the encoded stimulus quality or the neural code. For this purpose, we generate synthetic data from a simplified neural model that allows us to define the relationship between stimulus and neural signal and impose restrictions as necessary.

This controlled environment provides the means for an evaluation of various methods and detailed analysis of their ability to discover functional relationships. In this thesis, we propose mutual information as a measure that can fulfill the requirements stated above and that we are going to evaluate in a variety of settings and highlight problems faced by the experimenter who wants to use it.

Information theory provides a set of tools for the analysis of neural signals that has gained wide acclaim and popularity in neuroscience. Based on the intuitive idea that the brain transmits and processes information, it is intuitive to make use of a concept that was developed for the more technical setting of information transmission on analog and digital channels. Fortunately, the notion of information that we are going to use is sufficiently general and can be reduced to the idea that information transmission equals reduction of uncertainty.

1.2 Goals of this work

In this thesis, we are going to discuss approaches to the use of mutual information in the setting of neuroscientific experiments. Information theory provides a broad set of methods and the success of its application depends heavily on an adequate data representation. The major contribution of this work is a careful assessment of all processing steps in the controlled environment provided by the availability of synthetic data. Our results demonstrate how sensitive mutual information is to certain parameters that determine the transformation of neural spike trains into discrete random variables, e.g., the window size used for the estimation of instantaneous firing rates in particular.

We diverge from the path taken in previous studies that assume an extraordinarily controlled and homogeneous stimulus presentation protocol, but rather aim for the greatest generality and wide most applicability. In particular, this means that we use dynamic stimuli with probabilistic presentation intervals and without any intertrial intervals. The complications posed by this paradigm are severe but are exactly the ones faced by researchers in the field of brain computer interfaces. Even simple stimuli, such as the binary stimulus we are going to introduce later, transpire as interesting and insightful to study. We extend this model and consider a more general class of stimuli drawn from a discrete set. Our results shed light on the relationship of the temporal characteristics of the stimulus and optimal parameters to use for the extraction of the random variable from the spike train.

Finally, we analyze the amount of data required to obtain reliable estimates of mutual information. It is well known that insufficient amounts of data severely distort estimates of information-theoretic quantities. For this reason, we explain the mechanism that causes the upward bias in estimates and empirically study the effect on synthetic datasets.

2 Background

2.1 Role of the motor cortex for motion control and planning

The brain areas most heavily involved with motor control are Brodmann areas 4 and 6. Area 4, located between the frontal and prefrontal sulcus, is referred to as primary motor cortex (M1). Anterior to M1 sits area 6 that encompasses premotor cortex and the supplementary motor area (SMA) [16]. Mapping the functional organization of primary motor cortex goes back to the 1950s, when Penfield and Rasmussen systematically studied the somatotopic representation of the human body [38]. Their findings and those of similar studies led to the well-known homunculus of human somatotopy that describes how limbs are laid out on the motor cortex. More recent studies corroborated somatotopy in M1 for major body divisions, such as the arm, face, or leg but indicate that the homunculus model is an oversimplification on the smaller scale of individual digits, for example [42]. But nonetheless, somatotopic organization facilitates the development of neural prosthetics

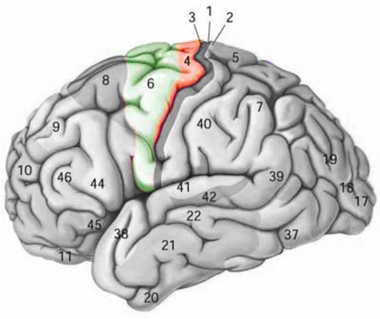


Figure 1: Brodmann area 4 (M1), located anterior to the central sulcus, is highlighted in red. Shown in green are premotor cortex and SMA, anatomically Brodmann area 6.

because the approximate position of relevant areas is known a priori.

The motor cortex is active in a variety of tasks, voluntary limb movement being the most obvious one. If this was the only scenario that elicits M1 activity, efforts to develop arm and hand prostheses were thwarted immediately. Fortunately, a series of neuroimaging studies in the mid- and late 1990s revealed that imagined movements cause significant activation of primary motor cortex, premotor cortex and SMA. One hypothesis relates activity during imagined movements to the similarity of imagining movements and planning movements [29]. A more detailed analysis of the variables encoded in M1 follows in the next section.

2.2 Coded variables in primary motor cortex

After the importance of Brodmann areas 4 and 6 for skeletomotor control had been discovered, diverging theories about the control variables encoded in motor cortex evolved. Electrophysiological recordings from single neurons in awake and behaving primates in a particular experimental setup — the center-out task — have revealed that certain directional properties are dominant in primary motor cortex. In the center-out task, a monkey grasps a manipulandum and moves it from the center position to one of eight equidistant target locations with an angular distance of 45° [17]. It was shown that responses of the bigger part of M1 neurons vary as a function of target location and thus direction of the intended movement. Consequently, the notion of directional tuning, similar to orientation tuning in visual cortex and other sensory areas, was introduced. It can be observed that neurons typically respond maximally to a certain movement direction and to a lesser degree if the movement direction diverts from the preferred one.

In order to quantify directional tuning, an experimentalist will elicit arm movements or wait for voluntary movements and record neural activity to determine whether a neuron is involved in the control of the limb under scrutiny. After this basic relationship has been established, neural recordings from a number of trials for each of the eight target locations are subjected to a statistical test (e.g., F-test) to determine whether there is a significant relationship between movement direction and neural signal. In many cases, neurons which exhibit directional tuning can also be characterized by their preferred direction which elicits the maximum response [18]. A tuning curve can then be fitted to the data via regression to interpolate neural response properties over a continuous range of directions. Sinusoidal curves, which have a small number of free parameters, are commonly used and provide a good fit of the data [17, 43].

The observation that neurons in M1 are broadly tuned in general implies that information about movement direction might be gained by looking at cell populations instead of single cells [18]. Support for this hypothesis is provided by the significant variation in unit responses from trial to trial under the same stimulus conditions [28]. Precluding more complex interactions between neurons of a population, averaging the response of several units is a viable mechanism to reduce noise at the very least. The amount of information that can be gained if an incremental number of neurons is taken into account, depends on the correlation structure and is still under heavy discussion. Hopefully, a rigorous application of the techniques in Section 3 will help to resolve the controversy.

Many different strategies for the decoding of population responses can be imagined. Georgopoulos et al. propose to model cell activity as the sum of a baseline firing rate and the cosine between preferred direction and movement direction of a particular instance. A decoding scheme to infer the movement direction from the population activity is devised. It is postulated that a vectorial linear combination of preferred directions, weighted by each neuron's response relative to baseline, closely approximates movement direction [18].

The directionally tuned neural response in M1 populations can be observed tens of milliseconds before the actual movement onset [31]. Georgopoulos et al. demonstrate that arm movement lags approximately 160ms behind the neural signal and that three-dimensional arm trajectories can be decoded from the neural response at sufficient precision [44]. It is only logical to extend the idea of tuning to the spatiotemporal domain, i.e., evaluate to what extent a neuron's activity correlates with some directional hand path parameter, e.g., velocity in a certain direction. Paninski et al. devised the spatiotemporal tuning curve as a way to represent this property in a compact form [36].

Given the complexity of the motor control problem, it is hard to imagine that a population of neurons which are solely tuned for the intended movement direction can constitute a neural substrate that is apt for the task. Furthermore, the limited motion range and stereotypedness of the center-out task — which served as the primary investigative tool — are good reasons not to endorse this point of view. And indeed, many studies have emphasized that other variables like limb position, velocity, force and acceleration are encoded by M1 neurons.

Those neurons which are tuned for movement direction will also exhibit a tuning for hand velocity direction in the center-out task, as all movements deviate only slightly from straight lines. For the same reasons, it is hard to devise an experimental setup which allows for a clear separation of limb velocity from limb position (cf. [36]). However, given that muscle spindles provide information about limb position, the hypothesis that position is the variable encoded in M1 is not a far-fetched one. Wise and Tanji found that about half of the neurons they studied in M1 exhibit an activity level that reflects static foot position. Limbs were also displaced by external force application in the form of ramps and it was discovered that neurons in caudal and rostral M1 respond to ramp displacement and static displacement [58]. It follows that not only static limb posture is represented in M1 but also the dynamic properties of this variable. Still, there are drawbacks when limb position is used as the primary coded variable for control purposes. External forces may deflect the hand or foot trajectory from the intended one. Before this control error is reflected by the limb's position, acceleration has to be integrated twice over a certain time window. The ensuing delay could decrease performance and introduce instabilities into the control system.

By contrasting neural response properties in the center-out task with those in an isometric task, Sergio and Kalaska strengthened the idea of directional tuning in M1 neurons, but they also found a pronounced relationship between hand force and neural activity [46]. Coding hand force in the motor cortex is a suggestive idea because the effectors in motor control are muscles. It follows that one should expect to find neurons which encode the force exerted by a particular muscle in motor cortex. The latter represents a coding scheme based on a more intrinsic view of the motor system, while coding of hand path and directional velocity are extrinsic parameters. No clear evidence in favor of either of the two paradigms was found in a study by Kakei. On the contrary, the authors claim that two distinct populations appear to exist in the primary motor cortex that separately encode variables that live in the intrinsic and extrinsic coordinate frames [23]. Such a clear-cut discrimination is not supported by the findings in studies which analyzed the effect of hand and arm orientation on cell discharge in M1 [45]. As microstimulation of neurons in both populations was able to elicit muscle contractions at comparable thresholds and the time delay between neural signal and movement onset was identical in both groups of neurons, the hypothesis that these neurons represent different stages in the control hierarchy has to be rejected [23]. While force is certainly a good representation to allow for fast reaction to external disturbances, it can be speculated that it may adopt badly to varying conditions (different loads, muscle fatigue) and requires an intermediate control system to compensate for the intrinsic non-linearities of the skeletomuscular apparatus.

In summary, the primary motor cortex eludes straight-forward paradigms about the nature of the primary variable encoded — a property it shares with many other neural systems. Diligent investigations succeeded in revealing an important basic principle in the form of directional tuning. But from that point on, one has to accept that there is no simple system for the highly complex task of motor control. As expected for a control system, intended movement direction and limb velocity are heavily represented in primary motor cortex. Limb position, in its static property of posture and the dynamic component of displacement, as well as muscle force also appear to be part of the equation. If one acknowledges that all aforementioned variables are crucial for an efficient motor control system, it must not come as a surprise that experimental data supports that each one is represented to some degree in cortical motor areas.

A fundamental property of the motor system is the extremely high number of degrees of freedom to control. Additional complexity is introduced by the infinite number of hand paths to perform a reaching movement from position X to position Y. Given the inherent complexity and ambiguity in control strategies for the musculoskeletal system, it has been proposed that predefined patterns of muscle activation are used to simplify the control problem. The function of the motor cortex reduces to a higher level control of activation patterns that help to make the control problem more manageable and reduce the latency. While differing in the details, many studies support a model in which the neural substrate storing activation patterns is located in spinal modules and driven by unit burst generators [5, 25, 52]. It follows that signals in motor cortex might live in a space of significantly lower dimensionality than expected if all degrees of freedom were treated independently. D'Avella et al. have shown that EMG activity recorded from the hind legs of bullfrogs executing kicking and jumping movements can be modelled by a few time-varying activation patterns at high accuracy [7]. This impressive result underlines that advanced techniques for the quantification of transmitted information or relevance can help to devise new theories about the way motion is planned and encoded in primary motor cortex.

2.3 Coding in the visual system

Of all cortical brain areas, those that are involved in the processing of visual information have been studied the most intensely. First of all, primates — and humans are no exception to that rule — are "visual animals". An unexpected stimulus, a loud sound for example, will trigger a stereotyped reaction in humans: a turn of the head that brings the source of the stimulus into the visual field. Given the importance of visual input for human behavior, it is not surprising that such enormous effort has been made to understand the inner workings of the system that processes this kind of input. A second factor makes the visual system attractive from the experimenter's point of view. It is very easy to present stimuli to the visual system, control experimental conditions and make them sufficiently reproducible. Modern eye tracking systems allow for the precise measurement of gaze direction down to one arc minute at high temporal resolution [8, 56], so the experimenter can tell with certainty where the subject is looking. As neural recordings are generally performed in experimental animals, the availability of species with a highly similar neural system can be essential for the generality and significance of any result obtained. In the case of the visual system, macaques (*macaca*) are frequently chosen as experimental animals for this reason [54].

From the point of view of coding, the visual system is harder to grasp than the motor system. Limb movement can be characterized by position, velocity, acceleration or muscle force in miscellaneous coordinate systems (polar or cartesian, various origins, etc.). Visual stimuli, however, live in an exceptionally high-dimensional space and we lack an expressive language to describe stimulus properties that go beyond basic attributes. Establishing what stimulus attributes a neural system responds to, is intractable under these circumstances. While we leave the problem of a parsimonious stimulus description to others, the difficulties faced here demonstrate the necessity of methods that characterize how relevant a neuron is for the coding of a particular stimulus property and, vice versa, by what set of stimuli the neuron is strongly driven. Before we proceed to techniques that can potentially answer this question in Section 3, we will give a short overview of encoded stimulus properties in the visual cortex.

The most low level cortical area that processes visual information is primary visual cortex V1, cytoarchitectonically Brodmann area 17 and often called "striate cortex". It receives input from the lateral geniculate nuclei via the optic radiation and is organized retinotopically. Hubel and Wiesel discovered that neurons in V1 are tuned for the orientation of lines presented in their receptive field. There is one orientation that elicits the maximum response, the preferred orientation, and the firing rate change observed for other orientations can be described by a tuning curve that is well approximated by a rectified cosine [41].

This discovery led to the famous “ice cube model” of orientation and ocularity selectivity in primary visual cortex [20]. To adequately represent a visual scene, many neurons have to encode the same stimulus quality for a given region of the visual field but exhibit different tuning properties.

From V1 on to higher visual areas, a general trend can be observed. Receptive fields get larger as input from more and more low level areas gets integrated and the visual features represented become more complex. A typical example that fits well into the hierarchical model of the visual system is the middle-temporal area (MT) that plays an essential role in motion perception. Receptive fields are, depending on their eccentricity, about 4° to 25° in diameter and binocular, whereas neurons in V1 are monocular and exhibit smaller receptive field sizes at equivalent eccentricities [13]. These findings agree with the idea of MT being a higher level area for motion discrimination that integrates input from low level areas and V1 in particular. Albeit at different levels of the hierarchy, the concept of a tuning curve can be applied equally well to area MT and striate cortex. Neurons in area MT are tuned for motion direction, motion speed, binocular disparity and object size but the tuning curve for motion direction is at least roughly sinusoidal [3]. Orientation and motion direction tuning shall only serve as examples of coding strategies in the visual system, an exhaustive overview for the about ten cortical areas involved in visual processing is beyond the scope of this work.

Let us take a step back and think about the major difficulties a researcher faces in the visual system. Because of the high dimensionality of visual scenes, the space to explore beyond the most simple stimuli, such as lines of different orientations, sinusoidal gratings or random dot fields, is huge. Nonetheless, the aforementioned stimuli were sufficient to gather an impressive amount of knowledge about low level visual areas, foremost V1. But further up in the hierarchy, a "Trial & Error" approach or reasoning on the basis of connectivity are likely to fail and a more controlled method for sampling the vast space of visual stimuli is called for. Leaving aside the choice of an appropriate search strategy, the ability to quantify how much information the neuron or population under study transmits about a particular stimulus is quintessential. It is the objective of this thesis to explore strategies that are potentially suitable for the task and point out caveats in their application.

