RICE UNIVERSITY

Performance Limits of Brain Machine Interfaces

by

Ilan N. Goodman

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE

Doctor of Philosophy

APPROVED, THESIS COMMITTEE:

Don H. Johnson, Chair
J.S. Abercrombie Professor Emeritus of Electrical and Computer Engineering

Ashutosh Sambarwal
Assistant Professor of Electrical and Computer Engineering

Steven J. Cox
Professor of Computational and Applied Mathematics

Houston, Texas
April, 2010
Abstract

Performance Limits of Brain Machine Interfaces

by

Ilan N. Goodman

Understanding the constraints governing information transfer between electrodes and neurons is crucial to the effective design of neural prostheses. In sensory prostheses such as cochlear implants, information is transferred to the brain by stimulating neurons to produce sensation. In motor prostheses such as cortically controlled bionic limbs, neural recordings are processed to extract information needed to control a computer or mechanical device. In each case, performance of the prosthesis hinges on how effectively information can be conveyed to or from the device at the interface between brain and machine.

In this thesis, we investigate the performance capabilities and constraints of brain machine interfaces (BMIs) using an information theoretic approach. Modeling the BMI as a vector Poisson process channel, we compute the information capacity of several different types of BMI channels. Since capacity defines the ultimate fidelity limits of information transmission by any system, this approach gives us an objective way of evaluating and comparing different types of BMIs by determining the best possible performance of each system given its unique constraints. For stimulation BMIs, we examine how the capacity
of the system scales with the number of inputs, the constraints on the inputs, and inter-neuronal dependencies. For control BMIs, we quantify the loss in performance that results from using extracellular recordings, where signals from multiple neurons are received on a single electrode. This performance loss can be mitigated through spike sorting, and we show how the properties of the spike sorting algorithm have direct consequences for the resulting BMI capacity. We also provide extensions to the basic models to account for signal attenuation, cross-talk, and measurement noise.

Finally, we discuss the real-world significance of BMI capacity in the context of Rate-Distortion Theory, and interpret the capacity results using performance criteria that are relevant to BMIs. This framework provides a direct way to compare competing systems, and allows us to make predictions about the specific conditions necessary for a BMI to achieve a desired performance level.
Acknowledgments

To my family and friends, who encouraged me always; and to my advisor, who taught me never to give up (and a million other things).
3 The Neural Stimulation Channel 33

3.1 Extracelluar Stimulation 35
  3.1.1 Independent Neurons 35
  3.1.2 Dependent Neurons 38
  3.1.3 Z Channel Equivalence 41

3.2 Direct Stimulation 44
  3.2.1 Independent Neurons 44
  3.2.2 Dependent Neurons 45

3.3 The General Model 50

4 The Neural Control Channel 55

4.1 Extracelluar Recording 56
  4.1.1 Single Electrode 56
  4.1.2 Multiple Electrodes 61

4.2 Spike Sorting 62

4.3 The General Model 68

5 From Theory to Practice 71

5.1 Minimum Distortion: The Optimal Performance of BMIs 72
  5.1.1 Example 1: Auditory Prosthesis 73
  5.1.2 Example 2: Visual Prosthesis 77
5.1.3 Example 3: Brain-Computer Interface ........................................ 78

5.1.4 Example 4: Prosthetic Limb ..................................................... 79

5.2 Limitations of the Theory ......................................................... 82

6 Conclusions ...................................... 85

Bibliography ....................................... 88

A Infinitely divisible vector Poisson processes ................................ 98

B Cumulant correlation coefficients .............................................. 104

C Convergence of the vector Bernoulli process to the infinitely divi-

visible vector Poisson process .................................................. 108
## Illustrations

1.1 General system models of BMIs ........................................... 2

2.1 The classic information theory model ................................. 8
2.2 Capacity and rate-distortion ............................................. 15
2.3 The point process channel ................................................ 18

3.1 The stimulation BMI channel ............................................ 34
3.2 Capacity of the common input stimulation channel ............. 40
3.3 The Z channel model of extracellular stimulation ............... 43
3.4 Capacity of the stimulation channel with separate inputs .... 46
3.5 Scalar version of the intracellular stimulation channel ....... 48
3.6 Capacity of the intracellular stimulation channel with correlated inputs ... 49
3.7 The stimulation channel with signal attenuation ................. 51

4.1 The single electrode recording channel ............................. 57
4.2 Scalar version of the single electrode recording channel .... 58
4.3 Capacity of the single electrode recording channel ............. 60
4.4 The multi-electrode recording channel ............................... 61
4.5 The spike sorting channel ............................................... 63
4.6 Spike sorting with mislabeled and deleted spikes ............... 65
4.7 Spike sorting with inserted spikes .................................... 67
4.8 The general control BMI channel ............................................. 69
4.9 Capacity of the noisy recording channel ................................. 70
5.1 OPTA curves for an auditory prosthesis model ....................... 75
5.2 OPTA curves for a visual prosthesis model ............................. 78
5.3 OPTA curves for a prosthetic limb model .............................. 81
Chapter 1

Introduction

Advanced techniques for electrode implantation, neural recording, and signal processing have enabled a suite of new applications for brain-machine interfaces (BMIs). These devices generally fall into two categories. The first type, which we call the stimulation BMI, consists of one or more electrodes that stimulate a population of neurons. Applications of stimulation BMIs include sensory prostheses such as cochlear implants, which have been used successfully to restore hearing in deaf patients for decades [46], and retinal implants, which show promise for restoring sight in blind patients [52]. In addition, stimulation BMIs have been used to treat or alleviate symptoms from a variety of conditions such as Parkinson's disease, chronic pain, Tourette's syndrome, and clinical depression [37]. Here, rather than supplying sensory information, the BMI is used to replicate or facilitate some other neurological function.

The second type of device, which we call the control BMI, consists of one or more electrodes that record multiple neural signals in real-time. Prostheses involving control BMIs include motor prostheses such as bionic limbs, and brain-computer interfaces that have restored communication to patients with neurodegenerative diseases [46]. Larger-scale brain activity, such as electroencephalographic (EEG) signals recorded by electrodes distributed over the scalp, have also been used successfully by paralysis patients to control
Figure 1.1 : System model of the BMI. (a) A sensory stimulus $S$ is transmitted through a stimulation BMI to be decoded elsewhere in the brain. (b) The brain transmits an intended action through the control BMI where it is decoded as the control signal $\hat{S}$. $X$ and $Y$ are inputs and outputs (respectively) to the neural population channel.

Currently, no general consensus has emerged as to the “best” approach to designing BMIs, with researchers devising a variety of novel techniques with varying degrees of success. However, though many existing BMIs may perform qualitatively well, evaluating their performance objectively requires knowledge of the ultimate performance benchmark. In other words, we need to know how well the operational system works compared to the optimal ideal, given constraints such as the number and placement of electrodes, and the quality of pre- and post-processing algorithms.

Figure 1.1 depicts our two basic models for BMIs. In stimulation BMIs, the device is the encoder, transforming sensory stimuli (sound, light, etc.) into signals that can be transmitted and decoded by cortical neurons. In control BMIs, the device is the decoder,
transforming neural signals into signals that can be used to produce actions. From an information theoretic viewpoint, how well a BMI can perform depends on, in the stimulation case, how effectively information about a stimulus can be conveyed through a neural population, and in the control case, how accurately the interface can extract information from multi-unit recordings. In both cases, the use and structure of the BMI — how the device interfaces with cortical neurons — constrains the effectiveness of the system in reproducing a stimulus or producing the intended control action. In this thesis, we use the information capacity and the complementary concept of rate-distortion to characterize the limits of BMI effectiveness. By studying how the structure and function of the brain-machine interface constrains the system's capacity, we provide useful guidelines for more effective design of BMIs.

1.1 Contributions

This thesis examines the performance limits of brain-machine interfaces systematically using an information theoretic framework, as follows:

**Joint probability models for BMIs.** In Chapter 2, we define an explicit point process model for neural populations that encompasses inter-neuronal correlations. In addition, we define a discrete-time version of the neural population channel that converges in distribution to the continuous-time one, enabling us to compute key information-theoretic quantities for problems that would be otherwise intractable.
We also introduce the basic information theory concepts of capacity and rate-distortion, and show how to apply them in the neural setting.

**Capacity of stimulation BMIs.** In Chapter 3, we define the stimulation BMI channel in terms of our neural population model, and derive the capacity of the channel under a variety of conditions. In particular, we characterize how the capacity changes with inter-neuronal dependence, and how the number and characteristics of the stimulating electrodes can affect the performance. We also elaborate the model to account for attenuation and cross-talk in the stimulation signals.

**Capacity of control BMIs.** In Chapter 4, we provide models for control BMIs when a limited number of electrodes are available to record the neural activity. Computing the capacity under various system constraints, we discuss how the performance is limited by the number and placement of electrodes at the recording site. We also show that spike sorting the recordings can improve the capacity in some cases, but its effects are heavily dependent on the type of errors that the sorting algorithm commits. Finally, we elaborate the model to account for recording noise and attenuation of the neural signals.

**Interpreting the results: Minimum distortion** In Chapter 5, we interpret the capacity results in the context of the rate-distortion function, which measures how accurately a source must be encoded to achieve any specified degree of error [3]. We show how to use this framework to compare BMIs using distortion measures relevant to both
sensory and motor prostheses. In addition, we discuss some limitations of the theory in terms of practical achievability of the performance bounds.
Chapter 2

Information in the Brain

Neurons represent and transmit information via electrical signals known as action potentials, or spikes. An action potential is literally a rapid (1ms) spike of around 100mV in the cell membrane electric potential, generated through the complex interaction of excitatory and inhibitory signals received at the neuron’s inputs (dendrites) and ionic currents that flow between the inside of the neuron and the surrounding extracellular medium. The spike propagates and is actively repeated along the neuron’s axon, which in turn provides inputs to other neurons via synapses [18].

Interestingly, spikes are highly stereotyped throughout the brain, leading neuroscientists to believe that minor individual variations between spike waveforms do not convey information; rather, what is important is when a spike occurs. Consequently, studying information processing in the brain amounts to studying spike trains, event sequences that are completely defined by their precise timing. There are various ways to model and study neural spike trains, but the mathematical framework most frequently employed is point process theory [44], which models neural signals as stochastic processes defined by the event timing [29]. To study the information processing and transmission capabilities of spike trains thus modeled, we turn to information theory, a tool that, although originally developed to describe communication systems, can be adapted to provide a unique insight
into neural processing. We begin by introducing the basic information theoretic concepts that we will be applying to the neural systems. Some of these concepts have already been introduced in [31]; for a more general discussion of information theory and its application to neuroscience, see [33].

2.1 Information Theory and Neuroscience

2.1.1 The Standard Model

Any discussion of information theory must begin with Shannon’s [57] fundamental model of communication systems, shown in Figure 2.1(a). In this model, an information source produces an information-bearing signal $S$. A fundamental assumption of information theory is that all sources are stochastic and consequently are described by their probability distribution $p_S(s)$. Though we will sometimes have to constrain $S$ in order to derive meaningful results about the system, in general the stimulus could be anything: $S$ could be a discrete or continuous-valued random variable, a random sequence, or even a continuous-time random process. In any case, $S$ is somehow transformed by the encoder to the signal $X$ to be input into the communication channel. This encoding may preserve some or all of the information contained in the source signal. The encoded signal then passes through the channel, which disturbs the signal in such a way that the channel’s input $X$ cannot be precisely determined from its output $Y$. The channel’s input-output relation is defined by the conditional probability distribution $p_Y|X(y|x)$. The decoder, which represents the final
(a) Information theory model for communications

(b) Information theory model for neuroscience

Figure 2.1: The classic information theory model of communication systems is shown as well as its translation to a neuroscience context.

processing stage, produces $\hat{S}$, an estimate of the original source signal.

In a similar manner, we can view a neuron as a communication channel, as shown in Figure 2.1(b). A neuron decodes the information expressed by its inputs, processes it and represents the information in its spike-train output. Here, the information source can be either the stimulus in sensory systems or an intended motion in motor systems. To simplify the presentation, we will use terminology from sensory systems, but information theoretic results developed here apply to motor systems as well. Note that we can also take $S$ to express some aspect of the actual stimulus that results from previous processing.

In neuroscience, the encoder represents neural coding: how the stimulus is represented in the firing pattern of one or several neurons. The channel represents a neuron or a population of neurons, wherein “channel disturbances” arise from the stochastic behavior of neural responses. The output of the decoder, $\hat{S}$, is an estimate of the stimulus. Although a stimulus estimate may not explicitly be produced in a neural system, the ability of the
channel to transmit information about the stimulus is characterized by how well the stimulus can be estimated from $Y$. As we shall see, the key to this analysis is the capacity of the channel relating $X$ and $Y$.

### 2.1.2 Information Capacity

To characterize the channel, we begin by defining the **mutual information** between two jointly defined random variables $X$ and $Y$.

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

(2.1)

If we use a base-2 logarithm in the definition, the mutual information has units of *bits*; if we use the natural logarithm, the result has units of *nats*. Note that we have defined mutual information for $X$ and $Y$ having discrete alphabets $\mathcal{X}$ and $\mathcal{Y}$ respectively. This quantity can be similarly defined for continuous random variables, with the sums in (2.1) replaced with integrals over the continuous alphabets [15].

Some important properties of mutual information include:

- $I(X;Y) \geq 0$,
- $I(X;Y) = 0$ if and only if $X$ and $Y$ are statistically independent,
- $I(X;Y)$ achieves its maximal value when $X = Y$.

Consequently, mutual information summarizes the degree of similarity between the statis-
tical properties of two random variables beyond the simple linear, pairwise dependencies captured by computing the correlation [23].

It is often useful to express mutual information in terms of a second quantity known as entropy:

\[ I(X;Y) = H(Y) - H(Y|X). \] (2.2)

Here, the entropy \( H(Y) \) and the conditional entropy \( H(Y|X) \) are defined as

\[
H(Y) = - \sum_{y \in Y} p_Y(y) \log p_Y(y)
\]

\[
H(Y|X) = - \sum_{x \in X} p_X(x) \left( \sum_{y \in Y} p_{Y|X}(y|x) \log p_{Y|X}(y|x) \right)
\]

Writing mutual information as a difference of entropies as in (2.2) highlights the fact that it depends upon both the channel’s input-output relationship and the probability distribution of the input. The dependence on the input is disguised: the probability distribution of the output \( p_Y(y) \) equals \( \sum_{y|X} p_{Y|X}(y|x)p_X(x) \), showing that \( H(Y) \) depends on both the channel’s input-output relationship and the input’s probability distribution. Consequently, mutual information does not summarize the behavior of the channel. The mutual information between the stimulus and a measured response depends on neural processing, the stimuli and the stimulus probabilities.

To separate the effect of the input from the channel, we need a quantity that takes into account how the channel acts on any possible input. This quantity is the information ca-
**Parity**, which is the mutual information between the channel input and output, maximized over all possible inputs within some constraint class $\mathcal{C}$:

$$C \triangleq \max_{p_X(\cdot) \in \mathcal{C}} I(X;Y).$$  \hspace{1cm} (2.3)

In communication systems, $\mathcal{C}$ might represent a constraint on the maximum input power. In neuroscience, the spike rate might be constrained to lie within some range due to the properties of the biophysical processes involved. Defining the capacity this way makes the system’s constraints explicit when considering how the channel limits the ability to estimate the stimulus.

For a digital communication system, Shannon showed that the information capacity captures everything required to determine when the channel can be used reliably to communicate information. More precisely, the Noisy Channel Coding Theorem states that, given a discrete memoryless channel, if the number of bits $R$ to be input to the channel is less than capacity, there exists a channel coding scheme such that all errors incurred in the channel can be corrected. Conversely, if $R > C$, no scheme exists that can prevent errors from occurring. Thus, capacity uniquely defines a sharp boundary between reliable (error-free) and unreliable digital communication [15].