2.4 Neural prosthetics

Based on a sufficient understanding of the neural code, one can hope to develop techniques that interface man-made technology with live neural tissue. The most obvious application lies in the field of neuroprosthetics, a field that aims at capacitating prostheses in such a way that they can be naturally controlled by humans. Developing a suitable communication interface is particularly challenging when the control signal cannot be extracted from the peripheral nervous system. Among the techniques that can gather input directly from the brain are electroencephalography (EEG), magnetoencephalography (MEG) and electrocorticography (ECoG) and multi-unit recordings via implants. For complex control tasks, only ECoG and multi-unit recordings promise sufficient spatial and temporal resolution. It has been recently demonstrated that data acquired form a multi-electrode array implanted in the motor cortex of a tetraplegic patient is sufficient for the rudimentary control of a robotic arm [19]. The development of a relevance measure is of high importance for the analysis of the neural code, a prerequisite for successful decoding efforts in neuroprostheses.

Once neuroprostheses leave the purely experimental clinical environment and have to be ready for everyday use, computational resources and power consumption will entail additional constraints. A multi-electrode array, like the one used by Hochberg et al., can easily pick up signals from 80 to 100 neurons [19]. Spike detection, spike sorting and decoding of nearly 100 neurons could be well beyond the means of a practical mobile implementation. An efficient measure of relevance could reduce the computational challenge. During an initial setup period, the prosthesis user could go through a set of imagined movements in order to determine the degree of relevance for efficient decoding of every unit. Based on this data, a feature selection scheme can select a significantly shrunk set of neurons with minimal impact on decoding accuracy.

2.5 Applications of a relevance measure

We will now give an overview of potential applications of mutual information as a relevance measure.

Encoded stimulus property: The review of current opinions about variables encoded in the motor cortex (Section 2.3) and the visual system (Section 2.2) has shown that impressive progress has been made, sufficient for coarse control of rudimentary artificial limbs. Notwithstanding these promising results, it has become apparent that the search for a single coding variable is probably futile and that it can be challenging to disambiguate the influences of different kinematic quantities, e.g. position, velocity and acceleration, in experiments. In the visual system, the major challenge is understanding the encoding of complex stimuli in high level visual areas.

This calls for the application of well-understood tools that can analyze whether the recorded neural signal is efficient at encoding the behavioral variable and transmits large amounts of information or not. Limb kinematics are highly dynamic variables and a putative relevance measure must be able to cope with this property. For this reason, we study mutual information in the context of stimulus models with different degrees of temporal complexity.

Coding schemes: Invariably tied to the encoded stimulus property is the method of encoding a given variable. We hope that a relevance measure can support research that aims at shedding light on this question. If suitable mechanisms for the conversion of spike trains into a random variable are available, mutual information is basically unrestricted with respect to the coding paradigm to be evaluated. In the framework of rate coding, for example, an estimator of the instantaneous firing rate can provide this functionality (cf. Section 3.4), while the temporal coding scheme [14, 10] would call for a representation that preserves precise timing information of the spike train.

Selection of optimally informative neural subpopulations: Advanced electrophysiological recording techniques have put researchers in a position where they can record from more than hundred neurons simultaneously [19]. The amount of data poses computational demands and high power requirements in embedded applications for neuroprostheses. Consequently, it is beneficial to minimize the number of neurons that are necessary for good control over the prosthetic device. As already mentioned in Section 2.4, a feature selection scheme based on mutual information could prove essential for the solution of the problem and identify small neural subpopulations that are yet informative.

The selection of optimally informative neural subpopulations can also help to gain insight into the computational architecture of the brain. Consider the case of a challenging visual discrimination task. Let us assume that it is possible to record from the sensory area that has been shown to provide sensory evidence on which the subject's decision is based (as done in [48, 39]). If the coding scheme of the sensory area is understood to a high degree, the subject's decision can be predicted from the neural signal. In area MT, neurons are sensitive to the direction of motion [3] and in a motion discrimination task, the signal from many recorded units can be turned into a hypothetical decision. The performance of this decision process — the neurometric performance — can subsequently be compared to the behavioral performance of the subject [33]. It is evident that the neurometric performance will correlate with the number of neurons that contribute to the computation. If a subset of neurons can be identified that has been maximized with respect to the information it transmits about the stimulus and exhibits the same (neurometric) performance as the subject, the size of this set is an indicator for the number of input neurons used in the relevant decision making area in the brain.

3 Techniques

3.1 Principles of information theory and mutual information

When Claude Elwood Shannon published the journal article *"A Mathematical Theory of Communication"* [50], it was not clear that he was about to lay the foundation of a new framework which allows for the analysis of information transmission over analog and digital channels. The concepts introduced by Shannon rely exclusively on the statistical properties of random variables to quantify the information content of a signal. *Entropy* covers the "surprise" that is inherent in the distribution of a random variable X. To calculate the entropy $H(X)$ for a discrete random variable, the discrete distribution must be known.

$$H(X) = -\sum_{x \in X} p(x) \log p(x) \tag{3.1}$$

Even though entropy is dimensionless, it is typically annotated with "bits" if the logarithm is to base 2, or "nats" for the natural logarithm $\log_e$.

This definition of entropy has some intuitive properties: A random variable which takes only one value, for example, will have an entropy of zero. If a random variable is defined over a set of two equiprobable elements, then it will have an entropy of 1 bit. The same number of bits is required for an optimal encoding of the random variable [6].

In our application, discrete distributions occur more often than continuous distributions. For the sake of completeness, note that the sum over the distribution support can be replaced by an integral to define *differential* *entropy* for continuous variables [6]:

$$H_{\text{diff}} = -\int_{-\infty}^{\infty} p(x) \log_2 p(x) dx \tag{3.2}$$

Extending our focus from one variable to two or more variables leads us to quantities that are well-suited to describe information transmission. One can ask the question how much knowledge about a random variable X can be obtained simply by knowing the value of another random variable Y. Or to use the notion of entropy as introduced above, by how much does knowledge of Y reduce the entropy of X. The mutual information $I(X;Y)$ of random variables X and Y can be calculated in different ways, either directly from the joint distribution $(X;Y)$ and the marginals X and Y, or expressed as the difference of entropies:

$$I(X;Y) = H(X) - H(X|Y) \tag{3.3}$$

$$I(X;Y) = H(Y) - H(Y|X) \tag{3.4}$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{3.5}$$

$$I(X;Y) = \sum_{x \in X, y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{3.6}$$

Mutual information can be interpreted as a measure of dependence between random variables. Statistically independent variables will have a mutual information of 0, because in this case it holds that $p(x,y) = p(x)p(y)$ and consequently $\log \frac{p(x,y)}{p(x)p(y)} = \log 1 = 0$ for all $x \in X, y \in Y$.

While the Pearson correlation coefficient, the standard quantity to measure dependence, can uncover linear relationships between scalar variables, mutual information can detect more general dependencies between variables and extends to the multi-dimensional case. Any deviation from independence, which, if fulfilled, guarantees that $p(x, y)$ can be written as a product distribution, will contribute to the sum in equation (3.6). More formally, mutual information is the Kullback-Leibler divergence between the joint distribution $(X; Y)$ and the product distribution of X and Y.

$$I(X;Y) = D_{KL}(X||Y) = \sum_{x \in X, y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \tag{3.7}$$

A neuroscientist studying a sensory system might be interested in the amount of information that can be extracted from the neural signal N about a specific stimulus S. Measures of information can give insight into the predictive power of the neuronal population under study. A similar paradigm can be developed for motor areas of the brain. The only difference is that, in the motor system, the neural signal does not reflect the reaction to the stimulus but a control signal that exerts a causal influence on motor output. For the sake of simplicity, we will refer to the behavioral variable as "stimulus" without implying causation. The discussion will be equally applicable to sensory and motor paradigms.

3.2 Different mutual information paradigms in experimental setups

In the previous section, information theory was introduced as a general technique with a wide range of applications. The quantities entropy and mutual information can be calculated from two discrete random variables X and Y. This section is devoted to establishing a connection between information theory and its application to neuroscience.

Instead of X and Y, we will use variable S for the stimulus and N for the neural signal. The meaning of the stimulus depends heavily on the experimental setup. In the visual system, S could describe a parameter like contrast [51], direction of motion [32], or a category index (e.g., when faces are to be discriminated from houses and other objects [55]). When motor control is studied, S will be a limb velocity in space or a joint angle, but could also represent position or acceleration. In any case, it must be possible to map the stimulus to a discrete set of values. The case where S is continuous will not be covered in this thesis, but we are going discuss a model that approximates a continuous stimulus by a discrete set of stimulus values in Section 4.6.

The neural signal N must represent the activity of the neural system under study in some way. In Section 3.4, common choices for N will be discussed in greater detail. To give some intuition, N will typically be the instantaneous firing rate of one neuron or a group of neurons, or it can be a high-dimensional vector that conveys precise timing information about spike trains.

A heavily used experimental setup in the study of sensory systems is based on the repeated presentation of stimuli drawn from a set $\mathcal{S}$. When a stimulus $s \in \mathcal{S}$ is presented to the subject, the neural response within a fixed time interval $[0, T]$ is recorded and then represented numerically by some means of quantification, binning (Section 3.4), for example. For the sake of clarity, note that $\mathcal{S}$ is a set of stimuli and S a random variable. In most cases, S will be defined over $\mathcal{S}$, i.e., $S : \mathcal{S} \rightarrow \mathbb{R}$, but it is not necessarily the case and S could potentially be defined over some other set that is related to $\mathcal{S}$.

A first paradigm in the application of information theory was developed by Strong et al. and subsequently used in the study of H1 neurons in the fly visual system [53]. The entropy $H(N)$ of the neural signal and the conditional entropy $H(N|S)$ are estimated from the data. As $H(N|S)$ is a measure of the uncertainty about the neural signal that remains if the stimulus is known, it is characterized as neural noise. A look at the following expansion

$$I(N;S) = H(N) - H(N|S) = H(N) - \sum_{s \in S} p(S=s) \cdot H(N|S=s) \quad (3.8)$$

reveals that one can interpret $I(N;S)$ as the average information transmitted over all stimulus conditions. As Borst and Theunissen point out, this approach does not allow for the identification of stimulus conditions under which the neural system transmits particularly high amounts of information [4]. The same authors also emphasize that reliable estimates of $I(N;S)$ can only be gathered from very large data sets, as no assumptions are made about the probability distributions from which entropies and mutual information are estimated. While the latter is certainly true, it is not clear that better estimates can be obtained if the distributions are assumed to be of a certain type. A lack of well-founded arguments about what distribution to use often leads experimenters to choose the normal distribution. Without doubt, the amount of information required to estimate mean and variance (in the one-dimensional case) is magnitudes smaller than in the unconstrained case, but violations of normality will have detrimental effects on estimates of information-theoretic measures.

In spite of the aforementioned caveat, we will describe the "spectral" method that relies on assumptions about properties of the neural response. More specifically, amplitudes of the neural signal in the frequency domain are assumed to be Gaussian. Spike trains naturally live in the time domain and can be viewed as point processes such that only the times of spike occurrences are relevant. But one can also subscribe to the perspective that the timing of spikes relative to each other and relative to the stimulus is the more expressive quantity. For this purpose, the spike train is converted into the frequency domain via Fourier transform, the result of which are distributions of signals at a discrete set of frequencies. The essential assumption of the spectral method states that the distribution at each frequency is Gaussian [4]. Because mean and standard deviation are sufficient statistics of the normal distribution, only two parameters have to be estimated from the data at each frequency value. Compared to the direct method, the amounts of data required for reliable estimates is expected to be significantly smaller. Apart from that, the spectral and direct approaches are identical and the basis for similar conclusions. In Section 3.5, the spectral method will resurface because properties of the normal distribution make it a suitable choice for the derivation of an upper bound on mutual information.

The third method aims at the derivation of mutual information values for each stimulus condition. Potentially, this more fine-grained analysis tool can provide insight into stimulus properties that are coded for by the neural system under study. Rewriting Equation (3.6) as

$$
\begin{aligned}
I(X;Y) &= \sum_{x\in X, y\in Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} = && (3.9)\\
&= \sum_{x\in X}\left(p(x)\sum_{y\in Y} p(y|x)\log\frac{p(x,y)}{p(x)p(y)}\right) && (3.10)
\end{aligned}
$$

reveals that an information value can be extracted for each $x \in X$ [4]:

$$I(x;Y) = \sum_{y\in Y} p(y|x)\log\frac{p(x,y)}{p(x)p(y)} \tag{3.11}$$

The resulting information value for each $x \in X$ quantifies the reduction in uncertainty about x if Y can be observed. Borst and Theunissen point out that the breakdown into information values for each stimulus condition suggests an alternative to traditional tuning curves. Instead of plotting the neural response as a function of the stimulus, a stimulus-information curve can visually represent the discriminability of different stimulus conditions by the neural signal.

3.3 Techniques for MI estimation

Several techniques have evolved for the calculation of mutual information. If both variables X and Y are discrete, the joint distribution (X,Y) will also be discrete, and in that case, equation (3.6) already dictates the method to calculate $I(X;Y)$:

$$I(X;Y) = \sum_{x \in X, y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{3.12}$$

If Maximum Likelihood estimates of the probabilities $p(x)$, $p(y)$ and $p(x,y)$ are used, this approach is typically referred to as the "plug-in estimate" or the "naïve method" [34][1]. To obtain ML estimates of a discrete distribution, the relative frequencies of all pairs $(x,y) \in X \times Y$ are extracted from the data and treated as if they were the true probabilities.

In cases where the distributions of variables X and Y are continuous, binning can be used to approximate the true density functions by a discrete distribution. Subsequent steps in the estimation of mutual information will be based on the discrete estimate. The major shortcoming of histogram based methods is the arbitrary placement of bin boundaries that makes the estimate very sensitive to slight shifts of a few data points. To maximize the amount of structure in the data that is taken into account for the density estimate, one can resort to *kernel density estimates*. This class of estimators creates continuous distributions from a finite number of data points by spatially smoothing out every observation and locally summing their contributions. Graphically, the smoothing operation can be visualized as replacing every data point by a spread out blob of probability mass. If N iid observations $x_1, \ldots, x_N$, a *kernel function* $K(\cdot)$ and a bandwidth h are given, the

[1] The term "plug-in estimate" is often used in the context of entropy and mutual information interchangeably. So whenever MI is estimated directly from equations (3.6), (3.3) or (3.5) via ML estimates, it is of "plug-in" type.

kernel density

$$\hat{f}_h(x) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right) \tag{3.13}$$

is a consistent estimate of the true density function [37]. The kernel function $K(\cdot)$ determines the shape of the "bump" and the bandwidth how much it is spread out in space. A common choice for $K(\cdot)$ is the Gaussian kernel

$$K_{\text{Gauss}} = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} \cdot \exp\left(-\frac{1}{2} x^\top \Sigma^{-1} x\right) \tag{3.14}$$

with mean zero and covariance matrix Σ. In the one-dimensional case, only the ratio of the Gaussian kernel's variance and the bandwidth h has to be determined, but the multidimensional case poses more challenges. The covariance matrix Σ has many degrees of freedom that describe the distribution's orientation in space. To reduce the number of parameters, multidimensional kernels are typically chosen to be product distributions and Σ consequently a diagonal matrix. Estimation of h is aimed at the minimization of the *Asymptotic Mean Integrated Squared Error*. For a more detailed explanation of kernel density estimation (KDE), we refer to [37] and [11].

The approximate density function derived above can then be plugged into equations for differential entropy and mutual information. It has to be emphasized that kernel density estimates are geared towards continuous distributions and are not suited for discrete distributions because spreading out data points changes the entropy of the random variable.