Unfortunately, applying the Noisy Channel Coding Theorem to neural systems seems to present a conceptual difficulty. Although methods for estimating the capacity of neural systems have been widely discussed in the neuroscience literature [50], it is not immedi-

ately clear what relevance this quantity has to a neural system; whereas the goal of a digital communication system is optimal error-free communication, it is unlikely that neural systems share this goal, especially when achieving capacity could require infinite delay and complexity. Instead, the significance of capacity in this setting becomes evident only when viewed in the context of another of Shannon’s classic results: rate-distortion theory.

2.1.3 Rate-distortion

In the model shown in Figure 2.1, the source signal $S$ is encoded, transmitted and/or processed, and decoded to produce the estimate $\hat{S}$. To assess the fidelity of this estimate, we begin by defining a distortion measure $d(s, \hat{s})$. Presumably the distortion increases as the discrepancy increases between the stimulus and its reconstructed value. However, the only requirement is that the measure is non-negative and bounded; otherwise, the distortion measure can be chosen according to whatever criteria is most relevant in any particular scenario. A common distortion measure used in communications and signal processing is the squared-error measure: $d(s, \hat{s}) = (\hat{s} - s)^2$. More relevant to sensory neuroscience perhaps would be a perceptual error measure, such as one that reflects Weber’s Law [58]. The distortion measure could alternatively incorporate a desired processing of the stimulus, making $\hat{S}$ an approximation to a feature extracted version of $S$. In motor systems, the distortion measure could account for differences between intended motion $S$ and actual motion $\hat{S}$, and could even include a penalty for velocity as well as path and target errors.

The average distortion $D$ is the expected value of the distortion measure with respect to
the joint distribution of the stimulus and its estimate.

\[ D \triangleq \mathbb{E}[d(s, \hat{s})] \]
\[ = \sum_s \sum_{\hat{s}} d(s, \hat{s}) p_{S,S}(s, \hat{s}) \]
\[ = \sum_s \sum_{\hat{s}} d(s, \hat{s}) p_{\hat{S}|S}(\hat{s}|s) p_S(s) \]

Note that the conditional distribution \( p_{\hat{S}|S}(\hat{s}|s) \) depends on virtually everything in a neural coding scenario: how the stimulus is encoded, the neuron’s spiking characteristics and how the decoder works.

As we did when we studied the capacity, we want to separate the effects of the channel from the properties of the source. In other words, we seek a quantity that summarizes the distortion characteristics of the system under any possible transformation from \( S \) to \( \hat{S} \). This quantity is the rate-distortion function \( \mathcal{R}(D) \), which is defined to be the minimum of the mutual information between the stimulus and its estimate over all possible channels, encoders and decoders that yield a specified average distortion \( D \):

\[ \mathcal{R}(D) \triangleq \min_{p_{\hat{S}|S}(\cdot|s): \overline{D} \leq D} I(S; \hat{S}). \]  (2.4)

Note that the minimization is calculated over all possible relationships between a stimulus and its estimate, not just the one under study.

The rate-distortion function for any source has two important properties that are de-
picted in Figure 2.2:

- $\mathcal{R}(D)$ is a strictly decreasing and convex function.

- $\mathcal{R}(D) = 0$ for all values of $D \geq D_{\text{max}}$. The critical value $D_{\text{max}}$ is the maximal distortion, incurred when the decoder makes its "best guess" as to what the stimulus might be with no data.

The significance of the rate-distortion function is revealed by the Rate-Distortion Theorem, which states that, given a source distribution $p_S(s)$ and a distortion measure $d(S, \tilde{S})$, if the number of bits $R$ used to describe the source is greater than $\mathcal{R}(D)$, then the average distortion $D$ can be achieved. Conversely, if $R < \mathcal{R}(D)$, then no encoding exists that can achieve $D$ [15]. Thus, the rate-distortion defines the minimum number of bits required to represent a source with a given fidelity. In other words, the value of $R$, known as the rate, measures the quality of the encoding system. In general, the greater the rate, the more information about the signal can be conveyed and the smaller the achievable distortion.

Rate-distortion functions are notoriously difficult to calculate, with only a few results known. One example is the bandlimited Gaussian random process having power $P$ and bandwidth $W$. If the squared error distortion measure is used, the rate-distortion function equals [3, Chap. 4]

$$\mathcal{R}(D) = \begin{cases} W \log_2 \frac{P}{D}, & D \leq P \\ 0, & D > P. \end{cases}$$ (2.5)

So far, we have separated the study of communication systems into two parts: capacity,
Figure 2.2: Capacity and rate-distortion are defined in the context of the standard model of Figure 2.1. Capacity $C$ summarizes the channel that presumably introduces disturbances into the communication process. The rate-distortion function $R(D)$ depends solely on the source characteristics. Shannon’s Source-Channel Separation Theorem relates these two quantities, showing that the smallest possible distortion $D_{\text{min}}$ is determined by $C = R(D_{\text{min}})$, which is a property of the channel irrespective of the source, and rate-distortion, which is a property of the source independent of the intervening channel. Shannon’s crowning result, the Source-Channel Separation Theorem, unifies these results and provides the basis for our framework for studying neural communication.

### 2.1.4 Source-Channel Separation and Minimum Distortion

Figure 2.2 illustrates the separate study of channel (capacity) and source (rate-distortion) in the context of the basic communication system model. In order to draw any conclusions about the system as a whole, we need a way of marrying the two concepts into a whole-system analysis. This is accomplished through the Source Channel Separation Theorem, which states that, given a stationary, ergodic source to be transmitted across a memoryless channel, if $R(D) < C$, then the average distortion $D$ is achievable [20]. Furthermore, since the rate-distortion is convex and strictly decreasing, the distortion at which the rate-distortion function equals the capacity defines the smallest possible distortion $D_{\text{min}}$ any
encoder and decoder can obtain (see Figure 2.2): \( R(D_{\text{min}}) = C \).

Shannon's proof of this result was not constructive, providing no guidance on how to find the encoder/decoder pair that produces the smallest possible distortion. However, the value of \( D_{\text{min}} \) defined by the source and the channel determines how well a given system can perform, thereby serving as a benchmark.

When the source is discrete-valued, the value of the rate-distortion function is indeed the bit rate, the number of bits used to encode the signal that can result in a specified or greater distortion. However, the theorem also applies when no digital scheme is involved. We can interpret \( R(D) \) as the virtual bit rate in such examples, which would suggest that some equivalent digital scheme exists. Whether this equivalent system can be found or not is irrelevant. More importantly, "bits" are the fundamental unit of exchange in any communication or signal processing system. It is an intermediary value; what we really want to know is the distortion.

Because the rate-distortion function is always a decreasing function, increasing capacity always allows the possibility of a smaller distortion. In our Gaussian example (equation 2.5), the smallest possible distortion decreases exponentially with capacity.

\[
D_{\text{min}} = D_{\text{max}} 2^{-C/W}
\]

Note that this result applies no matter what channel intervenes between the encoder and decoder. It could be a radio channel, cable television or a group of neurons. Whenever the
Source-Channel Separation Theorem holds, the capacity determines the minimum achievable distortion, no matter how distortion is measured. It is this generality of information theory that makes it fundamental to the study of communication, signal processing and control systems, be they natural or man-made.

2.2 Modeling Neural Responses: The Point Process Channel

2.2.1 The Single Point Process Channel

In Section 2.1, we showed how neural systems can be viewed in the context of the information theory model, and how capacity can be used to assess the ability to extract information from the neural channel. In order to compute the capacity of a neural system, we need an explicit model for cortical neurons that reflects how information is encoded and transmitted. To do this we define the point process channel, based on the most widely used statistical model for neural activity [29].

The point process channel is depicted in Figure 2.3. Given a real-valued, time-varying input $X(t)$, the channel produces a sequence of events, represented by the counting process $\{N_t; t \geq 0\}$ which denotes the number of events that have occurred up to time $t$. For a regular point process, the probability of events occurring in a small time interval $[t,t + \Delta t)$
Figure 2.3: The point process channel is depicted. The channel produces a point process with intensity function $\mu(t; X(t), N_t, u_t)$ that is a function of the stationary input $X(t)$ as well as the point process history. The channel output is a sequence of events, represented by the counting function $N_t$.

is given by

$$
\Pr[N_{t+\Delta t} - N_t = 1 \mid N_t, u_t] = \mu(t; X(t), N_t, u_t)\Delta t, \\
\Pr[N_{t+\Delta t} - N_t > 1 \mid N_t, u_t] = o(\Delta t).
$$

In the latter expression, $\lim_{\Delta t \to 0} o(\Delta t)/\Delta t = 0$, meaning that the probability of more than one event in a small interval decreases superlinearly. The quantity $\mu(t; X(t), N_t, u_t)$, known as the intensity function, represents how the instantaneous event rate depends on the input $X(t)$ and on the process’s history, which includes the number of events $N_t$ and the times $u_t = \{u_1, \ldots, u_{N_t}\}$ at which they occurred. In other words, the probability of an event occurring in a small interval is proportional to the length of the interval, and depends on all previous events. Because of this, we say the channel has memory.

Note that the statistical structure of the point process, which is completely determined by the intensity function $\mu(t; X(t), N_t, u_t)$, may depend explicitly on the time $t$, and may therefore be non-stationary. In our model, dependence on $t$ is usually taken to be an ex-
trinsic property of the system, representing the encoding of \( X(t) \), which is produced by a source external to the neuron. On the other hand, history dependence is an intrinsic property of the neuron itself; for example, the refractory period that usually follows after a neuron has produced a spike is expressed in the point process model as a dependence on the last event time [29]. The simplest point process, the \textit{Poisson process}, expresses no history dependence. In that case, we can write the intensity function as an instantaneous rate, 
\[
\mu(t; X(t), N_t, u_t) = \lambda(t).
\]

Finding the capacity of the point process channel requires a slight modification to the definition given in (2.3):

\[
C = \lim_{T \to \infty} \max_{\mu(t; X(t), N_t, u_t) \in \mathcal{C}} \frac{1}{T} \mathbb{I}(X_{\{0 \leq t < T\}}; N_{\{0 \leq t < T\}}),
\]

(2.6)

where we have used the notation \( X_{\{0 \leq t < T\}} \) to denote \( \{X(t); 0 \leq t \leq T\} \). In words, capacity is the maximal asymptotic time-averaged mutual information rate between the input signal waveform and the point process output of the channel; it has units of bits per second. To use this definition, we require the input signal to be a stationary random process. The resulting channel output is known as a \textit{doubly stochastic}, or \textit{Cox} process [44], since the point process intensity is now itself a random process. Here, \( \mathcal{C} \) represents an intensity constraint class, which implicitly places constraints on the input. The choice of constraint(s) that define the class can strongly affect capacity results. We focus on constraints on the minimum and
maximum intensity:

\[ \lambda_{\text{min}} \leq \mu(t; X(t), N_t, u_t) \leq \lambda_{\text{max}}. \]

In all point-process capacity calculations, the minimum and maximum rates are con­strained; if the maximum rate were not constrained, the capacity would be infinite. Other constraints, such as a constraint on the average rate, can be added as necessary. Note that neuron models containing no inherent variability when the input is deterministic, like integrate-and-fire models [6], have an infinite maximal rate, and consequently infinite ca­pacity*.

The capacity of the Poisson process channel with constraints on the instantaneous rate is a known result in optical communication theory, where the channel output models photon counts received by an optimal detector [5, 17, 35, 67]. Assume, without loss of generality, that the channel input is the instantaneous rate function \( \lambda(t) \), \( 0 \leq t < \infty \), and impose minimum and maximum rate constraints: \( \lambda_{\text{min}} \leq \lambda(t) \leq \lambda_{\text{max}} \). Then, the capacity is

\[
C = \frac{\lambda_{\text{min}}}{\log 2} \left[ \frac{1}{e} \left( \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}} \right)^{\frac{\lambda_{\text{max}} - \lambda_{\text{min}}}{\lambda_{\text{max}}}} - \log \left( \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}} \right)^{\frac{\lambda_{\text{max}} - \lambda_{\text{min}}}{\lambda_{\text{max}}}} \right].
\]

The division by \( \log 2 \) leaves the capacity with units of bits/s. The capacity is achieved when the instantaneous rate is a random telegraph wave—the rate randomly switches between its minimum and maximum values—with the probability of being at the maximum rate at

\*Without additional constraints on the inputs, all deterministic models (even those having a non-zero refractory interval limiting the maximum event rate) have infinite capacity if the inputs can be uniquely determined from the outputs.
any given time equaling $1/e$. In most cases of interest in neuroscience, the appropriate minimum-rate constraint is zero, which greatly simplifies the capacity formula,

$$\lim_{\lambda_{\text{min}} \to 0} C = \frac{\lambda_{\text{max}}}{e \log 2}, \quad (2.7)$$

which means the capacity-achieving signal has an average rate of $\lambda_{\text{max}}/e$. When the average rate is constrained to equal $\bar{\lambda}$, the capacity $\bar{C}$ is smaller, and only equals $C$ when $\bar{\lambda} = \lambda_{\text{max}}/e$.

$$\lim_{\lambda_{\text{min}} \to 0} \bar{C} = \frac{\bar{\lambda}}{\log 2} \log \frac{\lambda_{\text{max}}}{\bar{\lambda}} \quad (2.8)$$

This result shows how much the capacity can change when the constraint class is changed.

The capacity of non-Poisson point process channels is difficult to compute in general, and results are only known in a few cases. Nevertheless, Kabanov [35] proved that the capacity of any point process satisfying the same constraints must be less than the Poisson channel's capacity.

Unfortunately, extending the single point process capacity results to several point processes is very difficult using point process theory, especially when exploring the effect on capacity of dependencies that would arise from interneuronal interactions. Rather than seeking a solution that is broadly applicable to all vector point processes, we focus on vector Poisson processes.
2.2.2 The Vector Poisson Process

To generalize the single point process result to population channels, we need a model for jointly defined point processes. Unfortunately, the joint probability distribution for a vector point process is unwieldy at best, especially when incorporating inter-process dependencies. We can construct a vector Poisson process, however, for the special case in which the collection has the property known as infinite divisibility [16], meaning that it can be infinitely decomposed into sums of independent vector Poisson processes.

Generalizing a method of Holgate [28], we can form $M$ jointly Poisson processes using superpositions of collections of no more than $2^M - 1$ statistically independent building-block Poisson processes [32]. For example, to construct a pair of dependent Poisson processes, we use three independent building block processes, which we denote by $B_{1,t}, B_{2,t}, B_{3,t}$ that have instantaneous rate functions $\nu_1(t), \nu_2(t), \nu_3(t)$, respectively. We form the pair according to the superposition

$$N_{1,t} = B_{1,t} + B_{3,t}$$

$$N_{2,t} = B_{2,t} + B_{3,t}.$$

All of the dependence between the constructed processes is expressed by the building-block process $B_{c,t}$ they share in common. The correlation function between the two processes can
be written in terms of the instantaneous rates:

$$\rho^{(2)}(t_1, t_2) = \begin{cases} \frac{v_3(t)}{\sqrt{(v_1(t) + v_3(t))(v_2(t) + v_3(t))}}, & t_1 = t_2 = t \\ 0, & t_1 \neq t_2 \end{cases}$$

(2.9)

The correlation between the two processes is thus instantaneous: given an event occurs in one process, it is correlated with the other process only at that event time, and uncorrelated at all other times. Correlations with non-zero time lags can be introduced by adding a time offset to the common building-block processes. In addition, the correlation lies in the interval $[0, 1]$, reaching 1 in the limit of large common rate $v_3(t)$. Although joint Poisson processes have been constructed with correlation functions having both a temporal extent and negative-valued correlations [25, 30], their joint distribution is not infinitely divisible.

In Appendix A we show how to generalize the Holgate construction to an arbitrary number of jointly defined Poisson processes such that the ensemble is infinitely divisible. Importantly, we show that infinitely divisible jointly Poisson processes can have dependencies of higher order than just pairwise correlation. For example, constructing three dependent Poisson processes requires up to seven building-block processes:

$$N_{1,t} = B_{1,t} + B_{4,t} + B_{5,t} + B_{7,t}$$

$$N_{2,t} = B_{2,t} + B_{4,t} + B_{6,t} + B_{7,t}$$

$$N_{3,t} = B_{3,t} + B_{5,t} + B_{6,t} + B_{7,t}.$$
Three building-block processes are shared among pairs and one process is shared in common by all. The shared processes $B_{4,t}, B_{5,t}, B_{6,t}, B_{7,t}$ introduce pairwise dependencies and $B_{7,t}$ produces an additional third-order dependence as well.