Initially designed for classification problems, a specialized class of kernels was developed for discrete distributions by Aitchison et al. [1]. The so called *integer kernels* spread out probability mass, in a similar fashion as kernels for continuous settings do, but obey the discrete nature of the distribution. Integer kernels guarantee that no probability mass will be distributed onto values that are not in the discrete support set.

We define a discrete support set $X = \{x_1, \ldots, x_N\}$ and assume that a valid discrete distribution is already given such that $0 \leq p(x_i) \leq 1, \ \forall i \in \{1, \ldots, N\}$ and $\sum_{i=1}^{N} p(x_i) = 1$. Following the method in [49], we define the kernel $k_h(x, x_i)$ that determines what fraction of the probability mass $p(x_i)$ at point x_i is spread out to some other $x \in X$:

$$k_h(x, x_i) = \frac{h^{\|x - x_i\|^2}}{\sum_{j=1}^{N} h^{\|x_i - x_j\|^2}} \tag{3.15}$$

As before, h is a bandwidth parameter that is typically chosen by some heuristic. As proposed in [49], one can concentrate the probability mass of the point x_i on its immediate proximity by setting $h = 0.05^{\frac{1}{\sigma^2(X)}}$ where $\sigma(X)$ is the standard deviation of the observations X. It follows that 90% of the probability mass will remain within one standard deviation of x_i. Given the distribution $p(x)$ and the kernel, a new distribution

$$\hat{p}(x) = \sum_{i=1}^{N} \left(p(x_i) \cdot \frac{h^{\|x - x_i\|^2}}{\sum_{j=1}^{N} h^{\|x_i - x_j\|^2}} \right) \tag{3.16}$$

can be obtained. If we assume that $p(x)$ is a discrete distribution directly obtained from the data, it must be justified why $\hat{p}(x)$ should provide a better representation of the neural signal than $p(x)$. One argument is that applying the kernel is equivalent to a smoothing operation of the distribution that respects the discrete support set. Smoothing the distribution can express uncertainty in the measurements. If the number of spikes emitted by a neuron is summed over a time window (cf. Section 3.4 for a description of binning), it can be reasonable to spread out the probability mass of an observation $f_i \in \mathbb{N}$. A slight shift of the window could have changed the observation to (f_i+1) or (f_i-1) and significantly alter the distribution because of the integrality of observed spike counts. Smoothing the distribution can express this phenomenon and thus result in a more natural representation of the response profile.

In the context of information processing, applying the smoothing operator is comparable to the transformation of a random variable. The information processing theorem dictates that $I(N;S) \geq I(f(N);S)$ for any function $f(\cdot)$ [6]. The mutual information of the stimulus and the neural signal is thus expected to decrease or remain constant if $\hat{p}(x)$ instead of $p(x)$ is chosen to characterize the response of the neural system. An evaluation of integer kernels within the framework of our synthetic datasets is presented in Section 5.5.

Kernel density estimation is a general technique to estimate probability distributions from a finite amount of data and not specific to the estimation of entropy or mutual information. An approach that completely circumvents the estimation of probability distributions is based on nearest-neighbor methods. Kraskov et al. developed a technique that estimates mutual information from continuous distributions and rests on the observation that the distance of a point to its k-nearest neighbor in the joint distribution in relation to the same distance in the marginal distributions $p(x)$ and $p(y)$ allows for reliable, unbiased estimates of mutual information [27].

If we make N joint observations of two random variables X and Y that can also be represented in the joint space $Z = (X, Y)$ and both X and Y are metric spaces[2], we define the metric

$$\|z\|_\infty = \max\left(\|x\|, \|y\|\right) \quad \forall z = (x, y) \in Z \tag{3.17}$$

on Z. Now let $\frac{\varepsilon(i)}{2}$ denote the distance between z_i and its k-nearest neighbor. Similarly, $\frac{\varepsilon_x(i)}{2}$ and $\frac{\varepsilon_y(i)}{2}$ are the distances between the same two points in the projected spaces X and Y. It follows that $\frac{\varepsilon(i)}{2} = \max\left(\frac{\varepsilon_x(i)}{2}, \frac{\varepsilon_y(i)}{2}\right)$. Then we count the number of points closer than $\frac{\varepsilon(i)}{2}$, measured in X and Y:

$$n_x(i) = \left|\left\{j : \|x_i - x_j\| < \frac{\varepsilon(i)}{2}, i \neq j\right\}\right| \tag{3.18}$$

$$n_y(i) = \left|\left\{j : \|y_i - y_j\| < \frac{\varepsilon(i)}{2}, i \neq j\right\}\right| \tag{3.19}$$

An estimator for the mutual information of X and Y is given by

$$I(X;Y) = \psi(k) - \frac{1}{N}\left(\sum_{i=1}^{N} \psi(n_x(i) + 1) + \psi(n_y(i) + 1)\right) + \psi(N), \tag{3.20}$$

where $\psi(x)$ is the digamma function [27]. Kraskov et al. demonstrate that the estimator is unbiased and converges rapidly for small sample sizes.

[2] As random variables are actually functions, we clarify that the space they map into shall be metric, i.e., for a random variable $X : \Omega \to \mathbb{R}$ it holds that $\|\cdot\|$ is a metric on $\mathbb{R}$ and on $\mathbb{R}^n$, respectively, for an n-dimensional random variable

The nearest-neighbor method can be readily extended to higher-dimensional spaces and has been shown to exhibit similarly positive properties as in the scalar case. For the application to neuroscience, the requirement that X and Y must be continuous random variables is often not satisfied, e.g., in the case where stimuli are drawn from a finite set. If one or both of the random variables X and Y are discrete, the nearest-neighbor method breaks down because the nearest neighbor will often have distance 0. Many methods to represent the neural signal will also result in the random variable N being discrete. For this reason, Victor et al. choose to project the neural signal into a continuous space via Legendre polynomials [57].

If exactly one of the variables is continuous, without loss of generality we choose N, the alternative formulation of mutual information as

$$I(N;S) = H(N) - H(N|S) = H(N) - \sum_{s \in S} p(S=s) \cdot H(N|S=s) \quad (3.21)$$

gives rise to a second nearest-neighbor estimator. Going back to works by Kozachenko et al., entropy can be estimated by [57, 26]

$$H(X) = \frac{1}{N} \sum_{i=1}^{N} \log_2 \frac{\varepsilon(i)}{2} + \log_2 \left(2(N-1)\right) + \frac{\gamma}{\ln(2)}. \quad (3.22)$$

As above, $\frac{\varepsilon(i)}{2}$ denotes the distance of x_i from its nearest neighbor[3] in X. The major drawback of the expansion in Equation 3.21, as identified by Kraskov et al. [27], is the erratic error accumulation of multiple entropy

[3] In the method by Kraskov et al., $\frac{\varepsilon(i)}{2}$ measured the distance to the k-nearest neighbor for some $k \geq 1, k \in \mathbb{N}$. Kozachenko's estimator of entropy can also be extended in this way.

estimators. There is no guarantee that errors in entropy estimates will cancel out in a favorable way and yield a small error in the mutual information estimate. Kraskov et al.'s direct estimator (Equation 3.20) is less prone to error accumulation and should be preferred whenever possible.

3.4 Choice of random variables in neuroscientific settings

So far, we have discussed techniques for the estimation of mutual information between two variables X and Y, without ascribing a specific meaning to them. Experimental paradigms for the use of mutual information were considered in section 3.2 without detailing the representation of the stimulus S and the neural signal N.

One situation commonly encountered in the study of sensory systems is that of a discrete set of stimuli $\mathcal{S} = \{s_1, s_2, \ldots, s_{|\mathcal{S}|}\}$ being repeatedly presented to a subject (cf. Section 3.2). A time window T is defined — either identical to the stimulus presentation interval or arbitrarily chosen — in which the response of the neural system is recorded. The data obtained that way can be naturally split into one discrete component, the number of spikes that occurred in the time window, and one continuous component for the precise time of every spike. Most experimenters adhere to the technique of binning spikes in order to transform the data into a more manageable representation without losing too much information. After splitting the time window $[0, T]$ (relative to stimulus onset) into bins of size b, the number of spikes that fall within the limits of a given bin are counted and represented as a vector $\left(n_1, n_2, \ldots, n_{\lceil \frac{T}{b} \rceil}\right)$ or a matrix if multiple neurons are analyzed.

If the bin size is chosen sufficiently small, no more than one spike will occur in any given bin due to the refractory period of neurons. It follows that this vector or matrix is one possible discrete and high-dimensional representation of the neural signal [53]. Part of the information content of the neural signal is lost due to binning and the magnitude can be expressed as a function of the bin size b [57]. Recall the definition of differential entropy:

$$H_{\text{diff}} = -\int_{-\infty}^{\infty} p(x) \log_2 p(x) dx \tag{3.23}$$

Now one can look at the discretized version of this distribution and calculate the entropy [40]:

$$\begin{aligned} H_{\text{disc}}(b) &\approx -\sum_i bp(x_i) \log_2 bp(x_i) \approx -\log_2 b - \sum_i bp(x_i) \log_2 p(x_i) \\ &\approx -\log_2 b - \int_{-\infty}^{\infty} p(x) \log_2 p(x) dx = H_{\text{diff}} - \log_2 b \end{aligned} \tag{3.24}$$

Note that this derivation was performed for entropy and not mutual information. However, as b approaches zero, the discretized mutual information $I_{\text{disc}}(N; S)$ converges to the true mutual information $I(N; S)$ [40].

A shortcoming of binning approaches is the arbitrary and hard to justify position of bin boundaries. Jitter of a few microseconds can move a spike from one bin to an adjacent one. An intuitive solution to this problem is the application of a sliding window which smoothes out the effects of bin positioning. The number of spikes that fall in the window of length w is counted and transformed into an estimate of the instantaneous firing rate by division by w. Then, the window is shifted by an offset s and the procedure repeated.

Binning is a special case of the sliding window method where w equals s, but more frequently, $s \ll w$ is chosen. Overlap of adjacent windows induces dependence between consecutive observations of the random variable and it has to be established that no bias is introduced into mutual information estimates. Accepting that no additional information about the stimulus can be introduced by having convolution windows overlap is easy. On the other hand, a potential decrease of the estimated mutual information cannot be dismissed and is in fact quite likely. In Section 5, the magnitude of this effect will be examined, but we will state here that we found the decrease in mutual information negligible.

If the window size w is smaller than the stimulus period T, the random variable of the neural signal is observed multiple times for a single stimulus presentation. The assumption about the neural code imposed by treating these multiple observations equally, is that neural responses are stationary within the time window $[0, T]$. In the framework of rate coding, such a restriction is of limited consequence, even though the well-known phenomenon of adaptation [24] already entails noticeable deviation from the assumption. Even more precarious are the implications for the temporal coding paradigm [14, 10]. If one embraces the idea that precise timing of single spikes is a means of information transfer in neural systems, then sliding windows are certainly an inappropriate analysis tool.

Small bin sizes of a few milliseconds ensure that the neural signal is binary and provide sufficient precision to capture the timing of single spikes. A random variable of dimensionality $\lceil \frac{T}{b} \rceil$ can be defined by concatenating

$\lceil \frac{T}{b} \rceil$ adjacent bins into a vector. It can be argued that of all approaches discussed so far, this one will cover the most information and should be considered exclusively. But the amounts of data required for reliable estimates of mutual information are prohibitive in the case of multi-dimensional representations of the stimulus variable.

Victor developed a method which aims at circumventing the information loss incurred by binning. His strategy is based on the aforementioned dual nature of neural responses as a discrete number of spikes and a continuous component which represents the exact timing information [57]. This property can be put to use for the calculation of mutual information in a two-stage approach.

$$\begin{aligned} I(N;S) &= I_{\text{count}}(N;S) + I_{\text{timing}}(N;S) = \\ &= I_{\text{count}}(N;S) + \sum_{n=1}^{\infty} p\left(d(x) = n\right) I^{(n)}_{\text{timing}}(N;S) \end{aligned} \tag{3.25}$$

$I_{\text{count}}(N;S)$ denotes the mutual information of the stimulus and the neural signal if the stimulus is represented solely by the number of spikes that occurred in a trial. If the precise timing in a spike train transmits additional information, it will be reflected by a non-zero contribution $I_{\text{timing}}(N;S)$. A further expansion of $I_{\text{timing}}(N;S)$ in the second line of Equation 3.25 conditions the timing information on $d(x)$, the number of spikes emitted in the trial. So $I^{(n)}_{\text{timing}}(N;S)$ stands for the mutual information of stimulus and neural signal, restricted to trials in which n spikes occurred.

Traditional methods are used for the estimation of $I_{\text{count}}(N,S)$ (cf. Section 3.3). It is noteworthy that all neural responses are stratified into different categories based on the number of spikes emitted during the stimulus presentation. One consequence is that neural responses are now represented in spaces of different dimensionality and evade a straight-forward method for MI calculation. For this reason, all spike trains with n spikes are embedded into an Euclidean space of dimension $r = \min(n, D)$ for some maximum dimensionality D. To homogenize the mapping and maximize the uniformity of spike trains in the time domain, all spike times are transformed via a monotonic map

$$\tau_j = -1 + 2\frac{j - \frac{1}{2}}{M}. \tag{3.26}$$

Legendre polynomials map every spike train into a continuous space of dimensionality r. Then $c_h(x_j)$ is the h-th coordinate in the codomain and P_h the h-th Legendre polynomial.

$$c_h(x_j) = \sqrt{2h+1}\sum_{k=1}^{n} P_h(\tau_k) \tag{3.27}$$

The h-th Legendre polynomial is defined as

$$P_h(z) = \frac{1}{2\pi i}\oint (1 - 2tz + t^2)^{-\frac{1}{2}} t^{-h-1} dt \tag{3.28}$$

and is orthogonal to any P_j for $j \neq h$ on the interval $[-1, 1]$ [15]. Now that the neural response is represented in a continuous space, the nearest-neighbor entropy estimator in Equation 3.22 can be plugged into Equation 3.25 after rewriting $I^{(n)}_{\text{timing}}(N, S)$ as the difference of two entropies (cf. Equation 3.21). We are not going to derive the extension to higher-dimensional spaces necessitated by neural populations. The method introduced by Victor is inherently limited by the representation of the neural signal as a fixed-length vector of spike counts or firing rates. A representation of this type is appropriate if different stimuli are presented sequentially for a fixed amount of time T. But it does not adapt well to more dynamic experimental conditions in which the stimulus changes rapidly.

3.5 Bounds on mutual information

As huge amounts of data are required to obtain reliable estimates of the probability distributions $p(S;N)$ and particularly $p(N|S)$, many experimenters have abandoned attempts to directly calculate mutual information and instead reverted to constructing lower and upper bounds on mutual information.

The data processing theorem in equation (3.29) states that no transformation of the observations Y can add information about X [6].

$$I(X;Y) \geq I(X;f(Y)) \tag{3.29}$$

It follows that $I(X;f(Y))$ will be a lower bound on the mutual information $I(X;Y)$. Choosing a single random function $f(\cdot)$ will probably not yield a tight bound. But as $\max_{f(\cdot)} I(X;f(Y)) \leq I(X;Y)$ is a direct consequence of equation (3.29), it is possible to tighten the bound. Under the assumption that $f(\cdot)$ is differentiable, gradient descent is one method to achieve the desired maximization [21].

It is precarious to counter insufficient amounts of data by resorting to the application of mutual information bounds. Even after the maximization of $f(\cdot)$, it is not possible to make any guarantees about the gap between the actual information and the bound or estimate its magnitude. Besides the convenient property of differentiability, there exists little guidance for an appropriate choice of $f(\cdot)$, neither the dimensionality of its codomain nor the type of function. Linear filters are easy to compute, well understood and are thus a good option to explore initially.

Bounding mutual information from above typically makes assumptions about the stimulus distribution or properties of the neural signal. Rozell et al. point out that the validity of these assumptions tends to be questionable and demonstrate that violations of the imposed restrictions can lead to heavily erroneous results [40]. Based on equation (3.3), one might suspect that $H(X)$ is a simple upper bound for $I(X;Y)$. If X and Y are continuous random variables, the proposed bound can be rejected immediately, as differential entropy (cf equation (3.2)) can be negative [6]. For the common situation where data is binned and X and Y are discrete random variables, it holds that $H_{\text{disc}}(X) \geq 0$ (as defined in Equation (3.24)) and $H_{\text{disc}}(X)$ is indeed an upper bound for $I_{\text{disc}}(X;Y)$. More important though, is the fact that $H_{\text{disc}}(X)$ does not bound $I(X;Y)$, the mutual information of the continuous variables, from above [40].

Different paradigms for the use of mutual information in neuroscientific experiments were discussed in Section 3.2. Among them, the spectral method lends itself to the derivation of an upper bound on mutual information. Recall that the neural signal can be Fourier transformed into the frequency domain and that the distribution at each frequency is assumed to be normal. It is known that, among all probability distributions with mean μ and standard deviation σ, the normal distribution $\mathcal{N}(\mu, \sigma^2)$ has the maximum entropy [6]. Correspondingly, a frequency spectrum that is assumed to be normally distributed at each frequency bin has a total entropy greater or equal than any other spectrum with identical mean and standard deviation values at each frequency bin. The mutual information value calculated via the spectral paradigm bounds the true mutual information from above [4].

Another Gaussian bound predicates on the hypothesis that the stimulus is normally distributed. Likewise, the neural response is expected to be the result of an additive Gaussian noise process independent of the stimulus [40]. After fitting normal distributions to the stimulus and the neural signal, unconditioned and conditioned entropies of the neural signal can be calculated and plugged into the formula for mutual information (Equation (3.3)). The same arguments as above show that the resulting value is an upper bound on mutual information. Obviously, it is hard to justify such restrictive assumptions as normality of both stimulus and neural signal and few experiments in practice will ever give rise to a dataset where they hold. Rozell at al. give an account of the shortcomings of bounds on mutual information and provide empirical evidence to confute their validity [40].

The restrictive and questionable integrity of bounding schemes for mutual information suffice to motivate the development of good estimators for mutual information. It also becomes evident that studying the proneness of estimators to exhibit artifacts in small datasets and analyze convergence is highly beneficial for a solid framework of information-theoretic methods in neuroscientific applications.

4 Generation of synthetic data

4.1 Model of spike generation

For the purpose of generating synthetic spike trains, the Poisson process is one of the most widely used stochastic processes. There is evidence that statistics of real neurons are well approximated by Poisson processes but also exhibit some deviations [47, 9]. What makes the Poisson process interesting is the fact that it is well understood. If some property or statistic cannot be derived for the Poisson process, it is generally unlikely that it can be done for any of the more realistic models of spike generation [22]. Point processes can be characterized by their intensity which is identical to the firing rate in the case of the Poisson process.

4.2 Bernoulli process approximation

In order to synthesize binned spike trains, the generation of spiking events can be reduced to sampling from a Bernoulli distribution by simulating coin flips. For a given firing rate r in $\frac{\text{spikes}}{\text{second}}$ and bin size b, we calculate the probability that the neuron under study generates one spike in a given bin: $P_{\text{fire}} = rb$. For each bin, the presence of a spike is determined based on a coin flip with the probability of success equal to P_{fire}.

Trivially, for a trial of length T, which we divide into $M = \frac{T}{b}$ bins, we expect a firing rate of

$$\mathbf{E}(\frac{n}{T}) = \frac{1}{T} \cdot \mathbf{E}(n) = \frac{1}{T} \cdot M \cdot P_{\text{fire}} = \frac{1}{T} \cdot \frac{T}{b} \cdot r \cdot b = r$$

4.3 Deviations from Poisson process statistics

The procedure described above will yield a spike train with Poisson process statistics in the limit of $b \rightarrow 0$. As b approaches zero, the number of bins M diverges to infinity but $M \cdot P_{\text{fire}} = \frac{T}{b} \cdot r \cdot b = T \cdot r$ remains constant. This justifies the transition from the Bernoulli distribution to the Poisson distribution. Consequently, the probability of n spikes to occur within period T is given by

$$P_T[n] = \frac{(rT)^n}{n!} \exp(-rT)$$

As the number of spikes that occur between time t_0 and $t_0 + T$ follows a Poisson distribution with constant rate r, the time series is a homogeneous Poisson process.

For spike trains modeled by a Poisson process, all spike patterns with a fixed number of spikes n within a time window T are generated with equal probability. The binary output of the given spike train generation mechanism apparently violates this condition as no more than one spike can occur in any bin. As a result, it follows that the statistics of our model deviate more significantly from a Poisson process for high firing rates (which make multiple spikes in a single bin more likely) and large bins.

4.4 Sampling from the interspike interval distribution

By abolishing binning, the aforementioned issues can be circumvented. It is a well known fact that waiting times between events follow an exponential distribution for a homogeneous Poisson process with rate r:

$$P[\Delta t = x] = r \cdot \exp(-rx)$$

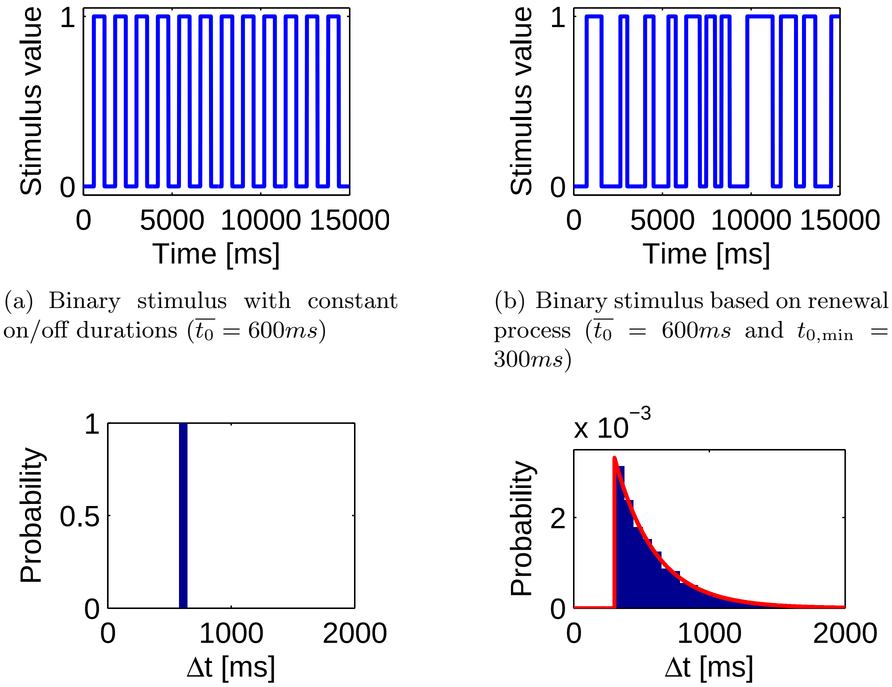
Spike trains can be generated by sampling inter-spike intervals from this distribution. It follows that the number of spikes assigned to a single bin is not limited anymore. One can argue that this approach will yield spike trains that follow the statistics of a Poisson process more closely because the coin flip based method relies on the properties of the random number generation over a small fraction of the unit interval.

As a drawback, nothing prevents two spikes from occurring within less than one millisecond, an unrealistic value due to the refractory period of biological neurons. Renewal processes, a generalization of the Poisson process, give more control over the properties of inter-spike intervals and allow for the incorporation of realistic refractoriness.

4.5 Binary stimulus

For the simplest and most approachable model, the stimulus S is taken from the domain $\mathcal{S} = \{0, 1\}$. For every neuron we define two firing rates f_0 and f_1. Whenever the stimulus has value 0, firing rate f_0 is used for spike generation and f_1 if the stimulus has value 1. Not all stimulus values have necessarily the same likelihood of occurrence, so we define a probability $P(S = 1)$ which determines the likelihood of the stimulus taking value 1.

At the lowest complexity level, the stimulus will have value 0 for a duration of $\overline{t_0}$ and value 1 in segments of length $\overline{t_1} = \frac{P(S=1) \cdot \overline{t_0}}{P(S=0)}$. Obviously, such a choice will cause the stimulus to be highly predictable (Figures 2(a) and 2(b)).



(a) Binary stimulus with constant on/off durations ($\overline{t_0} = 600ms$)

(b) Binary stimulus based on renewal process ($\overline{t_0} = 600ms$ and $t_{0,\min} = 300ms$)

(c) The trivial probability distribution for the simple binary stimulus with constant on/off times

(d) Probability distribution of the time between changes of the stimulus value. Theoretical distribution in red and empirical histogram in blue.

Figure 2: The binary stimulus

To counter the effects of a highly determined stimulus time series, we add an element of randomness into the stimulus choice at any given point in time. As before, $\overline{t_0}$ is the average time of segments where the stimulus has value 0. Additionally, refractory periods $t_{0,\min} < \overline{t_0}$ and $t_{1,\min} = \frac{P(S=1) \cdot \overline{t_0}}{P(S=0)} < \overline{t_1}$ for the stimulus set minimum times where the stimulus value remains constant.

To fully describe the temporal behavior of the stimulus, we define two series of variables $(t_1^{S_0}, t_2^{S_0}, \dots)$ and $(t_1^{S_1}, t_2^{S_1}, \dots)$. Starting at time $t = 0$, the stimulus will take value 0 for a duration of $t_1^{S_0}$, change to value 1 and keep it for $t_1^{S_1}$, then switch back to 0 for a segment of length $t_2^{S_0}$, and so on. The durations of the stimulus segments are given by the following formula:

$$
\begin{aligned}
t_i^{S_0} &= t_{0,\mathrm{min}} + d_i \\
&\quad \text{with } d_i \sim \mathrm{Exp}\left(\overline{t_0} - t_{0,\mathrm{min}}\right) \\
t_i^{S_1} &= t_{1,\mathrm{min}} + e_i \\
&\quad \text{with } e_i \sim \mathrm{Exp}\left(\overline{t_1} - t_{1,\mathrm{min}}\right)
\end{aligned}
$$

where

$$
\begin{aligned}
d_i &\sim \mathrm{Exp}\left(\overline{t_0} - t_{0,\mathrm{min}}\right) \\
e_i &\sim \mathrm{Exp}\left(\overline{t_1} - t_{1,\mathrm{min}}\right)
\end{aligned}
$$

and $\mathrm{Exp}(\lambda)$ is the exponential distribution with parameter λ and probability density function $f_\lambda(x)$:

$$
f_\lambda(x) = \begin{cases} \lambda \cdot \exp\left(-\lambda x\right) & \text{for } x \geq 0 \\ 0 & \text{else} \end{cases} \tag{4.1}
$$

First, we guarantee that the stimulus value will not change before the refractory period has expired. The remaining period is sampled from an exponential distribution and it follows that the time series $(t_i^{S_0} - t_{0,\min})_{i\in\mathbf{N}}$ constitutes a Poisson process with intensity $\overline{t_0} - t_{0,\min}$, and analogously for $(t_i^{S_1} - t_{1,\min})_{i\in\mathbf{N}}$. The properties of the exponential distribution also cause segment lengths to have standard deviations

$$\sigma_0 = \overline{t_0} - t_{0,\min} \tag{4.2}$$

$$\sigma_1 = \overline{t_1} - t_{1,\min}. \tag{4.3}$$

Figures 2(b) and 2(d) show an example stimulus generated via this procedure and its distribution of the time between changes of the stimulus value. For three arbitrarily defined neurons, spike trains were generated as described in Section 4.4 for a short segment of the binary stimulus and plotted in Figure 3.

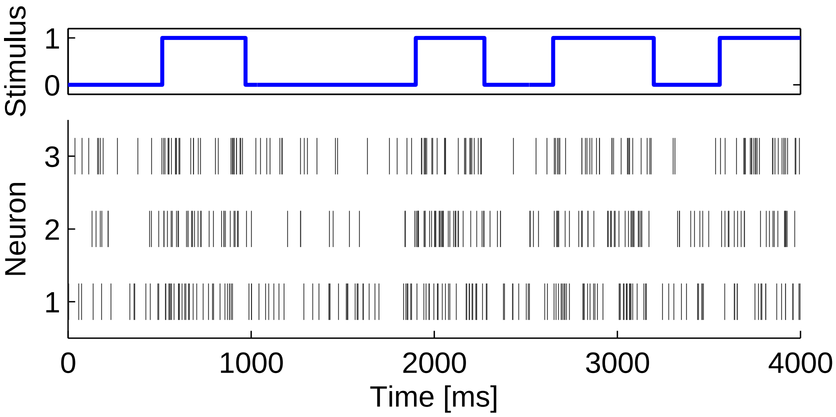


Figure 3: Raster plot of an example spike train generated from the binary stimulus using the Poisson process. The firing rates (f_0, f_1) for the three neurons were as follows: (30Hz, 35Hz); (20Hz, 45Hz) and (15Hz, 50Hz)

4.6 n-valued stimulus

It is obvious that the binary stimulus model introduced in the previous section is often not sufficient to cover the complexity of experimental setups encountered in practice. For this reason, we extend the concept to an n-valued stimulus and set $\mathcal{S} = \{s_1, \ldots, s_n\} = \{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\}$. In the limit of $n \to \infty$, a quasi-continuous spectrum of stimulus values can be achieved. In practice, the distinction between a truly continuous random variable and one that is defined over a fine-grained large set $\mathcal{S}$ can be negligible. To go back to the example of studying movement representation in motor cortex, the limiting factor is the resolution at which limb position can be reliably measured.

However, the n-valued stimulus is still far from being applicable to a setting where limb positions in a two- or three-dimensional space must be represented because of the restrictions we enforce. Similar to the binary stimulus, we define a firing rate f_i for $i = 1, \ldots, n$ that defines how many spikes per second a neuron fires when stimulus value $s_i = \frac{i-1}{n-1}$ is presented. To go even further, we study the case of a linear relationship between stimulus value and firing rate. Given a baseline firing rate f_0 and a maximum gain f_Δ, the neuron will emit $f_0 + \frac{i-1}{n-1} f_\Delta$ spikes on the average whenever stimulus s_i is presented.

The larger the stimulus set and the higher the frequency of stimulus change, the harder it becomes to obtain reliable firing rate estimates. At some point, assumptions about the smoothness of the neuron's response have to be made, and the random variable for the neural signal might have to take firing rate history into account. Whether the smoothness constraint can be justified or not will certainly depend on the neural system under scrutiny. In the study of hand reaching, for example, it is a sound assumption that limb positions won't change in a non-continuous fashion. It has been shown that neurons in motor cortex might not depend linearly on variables as position or velocity but observed maps are always smooth and typically not highly non-linear [35, 36].