To simplify the analysis we focus on the symmetric case, wherein the building-block processes unique to each constructed process have the same instantaneous rate $v^{(1)}(t)$, and the building block processes shared by any $m$ constructed processes have the same rate $v^{(m)}(t)$. The dependencies of every order are summarized by the cumulant correlation function (Appendix B), where the correlation function of order $k$ is given by

$$
\rho^{(k)}(t) = \frac{\sum_{m=k}^{M} \binom{M-k}{m-k} v^{(m)}(t)}{\sum_{m=1}^{M} \binom{M-1}{m-1} v^{(m)}(t)}, \quad k = 2, \ldots, M,
$$

The numerator equals the sum of the rates of the building block processes contributing to interactions of order $k$ and higher; the denominator equals the instantaneous rate of each constructed process. As in (2.9), the correlation function of any order is an instantaneous function of $t$, and has no temporal extent. Since the construction is symmetric, the correlation function between any $m$ processes is the same. Equation (2.10) implies that cumulant correlation coefficients of all orders are non-negative, less than one and smaller than all lower-order cumulant correlation coefficients: $0 \leq \rho^{(k+1)}(t) \leq \rho^{(k)}(t) \leq 1, \quad k = 2, \ldots, M - 1$. Additional properties can be derived by combining the cumulant correlation coefficients in light of the structure that equation (2.10) enforces. These relationships can be summarized
as
\[
\sum_{k=2}^{M} \rho^{(k)}(-1)^{k} \binom{M - 1}{k - 1} \leq 1, \\
\sum_{k=m}^{M} \rho^{(k)}(-1)^{k+m} \binom{M - m}{k - m} \geq 0, \quad m = 2, \ldots, M.
\] (2.11)

For example, the cumulant correlation coefficients for three symmetric jointly Poisson processes must obey the inequality chain

\[
0 \leq \rho^{(3)}(t) \leq 2\rho^{(2)}(t) - \rho^{(3)}(t) \leq 1.
\]

When \(M = 4\), the following relationships must hold.

\[
3\rho^{(2)} - 3\rho^{(3)} + \rho^{(4)} \leq 1 \\
\rho^{(2)} - 2\rho^{(3)} + \rho^{(4)} \geq 0 \\
1 \geq \rho^{(2)} \geq \rho^{(3)} \geq \rho^{(4)} \geq 0
\]

Furthermore, if no building block processes of order higher than two are present, which makes \(\rho^{(3)} = 0, \rho^{(4)} = 0, \ldots, \rho^{(M)} = 0\), the second-order correlation coefficient cannot be bigger than \(1/(M - 1)\). Thus, the cumulant correlation coefficients have well-defined properties, making them more useful for evaluating the vector Poisson process than the usual correlation coefficients derived from moments.

The generalized Holgate method thus provides a straightforward way of constructing jointly defined Poisson processes that are infinitely divisible. Despite this, direct capacity
calculations on the vector Poisson channel are difficult. Even if we could write out the joint probability function for the channel, the resulting mutual information calculation would be intractable. Fortunately, there is an alternative; in the next section, we discuss an analogous discrete-time channel, the vector Bernoulli channel, that is equivalent to the vector Poisson channel in the limit of small time intervals. Using this fact, rather than compute capacities for the vector Poisson channel directly, we can compute the capacity first for the vector Bernoulli channel, and then evaluate the limit as bin-width approaches zero to infer the capacity of the vector Poisson channel. This approach, similar to the one taken by Wyner [67] to obtain the capacity of the single Poisson channel, is justified because of the smoothness properties of mutual information, which allows us to evaluate well-behaved limits in any order.

2.3 The Bernoulli Channel

2.3.1 The Single Bernoulli Channel

A Bernoulli process $Y(n)$ equals either zero or one at each bin index $n$, and is statistically independent from any one bin to another. We define the Bernoulli channel by its conditional probability function

$$P(Y(n) \mid X(n)) = \begin{cases} 
X(n), & Y(n) = 1 \\
1 - X(n), & Y(n) = 0.
\end{cases} \tag{2.12}$$
In other words, the channel input $X(n) \in [0, 1]$ is the success probability for the channel's Bernoulli output at bin $n$. We assume the input is a stationary, memoryless process, making the channel output a *doubly stochastic Bernoulli process*. Since the channel is also stationary and memoryless, the statistical descriptions of both the input and the output Bernoulli process do not depend on $n$, so we suppress the dependence on bin index hereafter to simplify the notation.

The unconditional output probability distribution has a simple expression.

$$P(Y) = \begin{cases} \bar{X}, & Y = 1 \\ 1 - \bar{X}, & Y = 0 \end{cases}$$

Here, $\bar{X} = E[X]$, the expected value of the input.

The conditional entropy of the channel and the entropy of its output are therefore

$$H(Y \mid X) = -E[X \log X + (1 - X) \log(1 - X)]$$

$$H(Y) = -\bar{X} \log \bar{X} - (1 - \bar{X}) \log(1 - \bar{X}),$$

and we can find the mutual information between the channel inputs and outputs using the definition in (2.2):

$$I(X; Y) = E[X \log X + (1 - X) \log(1 - X)] - \bar{X} \log \bar{X} - (1 - \bar{X}) \log(1 - \bar{X}). \quad (2.13)$$
To compute the capacity, we have to maximize the mutual information with respect to the input distribution $p_X(x)$. Mimicking the maximum rate constraint we imposed in Section 2.2.1, we constrain the input to be in the interval $[0, x_{\text{max}}]$. Wyner [68] proved that the capacity achieving input distribution consists of impulses (probability masses) situated at the extremes of the input’s possible values:

$$p_X(x) = q \delta(x - x_{\text{max}}) + (1 - q) \delta(x),$$

with $\delta(\cdot)$ denoting a Dirac delta-function and $q$ the probability parameter. This input probability distribution allows easy evaluation of the expected value in (2.13):

$$I(X;Y) = qx_{\text{max}} \log x_{\text{max}} + q(1 - x_{\text{max}}) \log(1 - x_{\text{max}})$$
$$- qx_{\text{max}} \log qx_{\text{max}} - (1 - qx_{\text{max}}) \log(1 - qx_{\text{max}}).$$

Taking the derivative of (2.15) and setting it to zero, we find the maximizing probability $q_C$ equals

$$q_C = \frac{(1 - x_{\text{max}}) x_{\text{max}}^{-1}}{1 - x_{\text{max}} + x_{\text{max}}(1 - x_{\text{max}}) x_{\text{max}}^{-1}}.$$ 

The resulting expression for the capacity is too complicated to show here, but is easy to find by substituting the maximizing value of $q$ into (2.15).

It is well known that the Bernoulli process defined over discrete time bins $\Delta t$ converges to a Poisson process in the limit as $\Delta t \to 0$ [19]. To verify that the capacity of the Bernoulli
channel converges to the capacity of the Poisson channel we obtained in (2.7), we substitute $x_{\text{max}} = \lambda_{\text{max}} \Delta t$ into the formulas for capacity and the maximizing probability $q$, and let $\Delta t \to 0$. Doing so ensures that the rate stays the same as the time bins get smaller. Calculations show that $\lim_{\Delta t \to 0} q C = 1/e$. Substituting this result into the expression for mutual information, and dividing the result by $\Delta t$ (so that we obtain the result in bits/s), we find the capacity is

$$C = \frac{\lambda_{\text{max}}}{e \log 2} + o(\Delta t) \text{ bits/s},$$

the same expression as (2.7) for the point-process capacity under a maximum rate constraint.

If we want to impose an average rate constraint as well, note that the average value of the capacity-achieving input distribution is $\bar{X} = qx_{\text{max}}$. Thus the probability $q$ controls the average value of the input. To find the capacity under both maximal and average input constraints, we simply set $q$ in (2.15) to the value $\bar{X}/x_{\text{max}}$. Echoing the previous analysis, setting $x_{\text{max}} = \lambda_{\text{max}} \Delta t$ and letting $\Delta t \to 0$, we find that with $q = \bar{\lambda}/\lambda_{\text{max}}$, the average-rate-constrained point process capacity of (2.8) results:

$$\bar{C} = \frac{\bar{\lambda}}{\log 2} \log \frac{\lambda_{\text{max}}}{\bar{\lambda}} \text{ bits/s}.$$ 

In either case, the capacity-achieving input is a discrete-time "telegraph wave" switching randomly between zero and the maximum probability. With only a maximal rate constraint, the probability of being in the maximal-probability state equals $1/e$, the same as we found
2.3.2 The Vector Bernoulli Process

To elaborate the single Bernoulli channel to a vector channel, we need a model for the joint probability function of a Bernoulli random vector. We will begin with the simplest case, $M = 2$, and then elaborate to include an arbitrary number of jointly defined processes.

Let $X = (X_1, X_2)$ and $Y = (Y_1, Y_2)$ be the length-two vector inputs and outputs of the channel, respectively. As we did for the single channel, we define each channel's input $X_m \in [0,1]$ to be the success probability for the channel's Bernoulli output:

$$P(Y_m | X_m) = \begin{cases} X_m, & Y_m = 1 \\ 1 - X_m, & Y_m = 0. \end{cases}$$

When the channels are independent (no interneuron dependencies), the individual conditional probabilities thus defined are sufficient to characterize the vector channel's input-output relationship. In general, however, we also require a joint definition of the channels, which is captured by

$$P[Y | X] = P[Y_1 | X_1] P[Y_2 | X_2] \cdot \left[ 1 + \rho_{12}^{(2)} \frac{(Y_1 - E[Y_1 | X_1])(Y_2 - E[Y_2 | X_2])}{\sqrt{\sigma_{Y_1 | X_1}^2 \sigma_{Y_2 | X_2}^2}} \right].$$

This probability model is known as the Sarmanov-Lancaster expansion [24]. Here, $\rho_{12}^{(2)}$ is the simple pairwise correlation coefficient between the Bernoulli random variables $Y_1 | X_1$.
and $Y_2|X_2$. When $\rho^{(2)} \neq 0$, the channels are called conditionally dependent. Since $Y_m|X_m$ is itself a Bernoulli random variable, its variance has a simple expression:

$$\sigma_{Y_m|X_m}^2 = X_m(1 - X_m).$$

For $M$ Bernoulli channels, the conditional joint distribution can be written in its most general form as

$$P[Y|X] = \prod_{m=1}^{M} P[Y_m|X_m] \left[ 1 + \sum_{i_1=1}^{M-1} \sum_{i_2=i_1+1}^{M} \rho_{i_1i_2}^{(2)} \frac{(Y_{i_1} - E[Y_{i_1}|X_{i_1}]) (Y_{i_2} - E[Y_{i_2}|X_{i_2}])}{(\sigma_{Y_{i_1}|X_{i_1}}^2 \sigma_{Y_{i_2}|X_{i_2}}^2)^{1/2}} \right]$$

$$+ \sum_{i_1=1}^{M-2} \sum_{i_2=i_1+1}^{M-1} \sum_{i_3=i_2+1}^{M} \rho_{i_1i_2i_3}^{(3)} \frac{(Y_{i_1} - E[Y_{i_1}|X_{i_1}]) (Y_{i_2} - E[Y_{i_2}|X_{i_2}]) (Y_{i_3} - E[Y_{i_3}|X_{i_3}])}{(\sigma_{Y_{i_1}|X_{i_1}}^2 \sigma_{Y_{i_2}|X_{i_2}}^2 \sigma_{Y_{i_3}|X_{i_3}}^2)^{2/3}}$$

$$+ \cdots + \rho^{(M)} \prod_{j=1}^{M} \frac{(Y_j - E[Y_j|X_j])}{(\sigma_{Y_j|X_j}^2)^{M-1/M}}$$

In this expression, the coefficients $\rho_{ij}^{(2)}$ equal the simple pairwise correlation coefficient between the pair of Bernoulli random variables $Y_i$ and $Y_j$. The other coefficients account for higher order dependencies in the ensemble. In general, a collection of $M$ channels has $2^M - M - 1$ correlation coefficients.

Using this expression for the joint probability function, we can construct a vector Bernoulli process in the same way that we construct a single Bernoulli process. In Appendix C we show that the vector Bernoulli process converges in distribution to the in-
finitely divisible vector Poisson process, and there is an explicit connection between the correlation coefficients of the vector Bernoulli process and the cumulant correlation coefficients of the vector Poisson process. Consequently, we can compute the capacity of the general vector Poisson channel by first computing the capacity of the vector Bernoulli channel and then evaluating the small bin-width limit. In Chapters 3 and 4 we will use this approach to derive the capacities of the BMI stimulation and control channels, respectively.
Chapter 3
The Neural Stimulation Channel

The goal of a stimulation BMI is to transmit information about a sensory signal to the brain by stimulating a relevant set of neurons. For example, in a cochlear implant, sounds received by a microphone are processed and transformed into an activation pattern for electrodes implanted in the cochlea, which stimulate nearby nerve fibers with a modulated electrical current. Similarly, in a retinal implant, visual information received by a camera is encoded in the activation pattern of electrodes that stimulate retinal ganglion cells in the eye. Several factors contribute to the perceptual quality of such prostheses, including the number, size, spacing and location of the electrodes, the BMI’s electrical dynamic range, the encoding method used, and the patient’s history and physiology (e.g., duration of deafness or blindness, number of surviving sensory neurons, etc.) [41, 52]. In general, the effectiveness of any stimulation prosthesis will be limited by the method of stimulation as well as the size and characteristics of the population being stimulated.

Our two baseline models of the stimulation BMI channel are depicted in Figure 3.1. In the first model, a population of \( M \) neurons shares a single input \( X \). This can be seen as an idealized model of extracellular stimulation, where stimulation current is delivered identically to an entire population of neurons in the vicinity of a single electrode. In the second model, every neuron in the population receives a separate input; this represents the
Figure 3.1: Two configurations of stimulation BMI channels. (a) The entire population receives a single input, modeling extracellular stimulation of an entire population by a single electrode. (b) Each neuron in the population is independently stimulated, modeling intracellular stimulation. In both cases, connection induced dependencies between component neurons may be present.

The opposite extreme, in which intracellular stimulation current is delivered to each neuron directly, without interference. In both cases, lateral connections between neurons may result in inter-neuronal dependencies that can affect the system's capacity as well. Using the techniques we developed in Chapter 2, we will compute the capacity of the extracellular and intracellular stimulation BMI channels to compare the effectiveness of the different stimulation techniques, and to explore the effects of population size, inter-neuronal dependence, and crosstalk.
3.1 Extracellular Stimulation

3.1.1 Independent Neurons

We begin with the simplest extracellular stimulation channel, depicted in Figure 3.1(a), wherein a population of neurons is stimulated by a single input. We will use the Bernoulli channel approximation to derive the capacity of the vector Poisson channel. When the neurons are independent, the corresponding Bernoulli channel model is

\[ P(Y | X) = \prod_{m=1}^{M} P(Y_m | X), \]

(3.1)

where \( P(Y_m | X) \) is the single Bernoulli channel conditional probability given in (2.12).