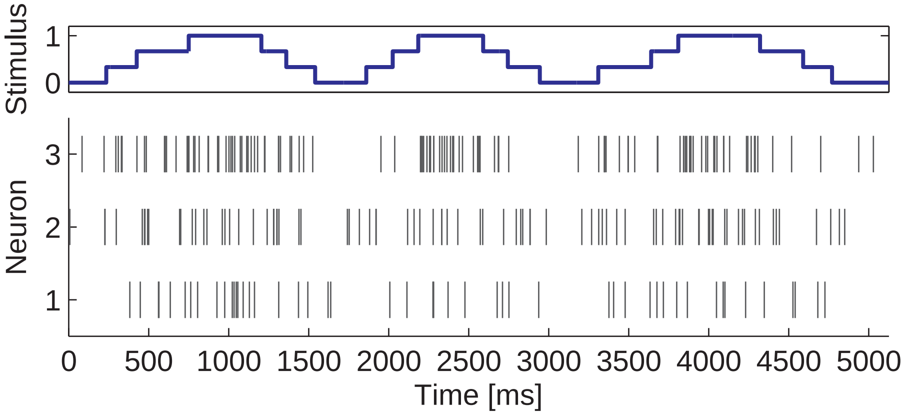


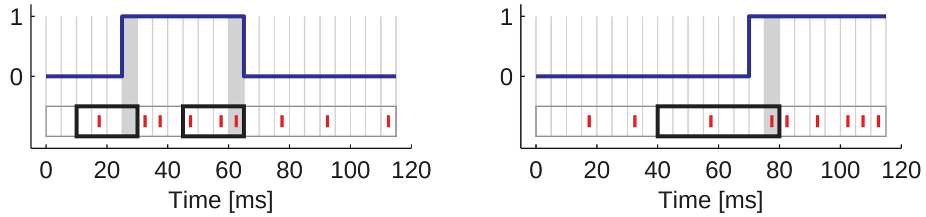
Figure 4: Raster plot of an example spike train generated from the n-valued stimulus with $\overline{t_0} = 500\text{ms}$ and $t_{0,\text{min}} = 250\text{ms}$. Baseline firing rates and maximum gain (f_0, f_Δ) for the three neurons were as follows: $(5\text{Hz}, 10\text{Hz})$; $(5\text{Hz}, 25\text{Hz})$ and $(5\text{Hz}, 40\text{Hz})$.

For the purpose of simulation, the smoothness constraint translates to the presentation of stimulus values in the predefined order (s_1, s_2, ..., s_n, s_n, s_{n-1}, ..., s_1). If the temporal pattern was fixed too, then the stimulus would be highly predictable and large sliding windows would extract an artificially increased mutual information value. A method that was introduced in Section 4.5 will reduce the predictability by randomizing the length of time segments in which the stimulus does not change. Stimulus change events are modeled as a renewal process. On the average, each instance of a stimulus s_i will be presented for the duration $\overline{t_0}$, but never for a shorter period than $t_{0,\text{min}}$. Figure 4 depicts an example stimulus and three spike trains. For sufficiently long data segments, all stimulus values are equiprobable and the entropy of the n-valued stimulus is $\log_2 n$ bits.

5 Results on synthetic data

5.1 Effects of window sizes in binning approaches on information measures

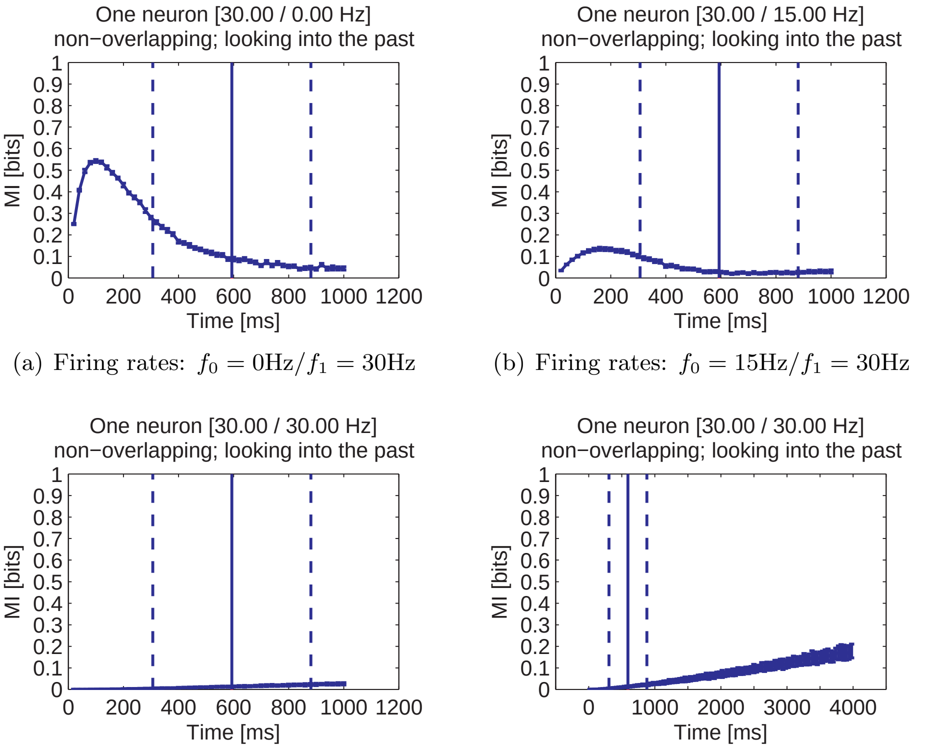
This chapter aims at studying the effects of different window sizes on the calculation of mutual information. In most experimental settings, it is impossible to derive appropriate bin and window sizes from theoretical principles. By contrast, we are in the exquisite situation of being able to control stimulus properties and neural responses. We put this to use and systematically analyze the effect of various parameters on mutual information. As we are able to generate experimental data in arbitrary quantities and use unbiased estimators, all mutual information values have been obtained via fully converged estimators and can be accepted as “ground truth”. Furthermore, we attempt to establish bounds on the amount of data required for reliable estimates depending on the estimator used.



(a) The first and last window that is associated with this instance of the $(S = 1)$ condition. Window size w is $20ms$ (4 bins).

(b) Contamination of a convolution window of length $40ms$ (8 bins) that is counted as a joint event with the $S = 1$ condition but has more overlap with the $S = 0$ condition.

Figure 5: Illustration of convolution windows used to obtain instantaneous firing rate estimates. Bin size of $5ms$, stimulus is drawn in blue, spikes in red and convolution windows in black.



(a) Firing rates: $f_0 = 0\text{Hz}/f_1 = 30\text{Hz}$

(b) Firing rates: $f_0 = 15\text{Hz}/f_1 = 30\text{Hz}$

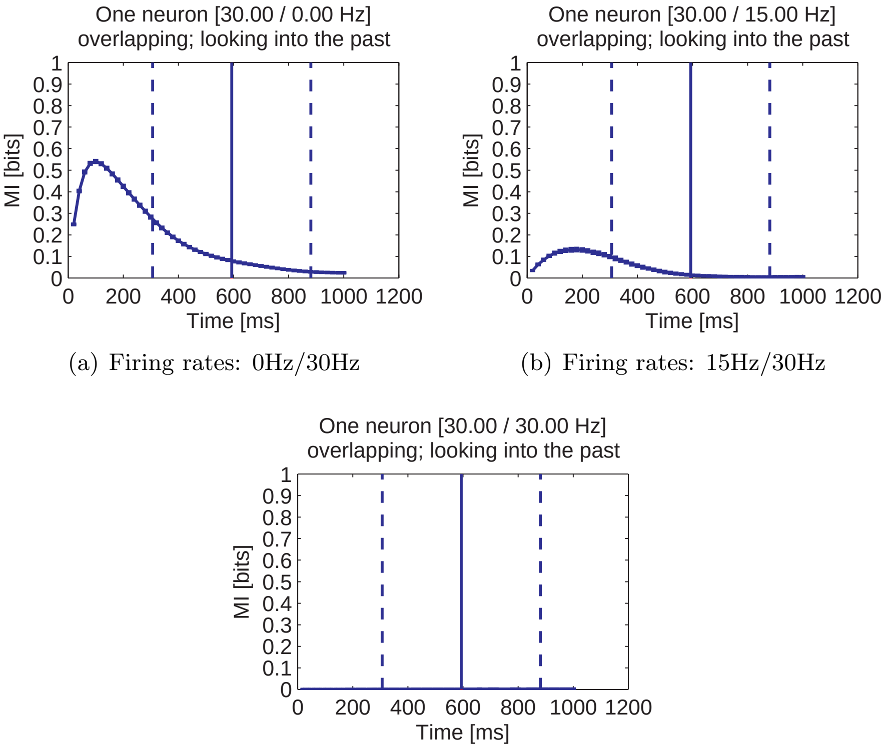
(c) Firing rates: $f_0 = 30\text{Hz}/f_1 = 30\text{Hz}$

(d) Firing rates: $f_0 = 30\text{Hz}/f_1 = 30\text{Hz}$ for a larger range of window sizes. The neuron is not sensitive to the stimulus but insufficient data causes a strong upward bias in the plug-in estimator

Figure 6: Mutual information of firing rate and stimulus. Firing rate was calculated from a window that only considered spikes in the past and was shifted by its own length (no overlap). Dashed lines indicate standard deviation $\sigma_0 = 300ms$ around $\overline{t_0} = 600ms$.

In a first experiment, we generated spike trains from the binary stimulus described in Section 4.5 and varied the neuron firing rates under stimulus conditions $S\!=\!0$ and $S\!=\!1$. Parameters were fixed to $P(S\!=\!0) = P(S\!=\!1) = 0.5$; $t_{0,\min} = 300ms$ and $\overline{t_0} = 600ms$. Three exemplary plots are shown in Figure 6, where the random variable of the neural signal was defined as the spike count in non-overlapping windows of varying lengths. The temporal reference point of the sliding window was at its end such that the stimulus at time t and the spike count in the time frame $[(t-w);t]$ were counted as joint events (illustrated in Figure 5(a)). To remove artifacts that are specific to certain generated spike trains, we repeated all experiments 50 times and plot mean curves. Some plots also feature error bars (e.g., Figures 6 and 7) but they are typically small and reduce to a jagged contour around the curve.

A first cursory inspection of Figure 6 reveals that neither very short nor very long windows convey large amounts of information about the stimulus. The first observation is attributable to the lack of contrast between different stimulus conditions in the case of short windows. Recall that the stimulus is guaranteed to remain constant for a duration of $t_{0,\min}$ in $S = 0$ segments and $t_{1,\min}$ in sections where the stimulus takes value 1. As $P(S{=}0) = P(S{=}1) = 0.5$, it holds that $t_{0,\min} = t_{1,\min}$. To make use of this property most efficiently, one would intuitively choose the window length w approximately equal to $t_{0,\min} = 300ms$. That choice should prove optimal in the sense of maximizing the number of spikes that go into the firing rate estimate while providing decent coverage of the average stimulus presentation interval. However, contaminations of the convolution window with spikes that occurred during the other stimulus condition (cf. Figure 5(b)) undermine the hypothesis and ideal window lengths are found to be in the range from $50ms$ to $200ms$ for the given parameters. In Figure 6, graphs for different amounts of firing rate contrast $f_\Delta = f_1 - f_0$ between stimulus conditions are shown.



(a) Firing rates: 0Hz/30Hz

(b) Firing rates: 15Hz/30Hz

(c) Firing rates: 30Hz/30Hz

Figure 7: Mutual information between firing rate and stimulus. Firing rate was calculated from a window that only considered spikes in the past and was shifted by $5ms$. Error bars represent one standard deviation (only visible as jagged contours around the curve).

A worrisome artifact is exhibited in Figure 6(c) and, more drastically, in Figure 6(d). Even though the neuron does not respond to the stimulus and has a firing rate of 30 Hz under both conditions, the plug-in estimator extracts an increasing amount of information for larger windows. It is necessary to clarify one detail of the experimental conditions at this point. The hypothetical experimental session in this experiment lasted for $1000s$, independent from the window size. One consequence is that the amount of data available is inversely proportional to the window size. There are 10000 observations for a window of $100ms$ but only 250 observations for a four second window. Figures 6(c) and 6(d) do not only show effects of different window sizes but also artifacts attributable to insufficient amounts of data.

One side effect of using overlapping windows is that the number of observations is independent from the chosen window size[4]. This makes the analysis of the relation between window size and estimated mutual information possible, while limited data effects are suppressed. Results for this scenario are shown in Figure 7. Now the estimation of mutual information for the stimulus insensitive neuron in Figure 7(c) behaves as expected.

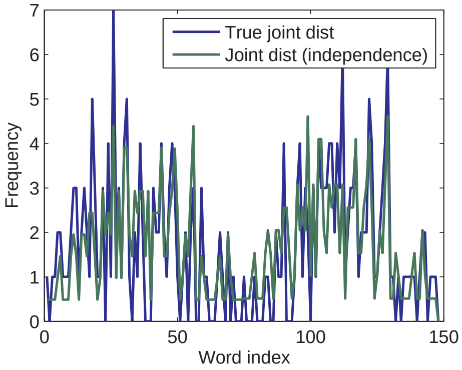


Figure 8: Joint distribution of (S, N) sorted in lexicographical order and hypothetical joint distribution under the assumption of independence for $1000s$ of data (250 observations for a $4s$ window).

On the other hand, it is instructive to illuminate reasons for the overestimation of mutual information when insufficient data is available. In the extreme case of $w = 4s$, the number of observations drops to 250. As the sliding window accumulates spikes, and hence discrete events, the distribution

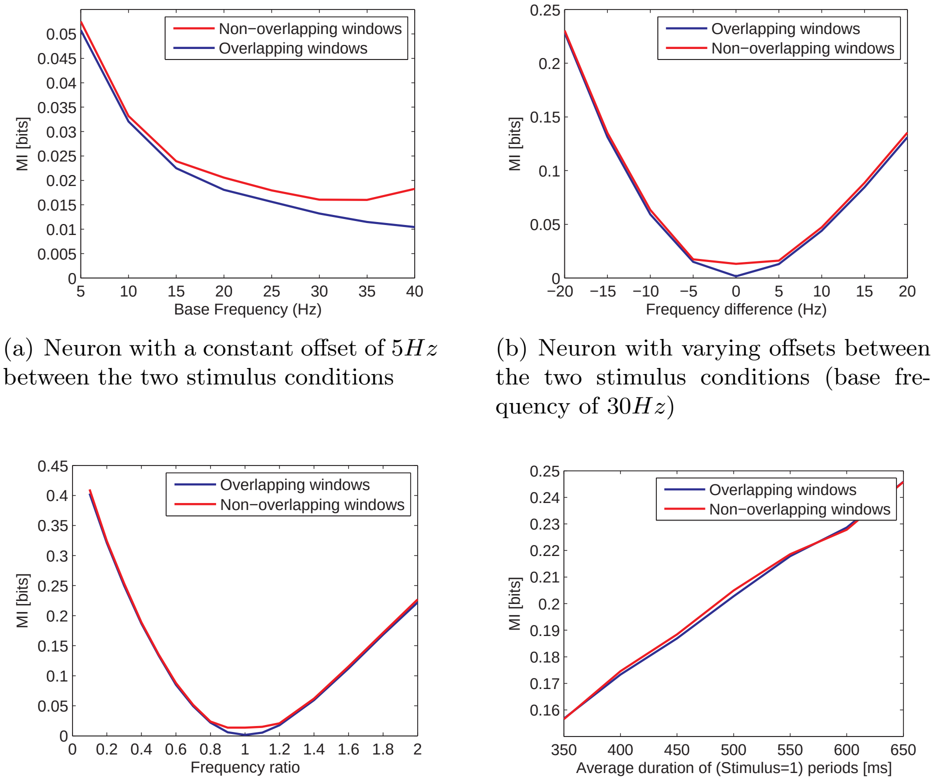
[4] With the marginal exception of border effects at the beginning and end of a data set.

over the instantaneous firing rate derived from the window will be discrete. In Figure 8, the relative frequencies of all observed pairs (s, n) sorted in lexicographical order are depicted. In addition, the distribution under the assumption of independence is plotted. Note that mutual information is a distance measure between the two distributions. There are many values n such that $(1, n)$ can be observed with non-zero frequency, but $(0, n)$ is not present at all, or vice versa. If we assume that $(1, n)$ was observed once and $(0, n)$ never, then $P\left((S, N) = (1, n)\right) = \frac{1}{250}$ and $P\left((S, N) = (0, n)\right) = 0$ in the observed joint distribution. Given that $P(S = 0) = P(S = 1) = \frac{1}{2}$, it follows that $P_{\text{indep}}(S = 0) = P_{\text{indep}}(S = 1) = \frac{1}{500}$ for the hypothetical distribution under the assumption of independence. Then each value n that fulfills the aforementioned condition will contribute

$$I_{\text{singleton}} = \frac{1}{250} \cdot \log_2 \left(\frac{\frac{1}{250}}{\frac{1}{2} \cdot \frac{1}{250}} \right) = \frac{1}{250} \cdot \log_2 2 = \frac{1}{250} \text{bits} \qquad (5.1)$$

to the estimated mutual information.

A straightforward conclusion from the comparison of overlapping and non-overlapping windows is that the former are always preferable. To provide more evidence for the advantages of this strategy, the differences between estimated mutual information values in the overlapping and the non-overlapping case are plotted in Figure 9. The comparison is based on the extracted mutual information peak value, i.e., the global maximum over all window sizes within a reasonable range. In Figures 9(a), 9(b) and 9(c), neuronal properties were varied. By contrast, stimulus properties were varied in Figure 9(d). In all cases, the mutual information estimate does not depend on the window shifting method, i.e., shifting by a fixed value of $5ms$ or the full window size to avoid any overlap.



(a) Neuron with a constant offset of $5Hz$ between the two stimulus conditions

(b) Neuron with varying offsets between the two stimulus conditions (base frequency of $30Hz$)

(c) Neuron with varying firing ratios between the two stimulus conditions (base frequency of $30Hz$)

(d) Neuron with firing rates of $[10Hz/30Hz]$ for different values of $\overline{t_0}$, $t_{0,\min}$ is kept constant at $300ms$.

Figure 9: Comparison of mutual information peak values obtained via the plug-in estimator for non-overlapping and overlapping (by $5ms$) sliding windows.

5.2 Effects of neuronal properties

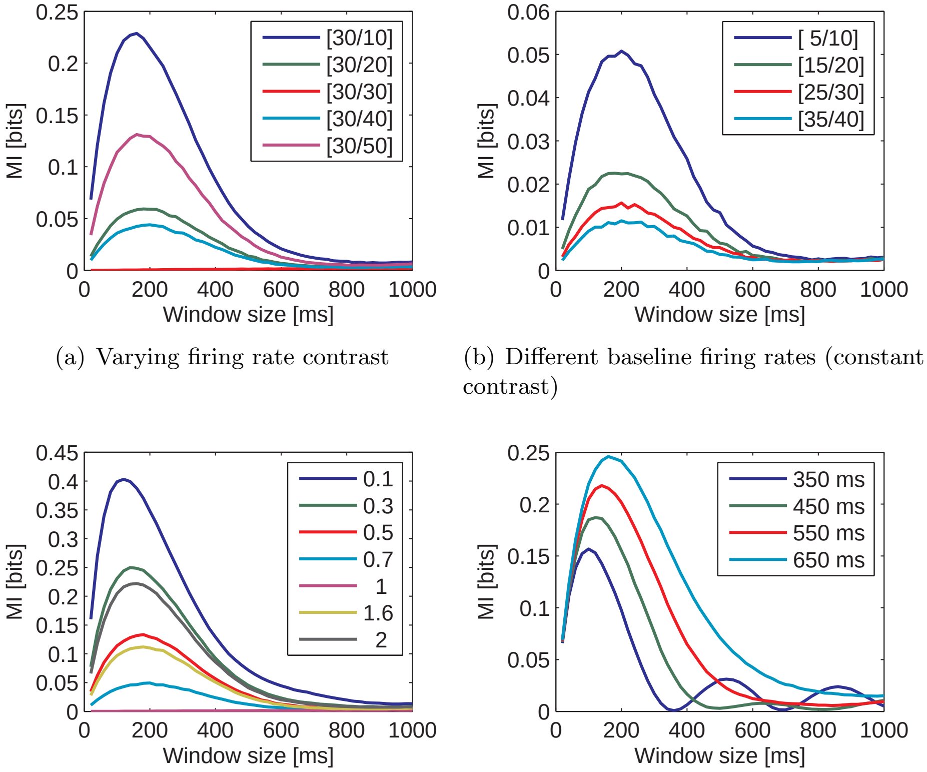
For consequent experiments, we exclusively used overlapping windows and fixed the window shift to $5ms$. In Figure 9, the effect of neuronal properties on mutual information is depicted. For Figure 9(a), the neuron's firing rate under the stimulus condition $S=1$ was set to $f_1 = f_0 + 5$Hz and the baseline firing rate f_0 varied. Obviously, information content does not only depend on $\Delta f = f_1 - f_0$, but also on f_0, as mutual information drops for increasing baseline firing rates. One possible interpretation of this phenomenon implicates the relative contrast between f_1 and f_0, i.e., $\frac{f_1}{f_0}$, as the relevant quantity for information transfer. Mutual information of stimulus and neural signal for $0 < \frac{f_1}{f_0} \leq 2$ is depicted in Figure 9(c). The shape of the curve implies a roughly quadratic dependence of $I(N;S)$ on $\frac{f_1}{f_0}$, but there is an evident asymmetry, as $I(N;S)$ grows less rapidly for increasing values of $\frac{f_1}{f_0}$ if $\frac{f_1}{f_0} > 0$ than in the $\frac{f_1}{f_0} < 0$ case. The same behavior is exhibited in Figure 9(b), where the x-axis is parameterized differently, namely as the absolute difference $\Delta f = f_1 - f_0$.

5.3 Effects of stimulus properties

Properties of the stimulus also have an impact on mutual information estimates, an effect that can be attributed to two major sources. When the presented stimuli change more rapidly, the presentation time of a single stimulus decreases and so does the amount of time that a neural system can spend encoding the stimulus. There are physiological constraints to neural spiking activity, among them the minimum refractory period of approximately one millisecond, the binary and discrete nature of action potential based information transfer, limited reliability of neurons and finite temporal precision in the form of jitter [30, 2]. Behavioral evidence from psychophysical experiments supports the inverse relationship of stimulus presentation time and subject performance in many different tasks, a very high-level indicator of information content in the neural signal.

The second influence on MI estimates is exerted by the analysis technique at hand. Even if a neural system has specialized in such a way as to transmit information about highly dynamic stimuli (such as the auditory cortex, for example, cf. [12]), devising methods to extract mutual information for dynamic stimuli is challenging. Estimates of instantaneous firing rates become less reliable because sliding windows will span across spikes emitted during different stimulus conditions. Reducing the window size, on the other hand, will make firing rate estimates more granular and decrease contrast between stimulus conditions. Figure 9(d) demonstrates how $I(N;S)$ increases as the frequency of stimulus change becomes smaller and, equivalently, as $\overline{t_0}$ grows.

More detailed plots are shown in Figure 10 where the dependence of $I(N;S)$ on neuronal and stimulus properties is illustrated for a few values in each model. It is interesting to note how the peak of the MI curve in Figure 10(d) shifts towards smaller window sizes as the stimulus becomes more dynamic, i.e., the value of $\overline{t_0}$ decreases. As explained above, the shift is to be expected because window overlap effects reduce the amount of information extracted if w is chosen too large. The dark blue curve for $\overline{t_0} = 350ms$ drops to zero in the range around $350ms$. For this parameter, the spike count in windows paired with a stimulus value $S=0$ will, on the average, be identical to the spike count in windows paired with a stimulus value $S=1$ due to window overlap. It follows that the neural signal N does not re-



(a) Varying firing rate contrast

(b) Different baseline firing rates (constant contrast)

(c) Varying firing rate ratios $\frac{f_1}{f_0}$

(d) Varying $\overline{t_0}$ values for constant $t_{0,\min} = 300ms$

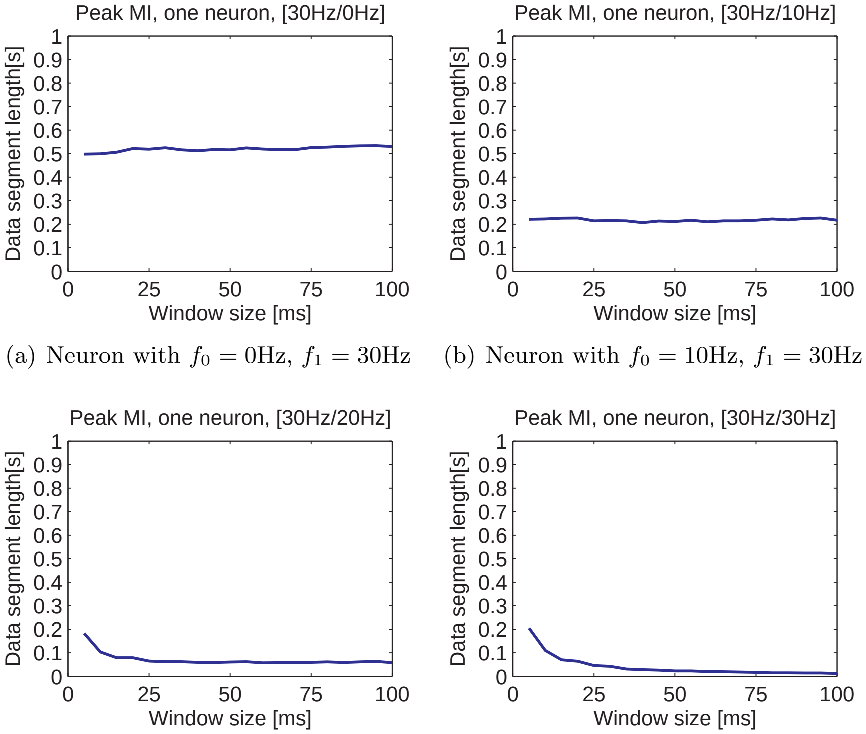
Figure 10: Effect of the window size on mutual information for different neuronal properties

duce uncertainty about the stimulus at all. But once w is increased beyond $350ms$, $I(S;N)$ begins to oscillate before it asymptotically approaches zero. The oscillation is more pronounced in the dark blue curve with $\overline{t_0} = 350ms$ than in any other curve shown in the plot. The length of segments in which the stimulus value is kept constant, is modelled probabilistically (refer to Section 4.5) and the standard deviation σ_0 of the segment lengths is given by Equation 4.2: $\sigma_0 = \overline{t_0} - t_{0,\min}$. It follows that σ_0 takes a small value of $50ms$ if $\overline{t_0} = 350ms$ and $t_{0,\min} = 300ms$. Due to the high regularity of the stimulus, another, albeit significantly lower, local maximum of mutual information can be observed for window sizes around $520ms$. The windows from which instantaneous firing rates are calculated span more than one stimulus presentation period in this case, but the ratio of spikes that were emitted during the $(S = 0)$ condition to those emitted during the $(S = 1)$ condition is still sufficient for some degree of discrimination. A similar oscillating behavior can be observed for larger values of $\overline{t_0}$ if the window size w is increased even beyond the range in Figure 10(d). The amplitude of secondary local maxima is smaller in relation to the amplitude of the primary local maximum because of the higher standard deviation σ_0, signifying the reduced stimulus regularity.

We can conclude from these observations that it is of high importance to match the window size for the estimation of instantaneous firing rates to the temporal characteristics of the stimulus. This point is less of a concern in highly controlled environments, e.g., when visual stimuli are presented for fixed amounts of time with adequate intertrial intervals. In more challenging experimental settings, when the sheer nature of the experiment precludes such an approach, more care has to be taken in choosing the optimal window size.

5.4 Effects of data set size

The fact that reliable estimates of mutual information depend on the availability of large amounts of data was mentioned in Section 3.5. Here we present examples how the upward bias in mutual information manifests itself in the case of the binary stimulus under different neuronal properties and amounts of data available.

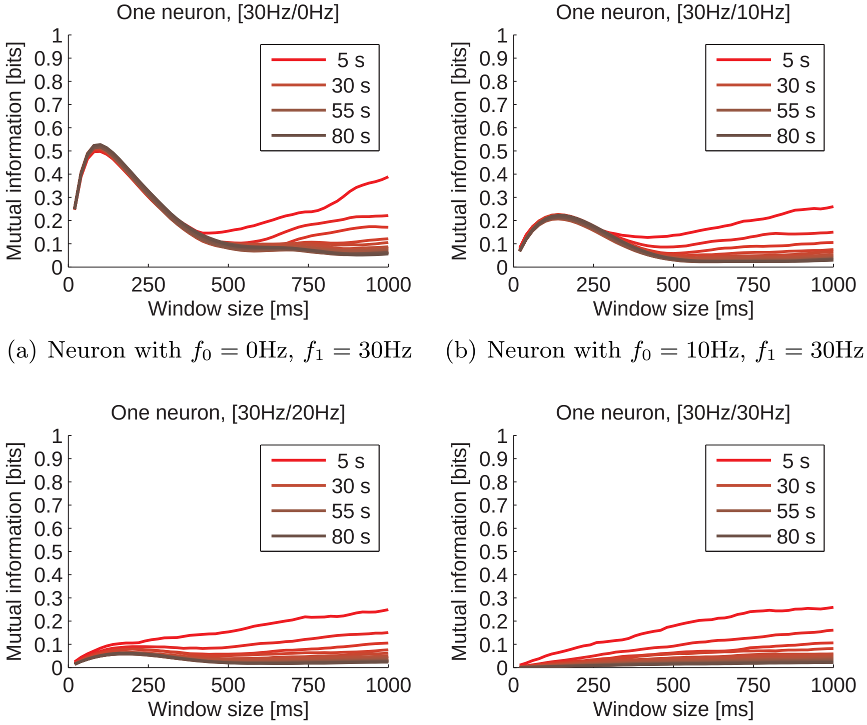


(a) Neuron with $f_0 = 0$Hz, $f_1 = 30$Hz (b) Neuron with $f_0 = 10$Hz, $f_1 = 30$Hz

(c) Neuron with $f_0 = 20$Hz, $f_1 = 30$Hz (d) Neuron with $f_0 = 30$Hz, $f_1 = 30$Hz

Figure 11: Different data set sizes for a binary stimulus ($300ms$, $600ms$), window shifted in $5ms$ intervals. Only MI peak value reported.

As before, spike trains were generated for $\overline{t_0} = 600ms$ and $t_{0,\min} = 300ms$. In order to obtain some kind of "ground truth", we estimated mutual information from a simulated experiment of about 20 minutes duration that yields $2 \cdot 10^5$ observations of the joint distribution $(N; S)$. To evaluate how the estimator behaves for smaller datasets, the experiment was artificially shortened in small steps, going down to a minimum of $5s$. Estimated mutual information values are plotted as a function of the dataset size in Figure 11.