To find the mutual information, we use (2.2), which expresses the mutual information as a difference of entropies. The conditional entropy term equals the expected value of the sum of all outputs’ conditional entropies,

\[ H(Y | X) = \mathbb{E}_x \left[ \sum_m H(Y_m | X) \right] \]

\[ = \mathbb{E}_x \left[ \sum_m \sum_{y_m} -P_{Y_m|X}(y_m | X) \log P_{Y_m|X}(y_m | X) \right]. \]

The conditional entropy \( H(Y | X) \) for the common-input case can now be easily found. When \( X \) has the probability distribution given in (2.14), we get

\[ H(Y | X) = -Mq \cdot (x_{\text{max}} \log x_{\text{max}} + (1 - x_{\text{max}}) \log(1 - x_{\text{max}})) \]

(3.2)
The first term in the expression for mutual information, the entropy of the joint output $H(Y)$, does not equal the sum of the component entropies because the common input makes the outputs statistically dependent on each other. However, when the input has the form of (2.14), the unconditional joint probability distribution of the output simplifies, and can be expressed in terms of the number of non-zero outputs:

$$P[Y = y] = P \left[ \sum_m y_m = m_{nz} \right] = \begin{cases} q(1 - x_{\text{max}})^M + (1 - q), & m_{nz} = 0 \\ q x_{\text{max}}^{m_{nz}} (1 - x_{\text{max}})^{M - m_{nz}}, & m_{nz} = 1, \ldots, M \end{cases}$$

The joint entropy consequently equals

$$H(Y) = - \left[ q(1 - x_{\text{max}})^M + (1 - q) \right] \log \left[ q(1 - x_{\text{max}})^M + (1 - q) \right] - \sum_{m_{nz} = 1}^{M} \binom{M}{m_{nz}} q x_{\text{max}}^{m_{nz}} (1 - x_{\text{max}})^{M - m_{nz}} \log \left[ q x_{\text{max}}^{m_{nz}} (1 - x_{\text{max}})^{M - m_{nz}} \right],$$

and the mutual information equals the difference of (3.3) and (3.2). Focusing on the asymptotic case $x_{\text{max}} \to 0$, we find that

$$I(X; Y) = M (-q \log q) \cdot x_{\text{max}} + o(x_{\text{max}}).$$

Ignoring the higher order term, the maximizing value of $q$ equals $1/e$, and the capacity
of the channel is the capacity of a single Poisson channel, scaled by the population size:

\[ C^{(M)} = M \frac{\lambda_{\text{max}}}{e \log 2} \]

Here, we have used the notation \( C^{(1)} = \frac{\lambda_{\text{max}}}{e \log 2} \) to denote the capacity of a single neuron channel, and \( C^{(M)} \) to denote the capacity of \( M \) neurons. We have also divided the result by \( \log 2 \) to yield the capacity in bits/s.

In order to derive this result, we had to assume a particular form for the input distribution, that of the random telegraph. Thus, in order to show that the value we obtained for mutual information is, indeed, the capacity, we must show that it is a maximum over all possible inputs that satisfy the rate constraint \( 0 \leq X \leq x_{\text{max}} \). To do so, we note that conditioning cannot increase entropy, and that the joint entropy of a collection of random variables is maximized when the variables are statistically independent [15]; hence,

\[
I(X; Y) = H(Y) - H(Y | X)
\]

\[
\geq H(Y) - H(Y | X, X_2, \ldots, X_M)
\]

\[
\leq \sum_{m=1}^{M} (H(Y_m) - H(Y_m | X_m))
\]

\[
= M I(X_1; Y_1),
\]

where inequality (a) follows because conditioning on more variables can only decrease the
entropy, and (b) is true when the variables $X_m$ are i.i.d. inputs to $M$ independent channels. Since the capacity of the single Poisson channel is the maximum of $I(X_1;Y_1)$ in (c), we have achieved the upper bound and thus (3.4) is the capacity of the extracellular stimulation channel. This result may seem surprising until one recognizes that independent Poisson processes driven by a common input act like a single Poisson process having a rate equal to the sum of the individual rates. As capacity is proportional to maximal discharge rate, the capacity results should agree.

### 3.1.2 Dependent Neurons

When inter-neuronal dependencies are present, the outputs are conditionally dependent, and (3.1) no longer applies. Instead, we use the Sarmanov-Lancaster expansion for the joint probability function. Since we are dealing with a single common input, we can simplify (2.16) slightly:

$$
P[Y|X] = \prod_{m=1}^{M} P[Y_m|X] \cdot \left[ 1 + \sum_{i_1=1}^{M-1} \sum_{i_2=i_1+1}^{M} \sum_{i_3=i_2+1}^{M} \rho_{i_1i_2i_3}^{(3)} \frac{(Y_{i_1} - E[Y_{i_1}|X]) (Y_{i_2} - E[Y_{i_2}|X]) (Y_{i_3} - E[Y_{i_3}|X])}{\left(\sigma_{Y_{i_1}|X}^{2} \sigma_{Y_{i_2}|X}^{2} \sigma_{Y_{i_3}|X}^{2}\right)^{2/3}} \right]$$

$$+ \sum_{i_1=1}^{M-2} \sum_{i_2=i_1+1}^{M-1} \sum_{i_3=i_2+1}^{M} \rho_{i_1i_2i_3}^{(3)} \frac{(Y_{i_1} - E[Y_{i_1}|X]) (Y_{i_2} - E[Y_{i_2}|X]) (Y_{i_3} - E[Y_{i_3}|X])}{\left(\sigma_{Y_{i_1}|X}^{2} \sigma_{Y_{i_2}|X}^{2} \sigma_{Y_{i_3}|X}^{2}\right)^{1/2}} \right]$$

$$+ \cdots + \rho_{M}^{(M)} \prod_{j=1}^{M} \frac{(Y_{j} - E[Y_{j}|X])}{\left(\sigma_{Y_{j}|X}^{2}\right)^{M-1}} \right]$$
The resulting mutual information expression is complex, but with the aid of the symbolic manipulation software MATHEMATICA, we were able to derive an exact analytical expression when the population is homogeneous and symmetric, meaning that all cumulant correlation coefficients of a given order are equal: $\rho_{i_1i_2}^{(2)} = \rho^{(2)}$, $\rho_{i_1i_2i_3}^{(3)} = \rho^{(3)}$, etc.

Consider first the case of two dependent channels, having cumulant correlation coefficient $\rho^{(2)}$. Just as before, assuming the input has the probability distribution given in (2.14) and evaluating the asymptotic behavior, we obtain

$$\lim_{x_{\text{max}} \to 0} I(X;Y_1,Y_2) = (2 - \rho^{(2)})(-q \log q) \cdot x_{\text{max}} + o(x_{\text{max}}),$$

which is maximized when $q = 1/e$. Consequently, we infer that the capacity of the two-component, common-input, vector Poisson channel equals

$$C^{(2)} = (2 - \rho^{(2)}) \frac{\lambda_{\text{max}}}{e \log 2}$$

$$= (2 - \rho^{(2)}) C^{(1)},$$

a quantity decreasing linearly with increasing correlation. At the extreme $\rho^{(2)} = 0$, we obtain the conditionally independent result; when $\rho^{(2)} = 1$, the channels are totally dependent and function like a single channel.

More generally, when we have $M$ conditionally dependent Bernoulli event generators driven by a common input, the capacity for any $M$ is achieved when the input probability
Figure 3.2: Capacity of the common input stimulation channel varies with population size $M$ and with the dependence parameters. The capacities for $M = 2, 3, 4, 5$ are plotted as a function of $\rho^{(2)}$. The vertical axis is the capacity normalized by the capacity $C^{(1)}$ of a single neuron. The broad spread for $M \geq 3$ occurs because dependencies of order higher than two are present in these situations; the range of capacity values for each $\rho^{(2)}$ represent how much capacity can vary as the other coefficients range over their allowable values.

is $q = 1/e$, and equals

$$C^{(M)} = \left( M - \sum_{k=2}^{M} \binom{M}{k} (-1)^k \rho^{(k)} \right) C^{(1)}.$$ 

The capacity for several values of $M$ is plotted in Figure 3.2. Because of the pecking order established by the inequality relationships among cumulant correlation coefficients, the capacity ranges between $MC^{(1)}$ (when all the cumulant correlation coefficients are zero) and $C^{(1)}$, which occurs when the cumulant correlation coefficients all equal one, modeling a completely redundant population (each component has exactly the same event pattern as all the others). In between these extremes, the capacity generally decreases as the population’s correlation coefficients increase.
If we fix the value of $\rho^{(2)}$, we can find the minimum and maximum capacity over all allowable values of the higher order cumulant correlations. The solution lies on the boundary of cumulant correlations' constraint region, and thus can be found directly from the expressions in (2.11). For example, when $M = 3$ we get

$$C_{\text{max}}^{(3)}/C^{(1)} = 3 - 2\rho^{(2)} \quad (\rho^{(3)} \to \rho^{(2)})$$

$$C_{\text{min}}^{(3)}/C^{(1)} = \begin{cases} 
3 - 3\rho^{(2)}, & 0 \leq \rho^{(2)} \leq \frac{1}{2} \quad (\rho^{(3)} \to 0) \\
2 - \rho^{(2)}, & \frac{1}{2} < \rho^{(2)} \leq 1 \quad (\rho^{(3)} \to 2\rho^{(2)} - 1)
\end{cases}$$

Similar results for larger populations can be seen in Figure 3.2. In general, the maximum capacity for a given value of $\rho^{(2)}$ is $C^{(M)} = (M - (M - 1)\rho^{(2)})C^{(1)}$, occurring when the cumulant correlations of every order are equal. It is particularly interesting that constraining the higher-order correlations decreases the capacity; for example, if $\rho^{(k)} = 0$, $k > 2$ and $\rho^{(2)} = 1/(M - 1)$ (the largest allowable value for pairwise dependence when all higher order dependencies are zero), capacity equals $(M/2)C^{(1)}$, half of its maximal value. Thus, for large populations with a single common input, small pairwise correlations can dramatically reduce capacity.

### 3.1.3 Z Channel Equivalence

If we take a slightly different view of the common input channel, we will be able to see intuitively why the inter-neuronal dependencies affect the capacity. Figure 3.3(a) shows the
same common input channel as before, except here we have mapped the vector Bernoulli output alphabet to a scalar (symbolic) one, where the number of output symbols is $2^M$. The arrows show the possible outcomes for each possible input value. From this diagram, it is immediately evident that the non-zero input can be reliably decoded (with probability 1) from any output other than the all-zero output symbol. In other words, if at least one neuron in the population produced a spike in a given bin, the input must have been non-zero. Consequently, we can collapse the output alphabet into two symbols, yielding a channel model known as the Z channel [15]. The crossover probability, which is the probability that a non-zero input gets "flipped" by the channel, is given by

$$p_c = (1 - x_{\text{max}})^M (1 + \alpha(\rho)),$$

where $\alpha(\rho)$ is a non-negative function of the cumulant correlation coefficients, given in the Sarmanov-Lancaster expansion (equation (2.16)). The capacity of the Z channel can be expressed in terms of the crossover probability:

$$C = \frac{\log \left( 1 + (1 - p_c) \frac{p_c}{1 - p_c} \right)}{\log 2} \quad (3.5)$$

As shown in Figure 3.3(b), this quantity is a strictly decreasing function of $p_c$. Thus, the capacity of the extracellular stimulation channel is maximized by minimizing its effective crossover probability, which is achieved when the neurons are independent ($\alpha(\rho) = 0$). In
Figure 3.3: The Z channel representation of the extracellular stimulation channel. (a) The output alphabet is collapsed to form the Z channel. (b) Capacity of the Z channel varies with the crossover probability, $p_c$. Minimizing $p_c$ maximizes the capacity; in the stimulation channel, this corresponds to having statistically independent neurons.
that case, $p_c = (1-x_{\max})^M$, and the capacity is $C^{(M)} = MC^{(1)}$. At the other extreme, when the neurons are maximally correlated, $p_c = 1-x_{\max}$ and consequently the channel behaves like a single neuron channel: $C^{(M)} = C^{(1)}$.

3.2 Direct Stimulation

3.2.1 Independent Neurons

Targeting specific individual neurons to stimulate is a common goal in the design of BMIs, as the ability to select individual targets for stimulation enables more precise modulation of the neural response, which could ultimately lead to better perceptual outcomes [43]. The channel depicted in Figure 3.1(b) is an idealized model of this scenario, which we term the direct stimulation channel to reflect the ideal case of direct stimulation of each individual neuron without interference or crosstalk.

When there is no inter-neuronal dependence, the basic properties of mutual information show that the population mutual information equals the sum of the individual mutual informations for each neuron. Expressing mutual information as a difference of entropies, we have

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

$$= \sum_{m=1}^{M} H(Y_m) - \sum_{m=1}^{M} H(Y_m \mid X_m)$$

$$= \sum_{m=1}^{M} I(X_m;Y_m),$$
and hence, the total capacity of the system equals the sum of the individual capacities:

\[ C^{(M)} = \sum_m C^{(1)}_m. \]

In other words, when independent, identically constrained neurons are independently stimulated, the capacity is the same as when they are stimulated by a single, common input. Thus, in terms of capacity (and consequently, minimum theoretical distortion), there is no advantage to stimulating independent neurons individually: one electrode is as good as many.

3.2.2 Dependent Neurons

A much more complicated situation arises in the presence of inter-neuronal dependencies, which makes the outputs of the channel conditionally dependent. To evaluate that case, we turn again to the vector Bernoulli model of (2.16), and as before consider the homogeneous symmetric case. Symmetry considerations suggest that the mutual information is maximized by identically distributed inputs, which we take to equal the bi-valued distribution expressed by (2.14). Beginning with the simplest case, \( M = 2 \), if we constrain the inputs to be statistically independent, we obtain the limiting expression for mutual information,

\[
\lim_{x_{\text{max}} \to 0} I(X;Y) = 2q \left[ q(1 - \rho^{(2)}) \log (1 - \rho^{(2)}) - (1 - q\rho^{(2)}) \log (1 - q\rho^{(2)}) - \log q \right] x_{\text{max}} + o(x_{\text{max}}).
\]
Figure 3.4: The capacities for $M = 2, 3, 4$ are plotted as a function of $\rho^{(2)}$ for the separate-input channel (intracellular stimulation). The vertical axis is the capacity normalized by its zero-dependence value ($MC^{(1)}$). The broad spread for $M = 3$ and $M = 4$ occur because dependencies of order higher than two are present in these situations; the range of capacity values for each $\rho^{(2)}$ represent how much capacity can vary as the other coefficients range over their allowable values.

To calculate the capacity as a function of $\rho^{(2)}$, we need to maximize with respect to $q$. Evaluating the derivative of the mutual information with respect to $q$ results in a transcendental equation for $q_C$, rendering impossible an analytic expression of the result. However, we can compute the maximum using numeric optimization, and the results are plotted in Figure 3.4. Somewhat surprisingly, the capacity increases with $\rho^{(2)}$, equaling 1.43 times its $\rho^{(2)} = 0$ value when $\rho^{(2)} = 1$. The maximizing value of $q$ also changes, going from $1/e$ in the independent case to 0.575 when $\rho^{(2)} = 1$.

For larger populations, the results are qualitatively similar. Mutual information increases as correlation increases, the opposite behavior of the common-input cases wherein dependence decreases capacity. In more detail, we find that values for $q_C$ and capacity de-
pend on the dimension $M$ of the vector channel. In the separate-input case, the percentage increase of the capacity at maximal correlation increases with the number of neurons; for example, the increases for $M = 2, 3, 4, 5$ are 43.3%, 67.5%, 78.2%, and 83.2% respectively. As $M \to \infty$, calculations show that the maximizing value of $q$ approaches $1/2$. In that case, the capacity of the maximally correlated stimulation channel with independent inputs approaches $M \lambda_{\text{max}}$, or $e \log 2$ times the capacity of a single channel: an increase of 88.42% over the independent neuron case.

In the preceding capacity calculations, we constrained the input signals to be statistically independent, and found that the capacity increases with inter-neuronal dependence. To consider the intracellular stimulation channel in its full generality, we must relax the independent input constraint and allow the inputs to be correlated. Once again, we turn to the scalar version of the vector Bernoulli channel, which is depicted in Figure 3.5 for the case $M = 2$. We can simplify the mutual information calculation by noting that, due to symmetry, the maximizing input distribution will have equal probability for the $(0,1)$ and the $(1,0)$ symbols. Consequently, $P_X(X)$ can be written in terms of only two parameters. The resulting expression for mutual information is complicated, but does yield an analytical maximum, which is plotted in Figure 3.6(a). In general, the capacity achieving input is correlated, and the resulting capacity is always higher than the independent input case.

When the neurons are maximally correlated, the channel simplifies as in Figure 3.5(b). In that case, for any value of $M$, the capacity achieving input has 0 probability of producing
The scalar version of the intracellular stimulation channel for $M = 2$ neurons.

(a) When both inputs are non-zero, the channel crossover probabilities depend on the inter-neuronal correlation $\rho^{(2)}$. (b) When the neurons are maximally correlated ($\rho^{(2)} = 1$), the channel simplifies greatly.

all zeros, and equal probabilities for all other input symbols:

$$P(X) = \begin{cases} 
0, & \sum_m X_m = 0 \\
\frac{1}{2M-1}, & \text{else.}
\end{cases}$$

The entropies $H(X)$ and $H(X \mid Y)$ are then straightforward to compute:

$$H(X) = \log(2^M - 1)$$

$$H(X \mid Y) = (1 - x_{\text{max}}) \log(2^M - 1).$$
Figure 3.6: The capacity of the intracellular stimulation channel. (a) Capacity of the two-neuron stimulation channel is plotted as a function of the inter-neuronal correlation $\rho^{(2)}$. The dashed line shows the capacity when the input signals are constrained to be statistically independent and the solid line shows the capacity when the inputs are permitted to be correlated. When $\rho^{(2)} > 0$, correlated inputs achieve a strictly greater capacity than independent inputs. (b) The normalized capacity of the stimulation channel with maximal inter-neuronal correlation ($\rho = 1$) is shown as a function of the number of neurons $M$. As $M \to \infty$, the capacity saturates at a value of $e \log 2 \cdot MC^{(1)}$. 
The capacity is thus

\[ C^{(M)} = \frac{\log(2^M - 1)}{\log 2} x_{\text{max}} \]

\[ = e \log(2^M - 1) C^{(1)}, \]

which approaches a maximum value of \( e \log 2 \cdot M C^{(1)} \) as \( M \to \infty \). Thus, while independent inputs result in a lower capacity than dependent inputs, they both approach the same limiting value.

### 3.3 The General Model

The two stimulation channel models that we have already studied represent two extremes of stimulation BMIs. In the intracellular case, we assumed that each neuron in the population could be stimulated independently, and from the point of view of capacity, this does represent the ideal case. Unfortunately, there are significant challenges to achieving this in practice, including constraints on the materials and technologies used in constructing microelectrode arrays, clinical safety considerations, and the resulting limitations on the proximity of electrodes to their target neurons and the stimulation charge densities [12].

At the other extreme, in the extracellular case we assumed that a single electrode could stimulate a population of neurons identically, and when the neurons are conditionally independent, we found that the single electrode strategy could achieve the same performance as the intracellular strategy. However, this situation is also difficult to achieve in practice, as
the current delivered to each neuron depends on its proximity to the charge injection site.

Moreover, typical microelectrode arrays place tens or even hundreds of microelectrodes in close proximity, which can cause a single neuron to receive stimulation from more than one electrode [49]. Thus, a more practically relevant model of stimulation BMIs should allow for multiple, correlated inputs.

Beginning with the common input case, we can modify the channel model by inserting an attenuation factor $a_m$ for each neuron’s event probability (Figure 3.7(a)):

$$P(Y_m \mid X) = \begin{cases} a_mX & Y_m = 1 \\ 1 - a_mX & Y_m = 0 \end{cases} \quad 0 \leq a_m \leq 1$$
To find the capacity of this attenuated channel, we note that the rate constraint in our definition of capacity (2.6) is a constraint on the channel, not the input. Indeed, in most of the capacity calculations, there is no loss of generality by considering the rate constraint as inducing a constraint on the input signal, as we have done. However, in considering the attenuated channel, when the input itself is not directly constrained, the capacity achieving strategy is immediately evident: Increase the amplitude of $X$ so that $a_m X \geq x_{\text{max}}$, $m = 1, \ldots, M$, and let the neurons clip at input values greater than $x_{\text{max}}$. In that case, the capacity is unchanged by the inclusion of attenuation.

In practice, arbitrarily increasing the amplitude of $X$ may not always be possible, as the maximum strength of the input stimulation is usually limited by safety considerations [55]. Including a direct constraint on the maximum input, we require as before that the input’s value ranges over the interval $[0, x_{\text{max}}]$; this could reflect a safety measure that limits the input to a level that, unattenuated, would elicit maximum firing rate in each neuron. In that case, the capacity-achieving input distribution is again bi-valued as expressed by (2.14), with $qC = 1/e$ regardless of the values for the cumulant correlation coefficients and the attenuations. The capacity equals

$$C^{(M)} = \left( \sum_{m=1}^{M} a_m - \sum_{k=2}^{M} (-1)^k \sum_{m_1 = k, \ldots, m_k = k}^{M} \sqrt[1]{a_{m_1} \cdots a_{m_k} \rho_{m_1, \ldots, m_k}} \right)c^{(1)},$$

where the cumulant correlation coefficients must be non-negative but obey a much more complicated version of the inequalities in (2.11). The combination of correlation coef-
ficients and attenuations that maximize capacity occur when the channels are identical
and conditionally independent: $a_m = 1, \rho^{(k)} = 0$. We conclude that statistical dependence
among neurons always decreases capacity for the common input, extracellular stimulation
channel.

A fully elaborated model of the stimulation channel is shown in Figure 3.7(b). Here, a
total of $K$ input signals are combined to form the inputs to $M$ neurons. Letting $X$ be a $K \times 1$
vector of inputs, the Bernoulli input probabilities are

$$X' = AX,$$

where $A$ is an $M \times K$ path gain matrix having bounded entries $0 \leq a_{mk} \leq 1$. We define
$a_m$ to be the row vector corresponding to the $m$th row of $A$. This model encompasses all
of the models we have discussed in the preceding sections. When $A = [1 \ 1 \ \cdots \ 1]_{1 \times M}$, we
obtain the single input extracellular stimulation channel. When $A = I_M$, we obtain the direct
stimulation channel. In general, the path gain matrix can consist of random or deterministic
entries, whose values could be known or unknown by the encoder or the decoder.

When the neurons are independent, we note once again that, absent an explicit con­
straint on the inputs, the capacity of the channel has a simple solution: $C^{(M)} = MC^{(1)}$, achieved when:

- The inputs are completely dependent (i.e. one single input)

- The amplitude of the input is high enough such that $X_m' = a_m X \geq x_{\text{max}}$, $m = 1, \ldots, M$. 
The second condition can always be achieved, whether $A$ is deterministic or random, by making the input signal amplitude equal to $M_{\text{max}}$. However, as we already noted, safety considerations usually dictate an explicit amplitude constraint on the input signal, which could make the above scheme impossible. Furthermore, when the population being stimulated has inter-neuronal dependencies, the single input does not achieve the capacity bound.

Recent work in wireless optical communications examined the capacity of a related channel, the shot noise limited multiple input multiple output (MIMO) optical channel [8, 26]. That model, like ours, consists of $K$ signals that are mixed via a path gain matrix to form the inputs to $M$ statistically independent Poisson event generators. In addition, each Poisson generator sustains a constant background noise rate, akin to a minimum rate constraint in our neural channel model. In that case, when the individual input signals are subject to peak and average constraints, they found that the capacity scales as $K \cdot M$. The fact that the capacity of the MIMO optical channel scales with the number of inputs as well as the number of outputs highlights a key difference in their model: Although they do constrain the individual inputs to the channel, they do not constrain the maximum rate of the Poisson generators themselves. Consequently, for a fixed $M$, the rate of the Poisson process output, and hence the capacity, is proportional to the number of inputs $K$. The capacity of the general stimulation channel thus remains an open problem.
Chapter 4
The Neural Control Channel

The neural control BMI is, in a sense, the complement to the neural stimulation BMI. Rather than encoding and transmitting a stimulus to the brain, the purpose of the control BMI is to receive and decode a goal or intended action from the brain (Figure 1.1(b)). For example, microelectrodes implanted in the motor cortex have been used to record spike trains from hundreds of neurons in animal and human subjects. This neural activity can be decoded to direct a computer cursor, effectively restoring function to patients with tetraplegia or ALS, or even to control a prosthetic limb [27].

In general, the same factors that affect stimulation BMI performance — number of electrodes and their physical characteristics, what encoding/decoding are used, patient history, etc. — also affect the performance capabilities of the control BMI. In fact, the stimulation models discussed in Chapter 3 can also serve as an idealized model for the control scenario: If the signals from each individual neuron in a population could be recorded reliably and without error (most likely via intracellular recording), then the capacity of the two systems would be the same. However, in most practical situations, obtaining reliable intracellular recordings is either infeasible or impossible. Rather, BMIs for control prostheses rely on extracellular recordings from one or more microelectrodes or, in some cases, even electroencephalographic (EEG) signals and local field potentials (LFP), which arise from the
simultaneous activity of many disparate neurons [1]. We elaborate the population channel models of the previous section to account for these effects.

4.1 Extracellular Recording

4.1.1 Single Electrode

Most neurally controlled prosthetics use aggregated recordings of simultaneous activity of many neurons. While techniques for separating such recordings into their constituent signals exist, in many practical cases of interest extracellular recordings are not or can not be teased apart into individual neural activities; instead, the summed activity is used as a surrogate for population activity [1]. To model this situation we sum the outputs of the baseline population to produce a single output (Figure 4.1):

$$Y = \sum_{m}^{M} Y_m.$$

Using our Bernoulli approximation approach, we can determine to what extent considering only aggregate behavior reduces the capacity of a population. When the population is homogeneous and symmetric, the probability distribution of the summed output conditioned
Figure 4.1: The single electrode recording channel model is shown. Here, the population response is summed into a single aggregate signal.

\begin{align*}
P(Y|m_{nz}) &= \binom{m_{nz}}{Y} x_{\text{max}}^Y (1 - x_{\text{max}})^{m_{nz} - Y} \\
& \quad \cdot \left[ 1 + \sum_{k=2}^{m_{nz}} \rho^{(k)} \sum_{l=0}^{k-1} \binom{m_{nz} - Y}{k-l} \binom{Y}{l} (-1)^{k-l} x_{\text{max}}^{1-l} (1 - x_{\text{max}})^{l-k+1} \right], \ Y = 0, \ldots, m_{nz}.
\end{align*}

Figure 4.2 shows the possible input-output configurations of the channel. When the neurons are conditionally independent, the capacity has a simple solution. In that case, setting the inputs to be completely dependent (i.e., a single common input) results in a Z channel (Figure 4.2(b)), with crossover probability $p_c = (1 - x_{\text{max}})^M$. Consequently, the capacity is given by (3.5), and equals $MC^{(1)}$, the same as for the individual output (intracellular recording) case. Thus, given the right inputs, aggregating the outputs of a population with no inter-neuronal correlations does not have to incur a decrease in performance.

At the other extreme, when the neurons are completely dependent, the channel has the form shown in Figure 4.2(c). To find the capacity, we need to find the input distribution
Figure 4.2: The scalar version of the single electrode (aggregate) recording channel is shown for $M = 2$ neurons. (a) In general, the crossover probabilities depend in a complex way on the inter-neuronal correlation $\rho^{(2)}$. (b) When the neurons are independent, the capacity is achieved by treating the channel as a Z channel. (c) When the neurons are maximally correlated, some of the crossover probabilities are zero, simplifying the channel from the general case.

that maximizes the mutual information in that case. By symmetry, we can assume that the input distribution has the form

$$P_X(X) = P \left[ \sum_m X_m = m_{nz} \right] = \begin{cases} 
q_0, & m_{nz} = 0 \\
q_1, & 0 < m_{nz} < M \\
1 - q_0 - q_1, & m_{nz} = M.
\end{cases}$$

Calculating the mutual information and taking the limit as $x_{\text{max}} \to 0$, we obtain

$$\lim_{x_{\text{max}} \to 0} I(X;Y) = \left( - (1 - q_0 - q_1) \log(1 - q_0 - q_1) - q_1 \log \left( \frac{q_1}{M-1} \right) \right) x_{\text{max}} + o(x_{\text{max}}).$$

When $M = 2$, the maximizing input distribution has $q_0 = 1/e, q_1 = 1/e$, and consequently the capacity is $C^{(2)} = 2C^{(1)}$. For $M > 2$, the maximizing input distribution has $q_0 = 0,
The natural text is as follows:

\[ q_1 = 1 - 1/M, \] and the capacity is

\[ C^{(M)} = (e \log M)^{C^{(1)}}. \] (4.1)

This result is plotted in Figure 4.3. Thus, when the neurons are maximally dependent, aggregating the outputs in a single recording results in a huge capacity loss as \( M \) gets large.

We can also compute the capacity when the inputs are constrained to be statistically independent. Although this is not the capacity achieving input in general, imposing this constraint allows us to compare directly with the unaggregated case, all other things being equal. When the inputs are independent, the input distribution has a binomial form,

\[ P_X(X) = P \left[ \sum_m X_m = m_{nz} \right] = \binom{M}{m_{nz}} q^{m_{nz}} (1 - q)^{M - m_{nz}}, \]

where \( q \) is the probability of an input being non-zero. The resulting mutual information expression is complex, but can be optimized numerically. As in the unconstrained case, the maximum mutual information depends on the cumulant correlation coefficients, with dependence resulting in higher capacity than when the coefficients are zero (Figure 4.3). As the size of the population grows, the aggregated-output capacity differs more and more from unaggregated values. In the worst case considered, that of independent neurons and independent inputs, the capacity of the aggregate channel for any size population is substantially less than that of two neurons having separate, unaggregated outputs, approaching
Figure 4.3: The normalized capacity $\frac{C^M}{C^{(1)}}$ of the aggregated population grows at a different rate depending on whether the population is conditionally independent ($\rho = 0$) or completely dependent ($\rho = 1$), and whether or not the inputs are constrained to be independent. When the population is conditionally independent but the inputs are permitted to have dependencies, the capacity of the aggregated channel is $MC^{(1)}$, the same as the unaggregated channel. Dependencies between neurons, and constraints on the inputs can only decrease capacity from that baseline.
Figure 4.4: The multi-electrode recording channel. Each of $L$ electrodes records the summed activity of a subpopulation of the $M$ neurons. In general, subpopulations may overlap, and may not contain equal numbers of neurons.

a maximum of $1.58C^{(1)}$. Consequently, depending on the form of the inputs, not separating an aggregate recording into its constituents can greatly reduce the information that can be gleaned.

4.1.2 Multiple Electrodes

The sharp capacity decrease sometimes incurred by aggregating the neural outputs in a single recording can be mitigated by forming multiple aggregations (Figure 4.4), each of which is obtained from a subpopulation; this situation can be seen as an idealized model for unsorted multi-electrode recordings. Assume we aggregate outputs from $L$ equal-sized subpopulations that have overlapping membership in the conditionally independent case. We could only compute the case where the inputs are independent. In that case, calculations show that the subpopulation size that maximizes capacity is $M \cdot L / (2L - 1)$ whereas equal non-overlapping subpopulations have a maximal size of $M/L$. Thus, recorded subpopula-
tions must overlap substantially to maximize capacity. For large populations with no inter-neuronal dependencies, the capacity approaches \((2L - 1)1.58C^{(1)}\), indicating that multiple aggregated recordings can greatly increase capacity, equaling the single-aggregation capacity (for independent inputs) multiplied by a factor of about twice the number of recordings. This asymptotic result breaks down when the factor \((2L - 1)1.58\) approaches \(M\).

When the neurons are conditionally dependent, we can bound the capacity when the subpopulations are equal and non-overlapping. Assume that \(\frac{M}{L} \in \{3, 4, \ldots\}\), and that the neurons within each subpopulation are maximally dependent. Then, generalizing (4.1), we obtain

\[
C^{(M)} = eL \log \frac{M}{L} C^{(1)}
\]

Thus, for example, when one third as many electrodes as neurons are used \((L = M/3)\), the capacity can be at least \(\left(e^{\frac{1}{3}\log 3}\right)MC^{(1)}\), a substantial improvement over the the single electrode case.