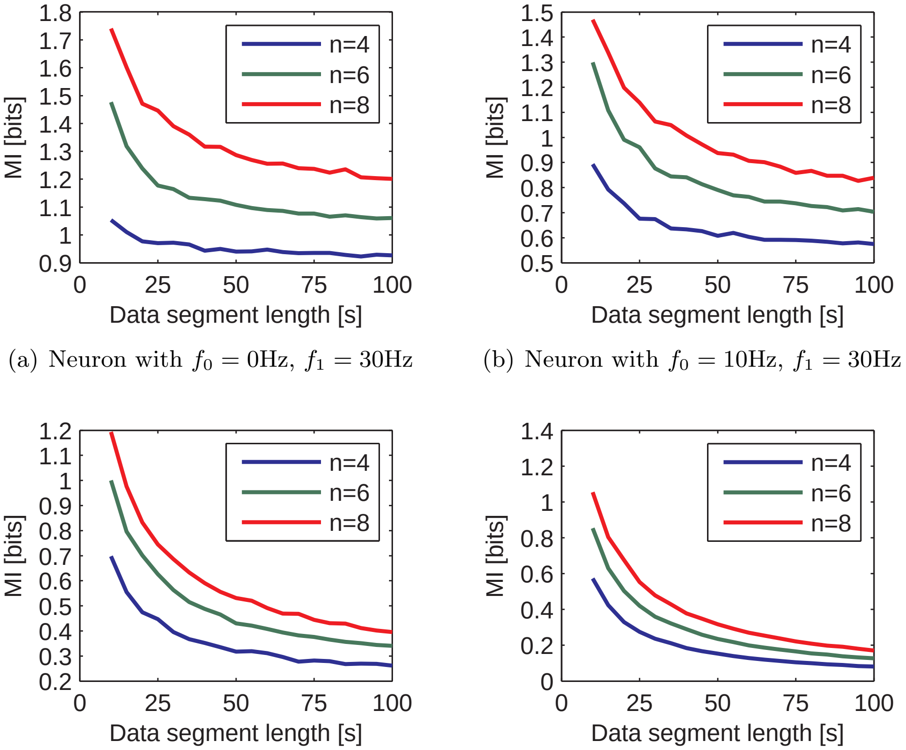


(a) Neuron with $f_0 = 0$Hz, $f_1 = 30$Hz (b) Neuron with $f_0 = 10$Hz, $f_1 = 30$Hz

(c) Neuron with $f_0 = 20$Hz, $f_1 = 30$Hz (d) Neuron with $f_0 = 30$Hz, $f_1 = 30$Hz

Figure 12: Different data set sizes for a binary stimulus $(300ms,\ 600ms)$, window shifted in $5ms$ intervals.

The previous arbitrary assumption that 20 minutes of data should suffice for an exact estimate appears to be justified and in most cases, the estimators have converged at session lengths of $80s$ to $100s$. Figure 11 also provides evidence that the extracted peak MI value remains quite stable even for extremely short sessions. This holds for experiments where the contrast Δf was sufficiently large, i.e., 20 Hz or 30 Hz. As discrimination between the two stimulus condition becomes harder at lower contrasts of



(a) Neuron with $f_0 = 0\text{Hz}$, $f_1 = 30\text{Hz}$

(b) Neuron with $f_0 = 10\text{Hz}$, $f_1 = 30\text{Hz}$

(c) Neuron with $f_0 = 20\text{Hz}$, $f_1 = 30\text{Hz}$

(d) Neuron with $f_0 = 30\text{Hz}$, $f_1 = 30\text{Hz}$

Figure 13: Different data set sizes for the n-valued stimulus ($n \in \{4, 6, 8\}$) with a presentation interval of $750ms$. Only MI peak value reported.

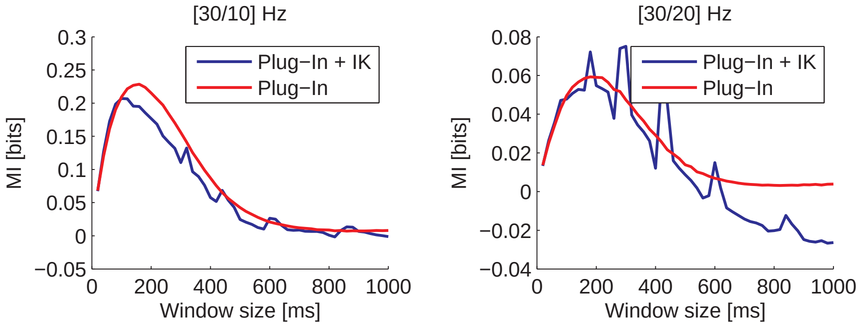
only 10 Hz or impossible at 0 Hz, insufficient data artifacts become more noticeable. In Figures 11(c) and 11(d), a strong upward bias affects the estimator if the number of observations drops below 5000 ($25s$ session) or 20000 ($100s$ session), respectively. The magnitude of this effect should not be underestimated, the bias reaches values of up to 0.2 bits for the stimulus insensitive neuron ($f_0 = 30$ Hz, $f_1 = 30$ Hz).

A more detailed dissection of the plug-in estimator's behavior is presented in Figure 12 that also plots the estimated MI values as a function of the window size. One surprising observation is the fact that estimates based on small window sizes are more resilient to small data sets than those based on larger window sizes. More specifically, mutual information peak values, typically extracted for $50ms \leq w \leq 200ms$, are extraordinarily stable. As w is increased well beyond the range that is relevant in practice for this particular experimental setup, a strong upward bias is evident.

The n-valued stimulus model is expected to aggravate upward bias effects relative to the binary stimulus. Even though the number of observations for the marginal distribution $p(N)$ of the neural signal is independent from the cardinality of the stimulus set, conditional distributions $p(N|S)$ are based on fewer samples if n is greater than 1. Specifically for the case of the n-valued stimulus, the number of observations of $p(N|S)$ is inversely proportional to n for each stimulus value. The expected behavior is well reflected by Figure 13. Especially in comparison with Figure 12, it becomes clear that the 100 second data segment is barely to sufficient for a convergence of the plug-in estimator. Even the neuron with a high firing rate contrast of 30Hz (Figure 13(a)) is severely afflicted by upward bias. The degenerate case in Figure 13(d) exhibits a bias of approximately 0.1 bits for the full $100s$ data segment even though the neuron is not responsive to the stimulus and has not converged to the true value of 0 bits.

5.5 Use of integer kernels

In Section 3.3, kernel density estimation was discussed as a viable approach to acquire estimates of the joint and marginal distributions required for the calculation of mutual information. It was also established that specialized kernels are necessary if the distributions to be estimated are discrete. We attempt to evaluate the adequacy of integer kernels (as defined in Equation (3.16)) for the estimation of mutual information within the framework of artificially generated data based on a binary stimulus (Section 4.5).



(a) $f_0 = 10\text{Hz}$, $f_1 = 30\text{Hz}$

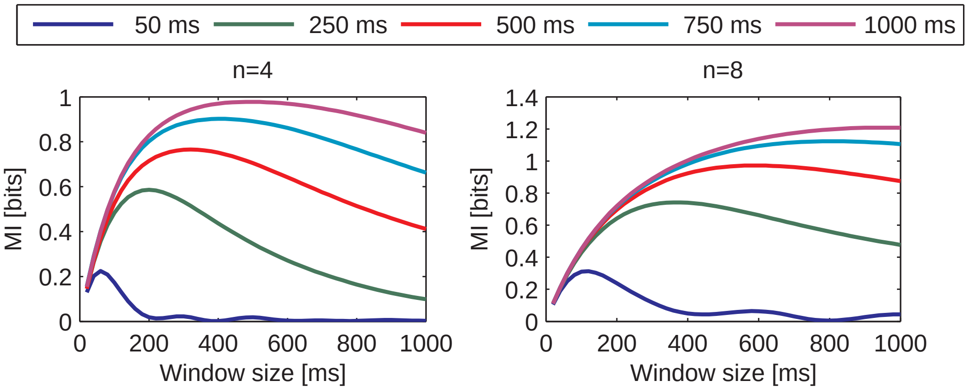
(b) $f_0 = 20\text{Hz}$, $f_1 = 30\text{Hz}$

Figure 14: Comparison of MI values for the plug-in estimator and the plug-in estimator after application of the integer kernel smoothing operator. Spike trains generated for the binary stimulus with $\overline{t_0} = 600ms$ and $t_{0,\min} = 300ms$.

The results of this experiment are shown in Figure 14. As described in Section 3.3, the kernel bandwidth h was chosen by a heuristic based on the standard deviation of the random variable. Unfortunately, several negative properties of the integer kernel density estimation method are observable in the plots. While the curve of the traditional plug-in estimate as a function of the window size is smooth, the curve for the integer kernel based estimator exhibits several spikes of large magnitude. As a result, MI values derived from IK estimates will suffer from a higher variance than plug-in estimates. If the spiky/jagged segments are excluded, both estimators behave roughly in the same way.

The use of integer kernels for the purpose of estimating information-theoretic quantities certainly deserves more attention than granted in this thesis. Though, within the limits of our model and data analysis framework, it appears doubtful whether integer kernel density estimates offer advantages over maximum likelihood estimates.

5.6 Results for the n-valued stimulus



(a) 4-valued stimulus

(b) 8-valued stimulus

Figure 15: Mutual information for n-valued stimuli plotted for different window lengths w and temporal characteristics of the stimulus.

The binary stimulus is a good tool to understand the dependence of mutual information estimates on neuronal properties, stimulus properties and the amount of data available. But few practical applications live in the microcosm of single neurons that are studied in the context of two stimuli.

In order to close the gap to continuous variables, such as quantities that describe limb kinematics, we introduced the n-valued stimulus in Section 4.6. Studying this type of stimulus is particularly interesting if the presentation protocol is highly dynamic in the temporal domain. The behavior of mutual information as a function of the window size is depicted in Figure 15. To emphasize the role of the temporal dynamics of a stimulus, the average presentation interval of the stimuli $\overline{t_0}$ was chosen from a set of values ranging from $50ms$ to $1000ms$. Likewise, the minimum presentation interval $t_{0,\min}$ was set to $\frac{\overline{t_0}}{2}$ so that the standard deviation of the presentation interval and its mean have a ratio of 1.

The general shape of the curves is quite similar to that observed for the binary stimulus (cf. Figure 10(d)). When stimuli are presented in less rapid succession, the window size that extracts the mutual information peak value increases. This is hardly surprising as larger windows allow for more reliable estimates of the instantaneous firing rates. In other words, the variance of the firing rate distribution conditioned on the stimulus decreases.

Recall that the entropy of the n-valued stimulus is $\log_2 n$ bits[5]. Even for relatively slow stimulus changes, the extracted peak MI value does not even come close to the entropy value. In fact, the increase in mutual information from the 4-valued stimulus (Figure 15(a)) to the 8-valued stimulus (Figure 15(b)) is small in comparison to the stimulus entropy's doubling. Another pointed difference pertains to the optimal window size in relation to the mean stimulus presentation interval $\overline{t_0}$. For $n = 4$, the window sizes

[5] This holds if stimulus values are equiprobable, a condition that is fulfilled by all discussed experiments.

that maximize mutual information are in the neighborhood of $\frac{\overline{t_0}}{2}$ as this value appears to optimize the reliability of instantaneous firing rate estimates without inducing too much contamination of the convolution window for the spike count (as illustrated in Figure 5). A drastic change occurs when n is increased to eight. Significantly larger windows now achieve the optimal trade-off between contamination and firing rate estimation. One can hypothesize that this shift is attributable to the decreased firing rate difference between "adjacent stimuli". In our example, the separation drops from 10Hz to $\frac{30}{7}$Hz $\approx$ 4.3Hz as we go from $n = 4$ to $n = 8$.

5.7 Results for multiple neurons

Since the technology for simultaneous recordings form multiple cells has become available, encoding of stimuli by populations of neurons has received a lot of attention. Mutual information readily adapts to the analysis of population data. A straight-forward choice for the random-variable is to concatenate the random variable for each neuron into one large vector. It follows that the random variable for the m-neuron case is m-dimensional. We picked a single configuration from the experiments performed on the n-valued stimulus with single neurons and examined the behavior of mutual information as the number of neurons is increased.

Obviously, the change of mutual information will depend on the neural properties in the population. In the name of simplicity, we confined ourselves to the simplest case in which all neurons are identical. The n-valued stimulus with parameters $n = 4$, an average presentation interval of $400ms$

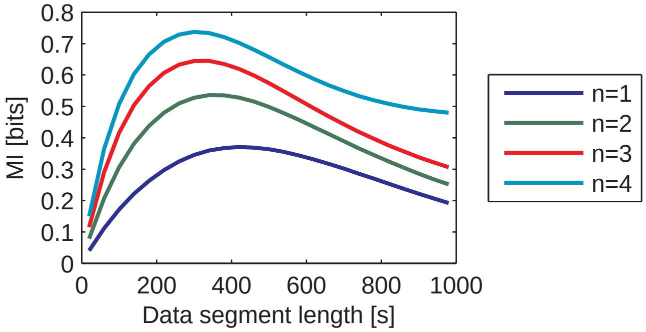


Figure 16: Behavior of mutual information as a function of the window size for up to four neurons. Model: 4-valued stimulus, average presentation interval of $400ms$, $f_0 = 10\text{Hz}$, $f_\Delta = 20\text{Hz}$

and $f_0 = 10\text{Hz}$, $f_\Delta = 20\text{Hz}$ was picked for this experiment. In preliminary experiments, it became clear that the amount of data has to be increased significantly in order to guarantee full convergence of the plug-in estimator.

In Figure 16, the results for $1 \leq m \leq 4$ neurons are depicted. As expected, mutual information increases monotonically as a function of the number of neurons. An interesting feature is the deceleration of information growth. Going from one to two neurons adds more than 0.15 bits, while the step from three to four neurons contributes less than 0.1 bits. It can also be observed that the optimum window size for the extraction of the mutual information peak value shifts markedly from $420ms$ for one neuron to $300ms$ for four units. This behavior is a clear indication for the obvious fact that firing rates can be estimated more easily from four identical neurons than from only one. Finally, a peculiar increase of mutual information for large window sizes around $1000ms$ is exhibited for bigger neural populations. This behavior is not understood yet and requires further investigation.

6 Conclusion

6.1 Summary

In the previous sections, we have introduced mutual information and justified its application in the field of neuroscience. Motivated by its widespread use in many studies, we moved on to present the most widely employed paradigms. Experiments performed on synthetic datasets provide evidence that great care must be taken in the choice of the random variable that represents the neural signal. Two different mechanisms for the estimation of instantaneous firing rates — overlapping and non-overlapping sliding windows — were compared and the results give rise to the conclusion that overlapping windows are less prone to artifacts if the number of observations is low, i.e., the dataset relatively small.

Taking advantage of the control over neural properties and stimulus characteristics offered by synthetically generated data, we examined the dependence of mutual information on several quantities. In the case of the binary stimulus, relative firing rate contrast $\frac{f_1}{f_0}$ and absolute firing rate contrast $\Delta f = f_1 - f_0$ appears to have the strongest influence on mutual information[6]. However, our results also indicate that mutual information is not independent from the baseline firing rate f_0 if identical contrast is assumed. In other words, mutual information $I(N; S)$ will have different values for a neuron with $(f_0 = 30\text{Hz}, f_1 = 40\text{Hz})$ than one with $(f_0 = 20\text{Hz}, f_1 = 30\text{Hz})$.

[6] As we fix the value of f_0 in our experiments, we can treat it as a constant and there is a linear relationship between $\frac{f_1}{f_0}$ and Δf, so both parameterizations are equivalent for any given $f_0 \neq 0$: $f_0 \left(\frac{f_1}{f_0}\right) - f_0 = f_1 - f_0 = \Delta f$.