### 4.2 Spike Sorting

As we saw in Section 4.1.1, aggregating the output of a population of neurons in a single electrode recording can effect a significant decrease in the system’s capacity. Thus, separating the constituent response signals from the gross recording — a technique known as spike sorting — may be required for effective use of control BMIs. A variety of spike sorting algorithms have been devised using techniques such as template matching and principal
Figure 4.5: The spike sorting channel. A single or multi-electrode recording is processed to produce estimates of the neural spike trains. Depending on the quality of the recordings and the accuracy of the spike sorting algorithm, the estimated signals could contain mislabeled, inserted, or deleted spikes.

components analysis, with each method suffering different detection and classification error rates [40].

To consider the effect of spike sorting on the overall capacity of the system, we pass the vector Bernoulli process modeling the population output through a spike sorting channel, depicted in Figure 4.5. Here, a single or multi-electrode recording is processed to produce estimates of the individual spike trains from each neuron in the population. Conceptually, the spike sorting channel "corrects" for information lost in the aggregation channel by using information such as amplitude differences and history dependence, which the aggregation channel decoder ignores. However, the spike sorting algorithm is limited by the available cues and external influences such as interference and noise; consequently, the spike sorting channel introduces errors of its own, resulting in a lower capacity than would be obtained if the decoder had direct access to the neural spike trains.

Since the output of the cascade is also a vector Bernoulli process, we can find the ca-
capacity of the combined channel using the same techniques we used for finding population capacities. We assume that errors incurred during spike sorting are independent of the neural population channel; consequently, the spike sorting channel transition probabilities multiply the neuron channel probabilities, again yielding a vector Poisson process output in the small Bernoulli probability limit. We further assume that sorting errors occur independently for each neuron and with equal probability. Figure 4.6 shows the scalar version of the spike sorting channel for three different types of sorting errors. We focus here on the case $M = 2$, but the approach generalizes easily to larger populations.

In the first case, shown in Figure 4.6(a), the spike sorter commits errors in labeling single spikes: With probability $p_{ml}$, a spike from neuron 1 is mistakenly said to have come from neuron 2, and vice versa. The capacity of the resulting channel is shown as a ratio to the capacity of the optimal channel, wherein the decoder has perfect knowledge of the original spike trains. Note that the capacity ratio for each channel is computed with respect to the optimal capacity of its particular equivalent neural channel having no errors. Hence, capacity ratios for two different channels can be compared to determine the relative effects of spike sorting on those systems, but cannot be used to compare absolute performance. Four cases are shown: the neurons are either independent ($\rho^{(2)} = 0$) or maximally dependent ($\rho^{(2)} = 1$), and the inputs are either independent or the input dependence is unconstrained. When the neurons are independent, the channel can be treated as a Z channel, with the crossover probability unaffected by the labeling errors. Thus, the capacity is equal to the optimal capacity, regardless of the error probability $p_{ml}$. However, when the inputs
Figure 4.6: Spike sorting channels with mislabeled and deleted spikes for $M = 2$ neurons. The right side of each panel shows the capacity of the channel shown on the left as a ratio to the capacity of the equivalent neural channel with no sorting errors. (a) Spikes from one neuron are mislabeled as having been produced by the other neuron, with probability $p_{ml}$. (b) Spikes from each neuron are deleted with probability $p_{ds}$ due to detection errors.
are independent, labeling errors can result in a substantial decrease in capacity, with the capacity ratio decreasing to a worst-case value of around 0.6. When the neurons are maximally dependent, a similar decrease in capacity from the optimum results, regardless of the input correlation.

In the second spike sorting channel, depicted in Figure 4.6(b), spikes are randomly deleted with probability $p_{ds}$, modeling detection errors that might result from noise or attenuation in the recorded signals. When the neurons are uncorrelated, the capacity is achieved with independent inputs, and the capacity ratio decreases linearly with increasing deletion probability $p_{ds}$. The effects are more severe in the dependent neuron case when $p_{ds}$ is small; however, the capacity ratios for both independent and dependent inputs have a higher worst-case value of around 0.65.

The third type of spike sorting channel is depicted in Figure 4.7. Here, all spikes are correctly detected and sorted, but false positives (extra spikes) are randomly inserted into each spike train. In order that the resulting channel output converge to a Poisson process, the insertion probability in each discrete-time bin must be proportional to the bin-width; consequently, the spike insertion channel is modeled simply by adding an independent Poisson process having rate $\lambda_0$ to each neuron's output.

The resulting capacity ratio is shown on the right side of Figure 4.7. When the neurons are independent, if the insertion process rate $\lambda_0$ is high relative to the neurons' maximum firing rate $\lambda_{max}$, the capacity goes to zero. Conversely, when $\lambda_{max} \gg \lambda_0$, the capacity approaches the optimal capacity. Interestingly, when the neurons are maximally correlated,
Figure 4.7: The spike insertion channel. The spike sorter erroneously inserts spikes when none were produced. In continuous-time, the inserted spikes are modeled as independent Poisson processes with rate $\lambda_0$ added to the neuron outputs. For independent neurons ($\rho = 0$), when the insertion process rate is high relative to the maximum rate $\lambda_{\text{max}}$ of the neurons, the resulting capacity goes to zero. However, when the neurons are maximally correlated ($\rho = 1$), the capacity never decreases below the capacity of a single neuron, regardless of the value of $\lambda_0$.

The capacity never decreases below the capacity of a single neuron, regardless of the value of $\lambda_0$. In that case, the capacity is achieved by a single common input, so that at any instant, either all neurons produce a spike, or none do. Since the probability of the every neuron's insertion process producing an event at the same instant is proportional to $(\Delta t)^M$, the resulting channel output can be decoded as well as for a single neuron with no insertions.

We conclude that the success of spike sorting for neural control depends heavily on the type and frequency of errors that are committed; the most potentially damaging errors are false positives, which severely degrade the capacity when the insertion rate is high relative to the firing rate of the neurons. In addition, it is important to note that maximizing capacity for the spike sorting channel does not necessarily preserve the statistics of the original spike...
trains. This non-intuitive result suggests that different strategies should be employed for spike sorting during experimentation versus spike sorting for prostheses. While accurate spike train reconstruction is important for experimental analysis, for neural control applications it may be better to intentionally bias the spike sorter to control one error rate over another, thereby increasing the capacity and improving the device's best-case performance. However, which type of errors to favor depends on the input and dependence characteristics of the neurons, which may or may not be known.

4.3 The General Model

As we have seen, our spike sorting channel model allows us to account directly for detection and classification errors that arise from noise and interference in the spike train recordings. In contrast, our aggregation model for unsorted extracellular recordings assumed that recordings are noise-free, and that all neurons in the population are recorded with equal amplitudes at the electrode. In practice, however, this is almost never the case [2, 47]. To connect these models to more real-world situations, we need to extend the analysis to allow for non-ideal recording situations. Figure 4.8 shows a general model for neural recording, in which $L$ electrodes receive signals from $M$ neurons, some or all of which may be subject to errors arising from cross-talk and attenuation in the conductive medium surrounding the target neurons. In addition, noise signals $N_1, \ldots, N_L$ get added to each aggregate; this models the noise that arises from many additive spike trains from distant cells, as well as electrical noise induced on the electrode or other recording apparatus itself [12].
Consider the simplest example, recording from a single neuron with white Gaussian noise $N$, where the channel input has the form given in (2.14). The channel conditional probability distribution is a Gaussian mixture:

$$p_{Y|X}(y|x) = \begin{cases} p_N(y; \mu = 0, \sigma^2_N), & x = 0 \\ x_{\max} p_N(y; \mu = 1, \sigma^2_N) + (1 - x_{\max}) p_N(y; \mu = 0, \sigma^2_N), & x = x_{\max} \end{cases}$$

where $p_N(y; \mu, \sigma^2_N)$ is the Gaussian density function with mean $\mu$ and variance $\sigma^2$. Computing the capacity thus requires calculating the entropy of a Gaussian mixture, which we can do computationally in MATHEMATICA, and the result is plotted in Figure 4.9. When the signal-to-noise ratio (SNR) is high enough, the capacity of the noisy aggregate is close to the capacity of the noiseless aggregate. However, there is a sharp drop-off in capacity when the SNR falls below roughly 15 dB, and for lower values of SNR the capacity of the system
Figure 4.9: Normalized capacity for a single neuron with Gaussian noise. There is a sharp drop in the capacity when the SNR is less than around 15 dB.

is almost negligible. While in some situations, well-isolated neurons may be recorded with 15 dB or more SNR [12], many BMIs must rely on SNRs that are much lower, particularly if using less differentiated signals such as EEG and LFP [10]. These results reinforce the need to improve the recording conditions, either by increasing the number of electrodes, running advance spike detection and sorting routines, and minimizing noise through proper electrode design and surgical implantation.

The capacity calculations can be extended to a population of neurons, however computing the required entropies becomes difficult as $M$ increases. The capacity of the general control BMI channel remains an open problem.
Chapter 5

From Theory to Practice

In Chapters 3 and 4 we derived the capacity for BMI devices, and showed how the capacity changes depending on the method of stimulation or acquisition used, as well as the size and characteristics of the neural population in question. As we noted in Chapter 2, capacity determines the maximum fidelity of information about a channel’s inputs that can be extracted from its outputs. More precisely, given a source signal and a distortion measure, the Source Channel Separation Theorem tells us that the point at which the source rate-distortion function equals the BMI capacity defines the minimum achievable distortion. This result holds for any stationary ergodic source and any distortion function. Because all rate-distortion curves are strictly decreasing and convex, increasing capacity always means reducing the smallest achievable average distortion. Consequently, all other things being equal, given two BMIs with different capacities, the BMI with the higher capacity is capable of better performance (lower distortion) for any source signal, regardless of how distortion is quantified. Thus, capacity provide us an absolute scale on which to compare BMI designs and to study the effects of various physical and computational constraints on BMI performance.

However, the implication of capacity differences to actual neural prostheses can only be properly understood in the context of what those prostheses are actually intended to accom-
plish. In that case, determining the ultimate capabilities of a prosthesis requires identifying a source signal that is of interest, and defining a distortion measure to characterize performance. For a cochlear implant, the perceptual quality of speech signals could be measured with a psychoacoustic model of speech processing, such as one used for evaluating speech codecs for telecommunication systems [51]. For a prosthetic limb, an intended motion defined by target coordinates and velocities might be compared to the actual motion produced by the artificial limb in terms of their mean squared difference [66]. In this chapter, we develop several simple examples of sources and distortion measures that are relevant to the study of neural prostheses, and show how the BMI capacities derived in the preceding chapters can be applied to study prostheses in terms of the optimal performance theoretically attainable (OPTA) [3, Chap. 5]. We also discuss other issues related to achievability in practice, including the delay and complexity of practical systems and the use of feedback.

5.1 Minimum Distortion: The Optimal Performance of BMIs

As we did for capacity in Chapter 2, we begin by modifying the definition of rate-distortion given in (2.4) to account for continuous-time sources. Let \( S_t \) be a stationary, ergodic process, and let \( d(s, \hat{s}) \geq 0 \) be a distortion measure. The rate-distortion function is defined as

\[
R(D) = \lim_{T \to \infty} \min_{p(\hat{s}|s): \mathbb{E}_T[d(s, \hat{s})] \leq D} \frac{1}{T} I(S_{0 \leq t < T}; \hat{S}_{0 \leq t < T}).
\]
The rate thus has units of bits/s (or nats/s, if the natural logarithm is used). Note that the distortion is averaged over the observation interval:

\[ E_T [d(s, \hat{s})] = \frac{1}{T} \int_0^T E [d(s_t, \hat{s}_t)] dt. \]

When the Source Channel Separation Theorem holds*, the OPTA is the minimum achievable distortion, given by the point at which the rate-distortion function equals the capacity: \( R(D_{\text{min}}) = C \). Thus, to find the OPTA for a given neural prosthesis model, we need to calculate the rate-distortion for the applicable source and distortion measure. Rate-distortion functions are difficult to calculate in general; however there are a variety of analytical and computational tools available, depending on the particular source and distortion functions in question [3, 4, 11, 48]. We provide here four examples of sources and associated distortion measures for neural prostheses, and compute the OPTA for each under the various BMI models discussed previously.

5.1.1 Example 1: Auditory Prosthesis

A simple example of a source stimulus for an auditory prosthesis is the band-limited Gaussian random process having bandwidth \( W \) and power \( P \), where the distortion measure is the mean-squared error. Although the Gaussian source is not a good model for natural sounds, and therefore may not be directly relevant to auditory processing, it remains a commonly

*As discussed in Chapter 2, the Separation Theorem is broadly applicable to a variety of situations, including stationary, ergodic sources and channels [61]. A discussion of when the theorem does and does not hold is outside the scope of this thesis.
used stimulus in studies of the auditory system [50]. Furthermore, the rate-distortion function for the Gaussian source is an upper bound on the rate-distortion for any source with the mean-squared error distortion measure [3, Chap. 4]. The rate-distortion function for this source is well known [3, Chap. 4]:

\[
\mathcal{R}(D) = \begin{cases} 
W \log \frac{P}{D}, & D \leq P \\
0, & D > P 
\end{cases}
\]

Note that the maximum distortion here is \( D_{\text{max}} = P \). Zero rate means that nothing about the signal is encoded, leaving the decoder to make an intelligent guess based on the input’s properties; for the mean-squared error distortion measure, using the signal’s expected value as the estimate minimizes the data-ignorant decoder’s distortion. The OPTA has a simple closed-form expression:

\[
D_{\text{min}} = D_{\text{max}} e^{-c/W}.
\]

Inserting our expression for the capacity of the BMI channel, we obtain

\[
D_{\text{min}} = D_{\text{max}} \exp \left\{ -\frac{\alpha(\rho) M \lambda_{\text{max}}}{e \cdot W} \right\},
\]

where the function \( \alpha(\rho) \) captures the effects of inter-neuronal dependence on the capacity of the vector channel.

Figure 5.1 shows the OPTA curves for the bandlimited Gaussian source.
Figure 5.1: The OPTA performance for a bandlimited Gaussian source with mean squared error distortion. (a) The minimum distortion (normalized by the maximum distortion) decreases exponentially with the ratio of the capacity and the Gaussian stimulus bandwidth. (b) The distortion ratio, shown here for the case $W = 500$ Hz and $\lambda_{\text{max}} = 700$ spikes/s, decreases with population size at different rates depending on the BMI characteristics. The baseline case (solid) is achieved when there are no inter-neuronal dependencies. When the neurons are correlated, the minimum distortion decreases faster with separate inputs (dotted), and slower with a single common input (dashed). The common input case shown has only pairwise inter-neuronal dependencies.
The ratio of minimum to maximum distortion decreases exponentially with the ratio of the capacity and the Gaussian stimulus bandwidth. The plot shows how large the capacity-to-signal-bandwidth ratio needs to be to produce small distortions. For example, if the capacity simply equaled the stimulus bandwidth, the best possible distortion would be no greater than about one third of that obtained by simply guessing. When the channel consists of a single neuron, capacity equals $\lambda_{\text{max}}/1.88$, which means to achieve this modest level of fidelity, the maximal rate would need to be almost twice the stimulus bandwidth. Thus, the maximal rate needs to be several times the bandwidth to obtain significant distortion reductions. Auditory-nerve fibers having a center frequency of 1 kHz have a bandwidth of about 500 Hz, and a maximum firing rate of around 700 spikes/s [34]. Consequently, a single neuron would only be capable of achieving a distortion reduction of about $1/2$. The analysis thus indicates that recording from a population is essential for accurate stimulus reconstruction.

Figure 5.1(b) shows the OPTA distortion ratio as a function of the population size, for a few different BMI models. As the plot shows, depending on the type of BMI, the number of neurons required to achieve a given level of distortion can vary significantly. For example, achieving a distortion reduction of $10^{-2}$ would require 9 independent neurons in the population. If the neurons are correlated, the population could be reduced to 5 neurons if separate inputs are used; however, if a single common input is used, even if the neurons are only pairwise correlated, a minimum of 18 neurons would be required to achieve the same level of distortion.
5.1.2 Example 2: Visual Prosthesis

To model an image source that might be presented to a visual prosthesis (e.g. a retinal implant), it is convenient to consider images in the wavelet domain. For natural images, the wavelet coefficients within a subband are well modeled by independent, identically distributed Laplace random variables [59]. We can further extend the 2D image model to encompass video by considering a sequence of images whose coefficients are assumed to be independent in time. The rate-distortion per coefficient for this source with an absolute-error distortion measure \( d(S, \hat{S}) = |S - \hat{S}| \) is given by [3, Chap. 4]

\[
R(D) = \begin{cases} 
W \log \left( \frac{\sigma}{\sqrt{2D}} \right), & D \leq \sigma/\sqrt{2} \\
0, & D > \sigma/\sqrt{2}
\end{cases}
\]

where \( W \) is the video frame rate, and \( \sigma \) is the standard deviation of the coefficients. It is easy to see that the OPTA expression is the same as (5.1) for the Gaussian source with mean-squared error distortion, and consequently the per-coefficient distortion is shown in Figure 5.1(a). For a video frame rate of 30 Hz, to achieve a distortion ratio of \( 10^{-2} \) a single neuron would need a maximum firing rate of about 375 spikes/s. Retinal ganglion cells can exhibit firing rates up to around 400 spikes/s [36]; thus, a single neuron is sufficient to provide this level of distortion.

Figure 5.2 shows the OPTA curves of various stimulation BMI models for a 12 x 12 image patch, assuming \( W = 30 \) Hz and \( \lambda_{\text{max}} = 400 \) spikes/s. Again, the population size
Figure 5.2: The OPTA performance for Laplace distributed transform coefficients of a 12 × 12 image patch. The distortion ratio, shown here for the case \( W = 30 \) Hz and \( \lambda_{\text{max}} = 400 \) spikes/s, decreases with population size at different rates depending on the BMI characteristics. The baseline case (solid) is achieved when there are no inter-neuronal dependencies. When the neurons are correlated, the minimum distortion decreases faster with separate inputs (dotted), and slower with a single common input (dashed). The common input case shown has only pairwise inter-neuronal dependencies.

required for a given performance level can be drastically different depending on the characteristics of the BMI. For example, if no inter-neuronal dependencies are present, a 200 unit population could achieve a distortion ratio of \( 10^{-3} \). With separate inputs, only half that number of neurons would be needed if they were maximally dependent. However, at least twice as many neurons would be required if a single common input was used.

5.1.3 Example 3: Brain-Computer Interface

Brain-computer interfaces (BCI) are devices that enable controlling a computer (e.g. moving a cursor, typing on a keyboard) without physically moving. The goal of such devices is to restore some communication and control functionality to people with neuromuscular dis-
orders [65]. For example, consider a virtual keyboard like the one described in [60]. The entropy rate of English has been estimated to be around 1.3 bits/letter [14], or 5.85 bits/word, assuming an average word length of 4.5 letters. Thus, error-free typing on a BCI could be achieved when the BCI capacity is \( C > \frac{5.85W}{60} \), where \( W \) is the typing speed in words/min. Typing 50 words/min would require a capacity of 4.875 bits/s, which could be achieved by a single neuron with maximum rate \( \lambda_{\text{max}} = 9.18 \). Neurons in the primary motor cortex can have maximum rates of over one hundred spikes per second [63]; thus, reliable use of a virtual keyboard is theoretically achievable by recording from a single neuron, or from a single population aggregate.

Indeed, experimental BCI systems have reported nearly error-free typing on a virtual keyboard using only scalp-recorded EEG activity [60]. There, typing speed was only around 3.8 words per minute, far below the OPTA limit predicted by our model. Performance likely suffered from recording noise and interference from non-target neural signals. Furthermore, achieving the OPTA limit requires an optimal encoder and decoder; in the case of a control BMI, the correct encoder must be "learned" by the subject. In fact, BCI performance typically improves through user training, suggesting a reorganization of cortical activity to optimize information transfer through the BMI [38, 65].

5.1.4 Example 4: Prosthetic Limb

In the last decade, there has been significant interest in the idea of decoding cortical signals to control robotic limbs, with a number of studies demonstrating BMIs that enabled
primates to control a robotic arm and perform reaching and grasping movements with remarkable accuracy [38]. In this example, we consider a simple discrete-time model of arm position along a single axis.

Let the source be memoryless, uniformly distributed in the interval $[-1, 1]$, and let the distortion measure be the absolute-error. The maximum distortion is then $D_{\text{max}} = 1/2$, and the rate-distortion function is [39]

$$ R(D) = \begin{cases} -\left(1 - \frac{D}{D_{\text{max}}}\right)^{1/2} - \log \left(1 - \left(1 - \frac{D}{D_{\text{max}}}\right)^{1/2}\right), & D \leq D_{\text{max}} \\ 0, & D > D_{\text{max}} \end{cases} $$

The OPTA does not have a closed form, but it can be expressed in terms of the Lambert $W$ function, which can be efficiently computed [13]:

$$ D_{\text{min}} = D_{\text{max}}(2W(-e^{-C/W-1}) - W^2(-e^{-C/W-1})). $$

Here, $W$ is the update frequency of the arm position. The OPTA is plotted in Figure 5.3. Figure 5.3(a) shows that the distortion decreases approximately exponentially with the ratio of the capacity to the update frequency. For example, to achieve a distortion ratio of $10^{-3}$, the capacity would have to be approximately 4.6 times the update frequency. In studies of primate BMIs, typical update frequencies are between 15 and 50 Hz [7, 56]. For an update frequency of 33 Hz, a single neuron would require a maximum firing rate of around
Figure 5.3: The OPTA performance for a memoryless uniform source with absolute error distortion. (a) The minimum distortion (normalized by the maximum distortion) decreases approximately exponentially with the ratio of the capacity to the position update frequency. (b) The distortion ratio, shown here for the case $W = 33$ Hz and $\lambda_{\text{max}} = 100$ spikes/s, decreases with population size at different rates depending on the aggregation BMI's characteristics. The baseline case is achieved when the neurons are statistically independent (solid), if either reliable intracellular recordings of each neuron are used, or if the outputs are aggregated given the optimal (dependent) inputs. If the inputs are independent, or if the neurons are correlated (dashed), the performance can be significantly worse for the same size population. For the independent input cases, results are only shown up to $M = 8$, which is the maximum we were able to compute.
286 spikes/s, more than twice the maximum rates typically observed for neurons in the primary motor cortex [63].

Figure 5.3(b) shows the OPTA performance of several different extracellular control BMI models, assuming an update frequency $W = 33$ Hz and a maximum firing rate of $\lambda_{\text{max}} = 100$ spikes/s. When aggregated recordings are used, the performance of the control BMI is highly dependent on the properties of both the neural population and the encoder. For example, error-free recording from a population of 8 independent neurons could achieve a distortion ratio of $10^{-3}$. Alternatively, a single aggregate could achieve the same performance if the encoder statistics match those given in Chapter 4 for the Z channel. However, when the neurons are correlated, at least 20 neurons would be required to achieve the same distortion. Furthermore, if the encoder produces independent inputs, the distortion ratio could never decrease below 0.1, no matter how large the population. In that case, multiple electrodes or spike sorting would have to be used to improve the BMI performance to an acceptable level.

### 5.2 Limitations of the Theory

We have seen how the complementary concepts of capacity and rate-distortion, together with a suitable model for the neural communication channel, can be used to evaluate how well information can be conveyed through a BMI. Indeed, the minimum distortion interpretation of capacity gives us an objective way of evaluating and comparing different types of BMIs by determining the best possible performance of each system given its unique
constraints. However, it is important to note the limitations of this theory as it relates to practical BMI performance.

**Non-Poisson spike trains.** Our analysis is confined to the Poisson case, whereas real neural spike trains usually exhibit memory effects [29]. A better model for the neural channel would be based on renewal processes or even Markov point processes having longer range history dependence. For a single channel, the capacity of the Poisson channel is an upper bound on the capacity for any other point process channel [35]. Consequently, the performance bounds implied by the Poisson results are still valid for neural information processing in a single neuron, although they may not be tight. Unfortunately, we have not been able to prove a similar bound for the vector point process channel.

**Achievability and delay.** Although the Source Channel Separation Theorem enables the analysis of the system’s ultimate performance limits, the theory is silent on how to achieve those limits. In general, achieving capacity may require infinite delay, and so the performance limits implied by the capacity may not be strictly relevant on a practical timescale. In that case, an analysis based on joint source-channel coding [21] or error exponents [67, 68] may offer some additional insight.

**Non-stationary inputs.** Our ergodic definition of capacity requires the inputs to be stationary. In a practical system dealing with real-world stimuli, this condition is unlikely to be met in general. Studying non-stationary inputs requires a different in-
formation theoretic approach. There are techniques in communication theory for dealing with non-stationary inputs, such as outage probability [9] and anytime capacity [54], however it is not clear that such approaches are relevant to the neural communication problem.

The role of feedback. Although certainly of great practical relevance, we have ignored the existence and use of feedback in our analysis of BMI performance. Our Poisson channel model is memoryless by definition, and it is well known that feedback does not increase the capacity of memoryless channels [15]. Nevertheless, feedback has been shown to be useful in a variety of situations, for example to reduce delay and complexity of channel codes [53], and it seems to play a prominent role in the practical performance of BMIs [7, 38, 56]. Moreover, real neural channels do exhibit some memory, and in that case, feedback may well increase the capacity [15]. Some recent developments in the study of directed information and channel causality show promise for evolving a better understanding of the role of feedback in communication systems [42].
Chapter 6

Conclusions

Computing the capacity of different BMI channel models provides an objective comparison of how effectively the different structures can convey information, and how their specific properties hinder or help their ability to drive prosthesis performance. We have constructed an information-theoretic framework in which to study these effects, and have provided models of both simple and more complicated BMI structures that can be evaluated with these techniques.

Importantly, we have shown that there are no fundamental barriers to creating viable replacement sensory inputs or motor outputs: under the right conditions, the same theoretical performance can be achieved with a single electrode or an array of electrodes. Here, the “right conditions” refers to the statistics of both the inputs and the neurons themselves. However, in unfavorable conditions, the capacity can be drastically reduced. Since the OPTA minimum distortion usually increases exponentially with decreasing capacity, the impact of capacity reduction on the system performance can be quite severe.

Thus, one goal of the BMI designer must be to construct BMI channels having high enough capacity to support their application. Unfortunately, this is very difficult to achieve in practice, since we can never have control over all elements of the system. For example, the important parameters for a stimulation BMI include the number and spacing of
electrodes, the target population (position of electrodes), and the encoding strategy used. Even if the optimal encoding strategy can be determined, the performance is limited by the properties of the neurons themselves, which may or may not be known and are usually not under our control. Furthermore, achieving the optimal performance also requires the right decoder, which is implemented in the brain itself. Even the most ably designed stimulation BMI’s performance hinges on the ability of the brain to learn the optimal decoding strategy given the BMI inputs. Feedback and learning may play significant roles here.

For a control BMI, the situation is even worse. There, not only is the performance limited by the brain’s ability to learn the optimal encoding strategy, but it is further reduced by the constraints inherent in recording neural signals. Spike sorting can be used to mitigate capacity loss for a control BMI using only a few electrodes, but spike sorting algorithms must be judiciously applied; sorting errors, depending on the type and frequency, can also result in a drastic reduction in capacity.

Finally, interpreting the capacity results in terms of the OPTA for a particular source and distortion measure provides even greater insight into the role the BMI plays in the ultimate performance of a neural prosthesis. It is clear that for most practical applications, the ability to target large numbers of neurons reliably is key to the device’s performance. Even still, finding the optimal encoder or decoder may be no easy task.

Shannon’s classic work on information theory determined the ultimate fidelity limits that communication and signal processing systems can achieve. Because of some obvious similarities, neuroscientists have long thought that the tools of information theory, so
successful in characterizing communication systems, should enable a deeper understanding of how neural systems process information. Indeed, in systems neuroscience, many of the same issues Shannon addressed have always been research issues: How is information encoded? What is the fidelity of information represented by neural signals? However, despite the apparent similarities, the issues and goals of the two communities have traditionally been very different. Communication engineers want to design systems; to that end, information theory provides the designer with computable performance criteria and performance limits, and illuminates the constraints and barriers that might prevent achieving those limits. Traditionally, neuroscientists’ main focus has been to analyze an existing system, the brain. Unfortunately, the key information theoretic quantities — capacity and the rate-distortion function — are solutions of mathematical optimization problems. Information theory is silent on how to judge a given system’s performance relative to these milestones.

In this thesis, we have taken on the role of “neuroscientist-designer.” Rather than analyzing an existing system, we have posed a classic information theory problem: What performance criteria are relevant to the design and operation of BMIs? What are the fundamental performance limits that we can aim for? What constraints are there to achieving the best performance in practice? This thesis thus represents an important first step towards a systematic analysis of the optimal performance of BMIs. We believe that quantifying the performance limits of BMI systems is crucial to realizing the potential to actually build and operate such devices.
Bibliography


[24] ———, "Orthogonal decompositions of multivariate statistical dependence measures,
Proc. 2004 International Conference of Acoustics, Speech, and Signal Processing,
May 2004.


[31] D. H. Johnson and I. N. Goodman, "Inferring the capacity of the vector Poisson chan­
13–33, 2008.


[47] Y. Nir, L. Fisch, R. Mukamel, H. Gelbard-Sagiv, A. Arieli, I. Fried, and R. Malach, "Coupling between neuronal firing rate, gamma LFP, and BOLD fMRI is related to


[68] ——, “Capacity and error exponent for the direct detection photon channel - Part II,”

Appendix A

Infinitely divisible vector Poisson processes

We derive here the infinitely divisible vector Poisson process. By infinitely divisible, we mean the random vector can be expressed as a sum of an arbitrary number of statistically independent random vectors [16]. The probability distribution of the sum is the convolution of the individual probability distributions. Consequently, infinite divisibility demands that a probability distribution be expressed as the n-fold convolution of a density with itself. In special cases, like the Gaussian and the Poisson, each of the constituent random vectors has the same distributional form (i.e., they differ only in parameter values) as do their sum.

We begin by defining the probability generating functional of the Poisson process. Given a regular point process \( \{N_t, t \geq 0\} \), the probability generating functional is defined as [62]

\[
G[u(t)] = E \left[ \exp \left( \int \log u(t) dN_t \right) \right], \tag{A.1}
\]

where \( u(t) \) is a real-valued function of bounded variation such that \( 0 \leq u(t) < 1, (-\infty < t < \infty) \). Like the characteristic function and the moment generating function for random variables, the probability generating functional completely determines the probability structure of \( N_t \) and can be used to evaluate properties of the random process such as moments.

For a Poisson process with instantaneous rate \( \lambda(t) \), the probability generating functional

\[
\text{}}
\end{align}
\]
reduces to

$$G[u(t)] = \mathbb{E}\left[ \exp\left\{ \int (u(t) - 1)\lambda(t)dt \right\} \right].$$  \hspace{1cm} (A.2)

To show that the Poisson process is infinitely divisible, we note that the product of arbitrary many probability generating functionals like (A.2) itself has the same form as (A.2), with a total rate function equaling the sum of the individual rates. Since the probability generating functional is uniquely determined by the process \( N_t \) [64], the superposition of an arbitrary number of Poisson processes is also Poisson.

The probability generating functional of a collection of point processes considered jointly generalizes as

$$G^{(M)}[u(t)] = \mathbb{E}\left[ \exp\left\{ \sum_{m=1}^{M} \int \log u_m(t) dN_{m,t} \right\} \right],$$

where the expected value is computed with respect to the joint distribution of the point processes, which is the quantity we seek. The probability-generating functional of component process \( j \) can be found from this formula by setting \( u_i(t) = 1, \ i \neq j \). If the processes are statistically independent, their joint probability functional equals the product of the marginal functionals. If the processes are added, the probability generating functional of the result equals the joint functional evaluated at a common argument: \( G[u(t)] = G^{(M)}[u(t), u(t), \ldots, u(t)] \).