If f_0 is kept constant and mutual information plotted as a function of the contrast f_Δ, the resulting curve is approximated well by a quadratic function for small values of f_Δ. As f_Δ increases further, the curve's slope decreases and curvature finally changes sign and becomes negative. For obvious reasons, this part of the curve is not modelled well by a quadratic function but can be fit well by an incomplete beta function. The aforementioned observation that neurons with $(f_0 = 30\text{Hz}, f_1 = 40\text{Hz})$ and $(f_0 = 20\text{Hz}, f_1 = 30\text{Hz})$ behave differently, is supported by the fact that slightly different parameters are necessary to model the negative part of the curve where $f_\Delta < 0$ versus the positive part where $f_\Delta \geq 0$.

One of the most significant contributions of this thesis is the analysis of the relationship between the temporal characteristics of the stimulus and the optimal window size that determines the width of the convolution window for the purpose of firing rate estimation. Surprisingly, optimal window sizes for the binary stimulus were found to be notedly smaller than $t_{0,\text{min}}$ and thus significantly below the expected value. Even more counterintuitive is the discovery that the relative magnitude of w and $t_{0,\text{min}}$ changes as we shift our attention to the n-valued stimulus model. If n is increased, the optimal window size w drifts to larger values. It is important to note that this phenomenon can only be observed is the stimulus fulfills certain assumptions concerning its smoothness and continuity.

In order to approximate continuous stimuli more closely, the n-valued stimulus model was studied under similar conditions as the binary model. We found that the gap between stimulus entropy and maximum mutual information increases significantly, notwithstanding favorable choices for the temporal characteristics of the stimulus (such as long presentation intervals) and high firing rate contrast for the neuron. Two different reasons can cause this result. Either the neuron's inherent properties (probabilistic spike generation, linear dependence on the stimulus value) forestall the efficient transmission of information about the stimulus or the toolchain to recover information from the spike train is lacking. In the case of the n-valued stimulus it seems as if the information output of a single neuron is not sufficient to allow for an observer of the spike train to discriminate between eight different stimuli.

We shortly evaluated mutual information in the context of neuron populations by replicating the same neuron several times. Not surprisingly, significant increases in mutual information can be observed as more neurons are added, even though the difference per neuron diminishes with each additional unit. Populations of multiple neurons constitute a more reliable system for information transfer such that notedly smaller window sizes permit the extraction of mutual information peak values. This phenomenon has to be attributed to the fact that shorter segments of the neural signal are sufficient to produce good estimates of firing rates. It is not perceivable whether the shift of the optimum window size would be of similar magnitude for more heterogeneous populations.

6.2 Outlook

Based on the results and conclusions in the previous section, we propose the following experiments to further evaluate the adequacy of mutual information as a measure of relevance in neural coding and analyze its behavior in a diverse set of environments:

Increasing the stimulus complexity: The n-valued stimulus is an approximation of continuous stimuli for sufficiently large n. Experiments were only performed in the range $n \leq 8$. Expanding the range to stimulus models that draw values from a larger discrete set in order to approximate a continuous stimulus via binning will show whether estimators for discrete distributions scale appropriately. Bigger sets of stimulus values reduce the number of observations in the distribution over the neural signal conditioned on a specific stimulus value. For practical applications, it is profoundly relevant to have knowledge of the amount of data required for reliable estimates of mutual information. Therefore, it could prove valuable to develop lower bounds on this quantity as a function of stimulus properties and characteristics of the neural signal.

Non-linear dependence of the firing rate on the stimulus: The simulator for the generation of synthetic spike trains assumed a linear relationship between the stimulus value and the firing rate of the neuron. To close the gap between artificially generated data and electrophysiological recordings, non-linear models should be implemented. Within the domain of the motor cortex, one could study hand motion in 2D or 3D and generate spike trains from hypothetical neurons that behave in accordance with spatiotemporal tunings function, as proposed by Paninski et al. [35].

Heuristics for the choice of window sizes: Our results in Section 5.1 have demonstrated the essential role of appropriately chosen window sizes for the estimation of instantaneous firing rates. The location of the mutual information peak value appears to vary as a function of several parameters, including temporal features of the stimulus and neural properties. A transformation of the stimulus and the neural signal into the frequency domain might lay the foundation for methods that can automatically choose optimal values for the window size that maximize mutual information.

Mutual Information Estimators: While there are several alternative ways to estimate mutual information from continuous distributions, e.g., nearest-neighbor methods and kernel density estimates, only the plug-in estimator is applicable to discrete distributions in general settings. Research could proceed in two directions to explore estimators that converge on smaller datasets. Firstly, nearest-neighbor methods could potentially be modified and adapted to overcome the ill-posedness of the nearest neighbor notion in discrete spaces. Secondly, discrete stimuli drawn from a large set can be interpreted as approximations of continuous stimuli. Whereas nearest-neighbor estimators still face the difficulties of integral distances between data points, kernel density estimates could outperform maximum likelihood estimates.

Study of multiple neurons: We have briefly peeked into the field of population analysis. Obviously, our experiments have to be repeated on a larger selection of datasets. Furthermore, the replication of identical neurons to simulate population behavior is a misrepresentation of population properties found in the brain. It is well known that many neural mechanisms rely on groups of neurons that span the stimulus space uniformly with overlapping tuning curves. A second, extremely interesting, field of research is related to the role of (temporal) correlations between neurons. Hopefully, a powerful measure of relevance can unite opposing views about their usefulness for information transmission.

References

[1] J. Aitchsion and C. G. G. Aitken. Multivariate binary discrimination by the kernel method. *Biometrika*, 63:413–420, 1976.

[2] W. Bair and C. Koch. Temporal precision of spike trains in extrastriate cortex of the behaving macaque monkey. *Neural Computation*, 8:1185–1202, 1996.

[3] R. T. Born and D. C. Bradley. Structure and function of visual area MT. *Annu Rev Neurosci*, 28:157–189, 2005.

[4] A. Borst and F. E. Theunissen. Information theory and neural coding. *Nat Neurosci*, 2(11):947–957, November 1999.

[5] J. Cheng, R. B. Stein, K. Jovanovic, K. Yoshida, D. J. Bennett, and Y. Han. Identification, localization, and modulation of neural networks for walking in the mudpuppy (necturus maculatus) spinal cord. *J Neurosci*, 18(11):4295–4304, 1998.

[6] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, Inc., 2006.

[7] A. d'Avella, P. Saltiel, and E. Bizzi. Combinations of muscle synergies in the construction of a natural motor behavior. *Nat Neurosci*, 6(3):300–308, 2003.

[8] J. De Bie. An afterimage vernier method for assessing the precision of eye movement monitors: Results for the scleral coil technique. *Vision Research*, 25(9):1341–1343, 1985.

[9] R. R. de Ruyter van Steveninck, G. D. Lewen, S. P. Strong, R. Koberle, and W. Bialek. Reproducibility and variability in neural spike trains. *Science*, 275:1805–1808, 1997.

[10] R. C. deCharms and A. Zador. Neural representation and the cortical code. *Annu. Rev. Neurosci.*, 23:613–647, 2000.

[11] R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. Wiley-Interscience Publication, 2nd edition, 2000.

[12] M. Elhilali, J. B. Firtz, D. J. Klein, J. Z. Simon, and S. A. Shamma. Dynamics of precise spike timing in primary auditory cortex. *J Neurosci*, 24(5):1159–1172, 2004.

[13] D. J. Felleman and J. H. Kaas. Receptive-field properties of neurons in middle temporal visual area (mt) of owl monkeys. *J Neurophys*, 52(3):488–513, 1984.

[14] E. E. Fetz. Temporal coding in neural populations? *Science*, 12:1901–1902, 1997.

[15] D. D. Fitts. *Principles of Quantum Mechanics: As Applied to Chemistry and Chemical Physics*. Cambride University Press, August 1999.

[16] M. S. Gazzaniga, R. B. Ivry, and G. R. Mangun. *Cognitive Neuroscience*. W. W. Norton & Company, 2nd edition, 2002.

[17] A. P. Georgopoulos, J. F. Kalaska, R. Caminiti, and J. T. Massey. On the relations between the direction of two-dimensional arm movements

and cell discharge in primate motor cortex. *J Neurosci*, 2(11):1527–1537, November 1982.

[18] A. P. Georgopoulos, A. B. Schwartz, and R. E. Kettner. Neuronal population coding of movement direction. *Science*, 233(4771):1416–1419, September 1986.

[19] L. R. Hochberg, M. D. Serruya, G. M. Friehs, J. A. Mukand, M. Saleh, A. H. Caplan, A. Branner, D. Chen, R. D. Penn, and J. P. Donoghue. Neuronal ensemble control of prosthetic devices by a human with tetraplegia. *Nature*, 442(7099):164–171, 2006.

[20] D. H. Hubel and T. N. Wiesel. Functional architecture of macaque monkey visual cortex. *Proc R. Soc. Lond. B*, 198:1–59, 1977.

[21] A. T. Ihler, J. W. Fisher III, and A. S. Willsky. Nonparametric hypothesis tests for statistical dependency. *IEEE Transactions on Signal Processing*, 52(8), August 2004.

[22] D. H. Johnson. Point process models of single-neuron discharges. *J Comp Neurosci*, 3(4):275–299, 1996.

[23] S. Kakei, D. S. Hoffman, and P. L. Strick. Muscle and movement representations in the primary motor cortex. *Science*, 285(5436):2136–2139, Sep 1999.

[24] E. R. Kandel. *Principles of Neural Science*. McGraw-Hill Education, June 2000.

[25] O. Kiehn and O. Kjaerulff. Distribution of central pattern generators for rhythmic motor outputs in the spinal cord of limbed vertebrates. *Ann N Y Acad Sci.*, 860:110–129, 1998.

[26] L. F. Kozachenko and N. N. Leonenko. Sample estimate of entropy of a random vector. *Probl. Inf. Transm.*, 23:95–101, 1987.

[27] A. Kraskov, H. Stögbauer, and P. Grassberger. Estimating mutual information. *Phys. Rev. E*, 69:066138, 2004.

[28] D. Lee, N. L. Port, W. Kruse, and A. P. Georgopoulos. Variability and correlated noise in the discharge of neurons in motor and parietal areas of the primate cortex. *J Neurosci*, 18(3):1161–1170, February 1998.

[29] M. Lotze, P. Montoya, M. Erb, and E. Hülsmann. Activation of cortical and cerebellar motor areas during executed and imagined hand movements: An fmri study. *Journal of Cognitive Neuroscience*, 11:5:491–501, 1999.

[30] P. R. Marsalek, C. Koch, and J. Maunsell. On the relationship between synaptic input and spike output jitter in individual neurons. *PNAS*, 94:735–740, 1994.

[31] E. M. Maynard, N. G. Hatsopoulos, C. L. Ojakangas, B. D. Acuna, J. N. Sanes, R. A. Normann, and J. P. Donoghue. Neuronal interactions improve cortical population coding of movement direction. *J Neurosci*, 19(18):8083–8093, 1999.

[32] W. T. Newsome. Deciding about motion: linking perception to action. *J Comp Physiol A*, 181:5–12, 1997.

[33] W. T. Newsome, K. H. Britten, and J. A. Movshon. Neuronal correlates of a perceptual decision. *Nature*, 341:52–54, 1989.

[34] L. Paninski. Estimation of entropy and mutual information. *Neural Comp.*, 15(6):1191–1253, June 2003.

[35] L. Paninski, M. R. Fellows, N. G. Hatsopoulos, and J. P. Donoghue. Spatiotemporal Tuning of Motor Cortical Neurons for Hand Position and Velocity. *J Neurophysiol*, 91:515–532, 2004.

[36] L. Paninski, S. Shoham, M. R. Fellows, N. G. Hatsopoulos, and J. P. Donoghue. Superlinear population encoding of dynamic hand trajectory in primary motor cortex. *J Neurosci*, 24(39):8551–8561, Sep 2004.

[37] E. Parzen. On the estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1065–1076, 1962.

[38] W. Penfield and T. Rasmussen. *The Cerebral Cortex of Man.* Macmillan, New York, 1950.

[39] R. Romo and E. Salinas. Flutter discrimination: Neural codes, perception, memory and decision making. *Nat Rev Neurosci*, 4:203–218, 2003.

[40] C. J. Rozell and D. H. Johnson. Examining methods for estimating mutual information in spiking neural systems. *Neurocomputing*, 65-66:429–434, 2005.

[41] E. Salinas and L. Abbott. Vector reconstruction from firing rates. *J Comp Neurosci*, 1(1–2):89–107, 1994.

[42] J. N. Sanes and J. P. Donoghue. Plasticity and primary motor cortex. *Annual Review of Neuroscience*, 23:393–415, 2000.

[43] A. B. Schwartz. Motor cortical activity during drawing movements: Population representation during sinusoid tracing. *J Neurophys*, 70(1):28–36, 1993.

[44] A. B. Schwartz, R. E. Kettner, and A. P. Georgopoulos. Primate motor cortex and free arm movements to visual targets in three-dimensional space. i. relations between single cell discharge and direction of movement. *J Neurosci*, 8(8):2913–2927, August 1988.

[45] S. H. Scott and J. F. Kalaska. Reaching movements with similar hand paths but different arm orientations. I. Activity of individual cells in motor cortex. *J Neurophysiol*, 77(2):826–852, Feb 1997.

[46] L. E. Sergio and J. F. Kalaska. Changes in the temporal pattern of primary motor cortex activity in a directional isometric force versus limb movement task. *J Neurophysiol*, 80(3):1577–1583, Sep 1998.

[47] M. N. Shadlen and W. T. Newsome. The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. *J Neurosci*, 18:3870–3896, 1998.

[48] M. N. Shadlen and W. T. Newsome. Neural basis of a perceptual deci-

sion in the parietal cortex (area lip) of the rhesus monkey. *J Neurophys*, 86:1916–1936, 2001.

[49] G. Shakhnarovich. Statistical data cloning for machine learning. Master's thesis, Technion, 2001.

[50] C. E. Shannon. *A Mathematical Theory of Communication*. CSLI Publications, 1948.

[51] S. M. Smirnakis, M. J. Berry, D. K. Warland, W. Bialek, and M. Meister. Adaptation of retinal processing to image contrast and spatial scale. *Nature*, 386(6620):69–73, March 1997.

[52] P. S. Stein, M. L. McCullough, and S. N. Currie. Spinal motor patterns in the turtle. *Ann N Y Acad Sci.*, 860:142–154, 1998.

[53] S. P. Strong, R. Koberle, R. R. de Ruyter van Steveninck, and W. Bialek. Entropy and information in neural spike trains. *Phys. Rev. Lett.*, 80(1):197–200, 1998.

[54] R. B. H. Tootell, A. M. Dale, M. I. Sereno, and R. Malach. New images from human visual cortex. *Trends in Neuroscience*, 19(11):481–489, 1996.

[55] D. Y. Tsao, W. A. Freiwald, R. B. Tootell, and M. S. Livingstone. A cortical region consisting entirely of face-selective cells. *Science*, 311(5761):670–674, 2006.

[56] D. van der Geest. Recording eye movements with video-oculography

and scleral search coils: a direct comparison of two methods. *J Neurosci Methods*, 114(2):185–195, 2002.

[57] J. D. Victor. Binless strategies for estimation of information from neural data. *Phys. Rev. E*, 66(5):051903, Nov 2002.

[58] S. P. Wise and J. Tanji. Neuronal responses in sensorimotor cortex to ramp displacements and maintained positions imposed on hindlimb of the unanesthetized monkey. *J Neurophysiol*, 45(3):482–500, Mar 1981.