Over thirty years ago, the probability-generating functional of two marginally Poisson processes that satisfied the infinite-divisibility condition was shown to have the unique
form \([45]\)

\[ G^{(M)}[u_1(t), u_2(t)] = \exp \left\{ \int (u_1(t) - 1)v_1(t)dt + \int (u_2(t) - 1)v_2(t)dt + \int \int (u_1(t)u_2(t) - 1)v_c(\alpha, \beta)d\alpha d\beta \right\}. \quad (A.3) \]

This joint probability-generating functional is easily interpreted. First of all, by setting \(u_2(t) = 1\), we obtain the marginal probability-generating functional of process 1, showing that it is a Poisson process having an instantaneous rate of \(\lambda_1(t) = v_1(t) + \int v_c(t, \beta)d\beta\). Similarly, process 2 is also Poisson with a rate equal to \(\lambda_2(t) = v_2(t) + \int v_c(\alpha, t)d\alpha\). Also, setting \(v_c(\alpha, \beta) = 0\) results in the product of the marginal probability-generating functionals, corresponding to statistically independent processes. Thus, the "common rate" \(v_c(\alpha, \beta)\) represents a joint rate variation that induces statistical dependence between the processes. The simplest example is

\[ v_c(\alpha, \beta) = v_c(\beta)\delta(\alpha - \beta), \quad (A.4) \]

indicating an instantaneous correlation at each moment in time. As we shall see later, this results in the constructed processes being jointly (wide-sense) stationary. The resulting dependence term in the probability generating functional equals

\[ \int \int (u_1(t)u_2(t) - 1)v_c(\alpha, \beta)d\alpha d\beta = \int (u_1(t)u_2(t) - 1)v_c(t)dt. \]
Consequently, statistically dependent Poisson processes having an infinitely divisible joint probability distribution can be constructed from two statistically independent “building-block” Poisson processes having rates \( v_1(t) \) and \( v_2(t) \) by adding a common Poisson process having rate \( v_c(t) \) that is statistically independent of the others, a method first described by Holgate [28].

We can now generalize the two-process technique to form an arbitrary number of infinitely divisible, jointly defined Poisson processes. Given \( L \) statistically independent building-block Poisson processes, represented by the column vector \( B_t \), we create a collection of \( M < L \) statistically dependent Poisson processes \( N_t \) by superimposing the building-block processes: \( N_t = AB_t \). The construction matrix \( A \) is an \( M \times L \) matrix whose entries are either 0 or 1. For example, the construction matrix underlying the two-process example is

\[
A = \begin{bmatrix}
1 & 0 & 1 \\
0 & 1 & 1 \\
\end{bmatrix}.
\]  
(A.5)

For three processes, we use 7 building-block processes, with the construction matrix

\[
A = \begin{bmatrix}
1 & 0 & 0 & 1 & 1 & 0 & 1 \\
0 & 1 & 0 & 1 & 0 & 1 & 1 \\
0 & 0 & 1 & 0 & 1 & 1 & 1 \\
\end{bmatrix}.
\]  
(A.6)

In general, to capture the full range of dependencies between the constructed processes, a vector of \( M \) dependent Poisson processes requires \( L = 2^M - 1 \) building-block processes.
The probability generating functional of $N_t$ can be written in matrix form as

$$G^{(M)}[u(t)] = \mathbb{E} \left[ \exp \left\{ \int (A' \log u(t))' dB_t \right\} \right],$$

where the logarithm of a vector is defined as an element-by-element operation. Each component of the vector $(A' \log u(t))$ expresses which combination of components of $u(t)$ are associated with each building-block process. This combination corresponds to the constructed processes to which each building block process contributes. Since the building block processes are statistically independent and Poisson, we have

$$G^{(M)}[u(t)] = \exp \left\{ \sum_{i=1}^{L} \left( \prod_{m=1}^{M} u_{m,i}^{A_{m,i}}(t) \right) - 1 \right\} v(t) dt$$

$$= \exp \left\{ \sum_{i=1}^{L} \int \left( \prod_{m=1}^{M} u_{m,i}^{A_{m,i}}(t) \right) - 1 \right\} v_i(t) dt \right\}. \quad (A.7)$$

Here, $u_{m,i}^{A_{m,i}}(t)$ means $u_m(t)$ raised to the power of the $(m, i)$ element of $A$. In other words, the corresponding term is only included if $A_{m,i} = 1$. Thus, the exponent of the probability generating functional consists of a sum of terms, one for each building block process, wherein the coefficient of each rate $v_i(t)$ is the product of arguments corresponding to those constructed process building block process $l$ helped to build. Equation (A.7) can be seen as a direct generalization of equation (A.3) and condition (A.4). For example, using
the construction matrix given in (A.6), we obtain three Poisson processes with rates

\begin{align*}
\lambda_1(t) &= v_1(t) + v_4(t) + v_5(t) + v_7(t), \\
\lambda_2(t) &= v_2(t) + v_4(t) + v_6(t) + v_7(t), \\
\lambda_3(t) &= v_3(t) + v_5(t) + v_6(t) + v_7(t).
\end{align*}

Thus, each pair of processes shares a rate component, and all three processes share a single rate component \( v_7(t) \).
Appendix B
Cumulant correlation coefficients

Like the moment-generating function for random variables, evaluating partial derivatives of the logarithm of the probability-generating functional yields the cumulants of the vector Poisson process. We can view the cumulants as coefficients of the multivariate Taylor series for $\log G^{(M)}[u(t)]$ centered at $u(t) = 1$. Because the $m$th term in (A.7) contains only multilinear combinations of $u_m(t)$, second-order and higher derivatives of these terms are zero. Consequently, the Taylor series consists only of multilinear terms having $(u_m(t) - 1)$ as its constituents, with the cumulants as the series' coefficients:

$$
\frac{\partial^k \log G^{(M)}[u(t)]}{\partial u_{m_1}(t) \cdots \partial u_{m_k}(t)} \bigg|_{u(t)=1} = \sum_{l=1}^{L} \left( \prod_{m=m_1 \ldots m_k} A_{m,l} \right) v_l(t) \quad (B.1)
$$

Because the elements of $A$ are either 0 or 1, the product $\prod_{m} A_{m,l}$ equals either zero or one, bringing in the $l$th building-block process only if it contributes to all of the constructed processes indexed by $m_1 \ldots m_k$. For example, the first partial derivative expresses the rate
of each constructed process, which for a Poisson process is both the mean and variance:

\[
E[dN_{m,t}] = \frac{\partial \log G^{(M)}[u(t)]}{\partial u_m(t)} \bigg|_{u(t)=1} = \sum_{l=1}^{L} A_{m,l} \nu_l(t) = \lambda_m(t).
\]

Similarly, the covariance between any two processes \(m_1\) and \(m_2\) can be found by evaluating the second mixed partials. For example, using the construction matrix given in (A.5) for \(M = 2\), the covariance between the two constructed processes is

\[
cov(dN_{1,t},dN_{2,t}) = \frac{\partial^2 \log G^{(M)}[u(t)]}{\partial u_1(t) \partial u_2(t)} \bigg|_{u(t)=1} = \nu_3(t),
\]

the rate of the common building-block process. The correlation between the two processes is thus

\[
\rho^{(2)}(t) = \frac{\nu_3(t)}{\sqrt{\lambda_1(t)\lambda_2(t)}} = \frac{\nu_3(t)}{\sqrt{\nu_1(t) + \nu_3(t) + \nu_2(t) + \nu_3(t)}}, \tag{B.2}
\]
which lies in the interval \([0, 1]\), with the maximal correlation occurring in the limit of large values for the common rate. Note that, due to the construction of the common components defined by (A.7), the cumulant correlation coefficients have no temporal extent; given that an event occurs in one process, it is correlated with the other process at that event time and independent at all other times. The construction can be modified slightly to allow for a non-zero time lag in the correlation by introducing delays among the common building-block processes; however, the correlation is still instantaneous at specific time lags. Consequently, the cross-covariance between the constructed processes depends only on the time difference, and the constructed processes are jointly wide-sense stationary. Although it is possible to construct Poisson processes having negative correlations and non-zero covariance for continuous time lags [30], the generalized Holgate construction seems to be the only method capable of yielding infinitely divisible, jointly stationary Poisson processes.

To capture the full range of dependencies between the constructed processes, we generalize the correlation coefficient by defining the **cumulant correlation coefficients** as the normalized cumulants evaluated as in (B.1). The normalization factor is the geometric mean of the constructed process variances (rates):

\[
\rho_{m_1 \ldots m_k}^{(k)}(t) \triangleq \frac{\partial^k \log G^{(M)}[u(t)]}{\partial u_{m_1}(t) \cdots \partial u_{m_k}(t)} \bigg|_{u(t)=1} \frac{1}{[\lambda_{m_1}(t) \cdots \lambda_{m_k}(t)]^{1/k}}
= \frac{\sum_{l=1}^L (\prod_{m=m_1 \ldots m_k} A_{m,l}) v_l(t)}{[\sum_l A_{m_1,l} v_l(t) \cdots \sum_l A_{m_k,l} v_l(t)]^{1/k}}.
\]

Because the numerator expresses which building block processes are in common with all
the specified constructed processes, they and others are contained in each term in the denominator. This property means that each cumulant correlation coefficient is less than one and, since rates cannot be negative, greater than or equal to zero. Similar manipulations show that \( \rho^{(k)}_{m_1\ldots m_k}(t) \geq \rho^{(k+1)}_{m_1\ldots m_k}(t) \); in other words, the size of the cumulant correlation coefficients cannot increase with the order of correlation.

In the symmetric case, the expression for the cumulant correlation coefficients simplifies:

\[
\rho^{(k)}(t) = \frac{\sum_{l=k}^{M} \binom{M-k}{l-k} \nu^{(l)}(t)}{\sum_{l=1}^{M} \binom{M-1}{l-1} \nu^{(l)}(t)}.
\]

Here, \( \nu^{(l)} \) is the rate of the building-block process shared between exactly \( l \) constructed processes. Thus, the numerator is the sum of the rates of the processes that induce the \( k^{th} \)-order dependence, and the denominator is the rate \( \lambda(t) \) of each constructed process.

In the symmetric case, the cumulant correlation coefficients obey the following inequality conditions:

\[
\sum_{k=2}^{M} \rho^{(k)}(-1)^k \binom{M-1}{k-1} \leq 1,
\]

\[
\sum_{k=m}^{M} \rho^{(k)}(-1)^{k+m} \binom{M-m}{k-m} \geq 0, \quad m = 2\ldots M. \tag{B.3}
\]
Appendix C

Convergence of the vector Bernoulli process to the infinitely divisible vector Poisson process

It is well known that the single Poisson process can be derived as the limit of a discrete-time Bernoulli process when the event probability becomes arbitrarily small [19]. In a similar manner, the infinitely divisible vector Poisson process constructed in Appendix A can be derived as the limit of a vector Bernoulli process. To begin, consider the case $M = 2$, and let $X_1, X_2$ be Bernoulli random variables with event probabilities $p_1$ and $p_2$, respectively. The joint distribution can be written using the Sarmanov-Lancaster expansion [22]

$$ P(X_1, X_2) = P(X_1)P(X_2) \left[ 1 + \rho \frac{(X_1 - p_1)(X_2 - p_2)}{\sigma_1 \sigma_2} \right], $$

where the standard deviation $\sigma_i$ of each random variable equals $\sqrt{p_i(1-p_i)}$. We construct a discrete-time Bernoulli process from an iid sequence of Bernoulli random vectors in time bins with bin-width $\Delta t$.

The moment-generating function for the jointly Bernoulli distribution is

$$ \Phi(z_1, z_2) = (1 + p_1(z_1 - 1))(1 + p_2(z_2 - 1)) + (z_1 - 1)(z_2 - 1)\rho \sigma_1 \sigma_2 $$

Letting the event probabilities be $p_i = \lambda_i \Delta t$ and evaluating this expression to first order in
the event probabilities, we get

\[ \Phi(z_1, z_2) \xrightarrow{(\Delta t \to 0)} 1 + \lambda_1 \Delta t (z_1 - 1) + \lambda_2 \Delta t (z_2 - 1) + \rho \sqrt{\lambda_1 \lambda_2} \Delta t (z_1 - 1)(z_2 - 1). \]

Evaluating the logarithm and using the approximation \( \log(1 + x) \approx x \) for small \( x \), we find that

\[ \log \Phi(z_1, z_2) \approx (z_1 - 1) \lambda_1 \Delta t + (z_2 - 1) \lambda_2 \Delta t + (z_1 - 1)(z_2 - 1) \lambda_1 \lambda_2 \rho \sqrt{\lambda_1 \lambda_2} \Delta t. \]

Now, letting \( N_{i,T} = \sum_{k=0}^{\left\lfloor \frac{Z_i}{2} \right\rfloor} X_{i,k} \) be the sum of the Bernoulli random variables in each process over a fixed time interval \([0, T]\), we obtain the number of events that occur in the time interval. Since the variables are independent bin-to-bin, the moment generating function of the sum is the product of the individual joint moment generating functions, which means its logarithm equals the sum of the logarithms of the individual functions. As the bin-width decreases, the sum becomes an integral to yield

\[ \log \Phi(z_1, z_2) = (z_1 - 1) \int_0^T \lambda_1(t) dt + (z_2 - 1) \int_0^T \lambda_2(t) dt \]

\[ + (z_1 - 1)(z_2 - 1) \int_0^T \rho(t) \sqrt{\lambda_1(t) \lambda_2(t)} dt. \]

If we let \( \lambda_i(t) = \nu_i(t) + v_c(t) \) and substitute (B.2) for the correlation coefficient \( \rho(t) \), we obtain the logarithm of the probability-generating functional for two jointly Poisson processes constructed using Holgate’s method, where \( u_i(t) \to z_i \).
Generalizing this result to $M$ variables is tedious but straightforward. We note, however, that a slight modification to the usual Sarmanov-Lancaster expansion is required to make the correlation coefficients in the discrete-time model approach the cumulative correlation coefficients in the vector Poisson model. In the Sarmanov-Lancaster expansion for Bernoulli random variables, the $k^{th}$ order correlation coefficients are defined as

$$
\rho^{(k)} \Delta \frac{E[(X_{m_1} - p_{m_1}) \cdots (X_{m_k} - p_{m_k})]}{\sigma_{m_1} \cdots \sigma_{m_k}},
$$

where the denominator, which is the product of the standard deviations for the variables included in the dependence term, arises from the construction of the Sarmanov-Lancaster expansion using products of orthonormal functions [22]. However, the higher order correlation coefficients using this definition have no guaranteed range as does $\rho^{(2)}$. The cumulant correlation coefficients of the vector Poisson process, on the other hand, do have an orderliness, as described by the inequalities in (B.3). To make the two models agree, we must change the normalization of the coefficients, redefining them as

$$
\rho^{(k)} \Delta \frac{E[(X_{m_1} - p_{m_1}) \cdots (X_{m_k} - p_{m_k})]}{(\prod_{i=1}^{k} \sigma_{m_i}^2)^{1/k}}.
$$

The denominator is thus the geometric mean of the variances, corresponding to the normalization of the cumulant correlation coefficients for the vector Poisson process. Consequently, the Sarmanov-Lancaster expansion must be modified as well; now, the $k^{th}$ order
term in the expansion has the form exemplified by

\[ \rho^{(k)} \frac{(X_{m_1} - p_{m_1}) \cdots (X_{m_k} - p_{m_k})}{(\prod_{i=1}^{k} \sigma_{m_i}^2)^{1/k}}. \]

Using this normalization in the Sarmanov-Lancaster expansion now creates a direct relationship between its parameters and those of the vector Poisson probability distribution. The inequality sets shown in (B.3) also guarantee existence of the Sarmanov-Lancaster model. This change does not affect the orthogonality so crucial in defining the Sarmanov-Lancaster expansion, only the normality.

Because of the correspondence between vector Bernoulli processes and vector Poisson processes, we can use the limit of the Sarmanov-Lancaster expansion to represent the joint distribution of vector Poisson processes. In particular, we can evaluate information-theoretic quantities related to Poisson processes using this correspondence. Since entropy and mutual information are smooth quantities (infinitely differentiable) for all cases of interest to us, the small-probability limit can be evaluated after they are computed for Bernoulli processes.