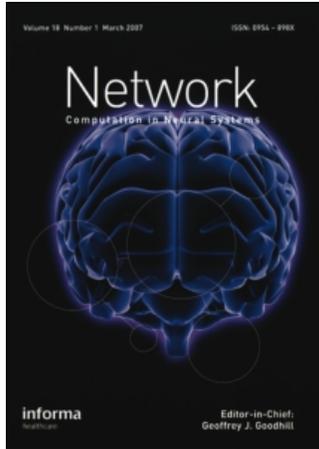


This article was downloaded by:[Rice University]  
On: 27 February 2008  
Access Details: [subscription number 776098947]  
Publisher: Informa Healthcare  
Informa Ltd Registered in England and Wales Registered Number: 1072954  
Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



## Network: Computation in Neural Systems

Publication details, including instructions for authors and subscription information:  
<http://www.informaworld.com/smpp/title~content=t713663148>

### Inferring the capacity of the vector Poisson channel with a Bernoulli model

Don H. Johnson<sup>a</sup>; Ilan N. Goodman<sup>a</sup>

<sup>a</sup> Electrical & Computer Engineering Department, MS380 Rice University, Houston, Texas 77005-1892, USA

Online Publication Date: 01 January 2008

To cite this Article: Johnson, Don H. and Goodman, Ilan N. (2008) 'Inferring the capacity of the vector Poisson channel with a Bernoulli model', Network: Computation in Neural Systems, 19:1, 13 - 33

To link to this article: DOI: 10.1080/09548980701656798

URL: <http://dx.doi.org/10.1080/09548980701656798>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.informaworld.com/terms-and-conditions-of-access.pdf>

This article maybe used for research, teaching and private study purposes. Any substantial or systematic reproduction, re-distribution, re-selling, loan or sub-licensing, systematic supply or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

## Inferring the capacity of the vector Poisson channel with a Bernoulli model

DON H. JOHNSON & ILAN N. GOODMAN

*Electrical & Computer Engineering Department, MS380 Rice University, Houston, Texas 77005–1892, USA*

*(Received 13 October 2006; accepted 31 August 2007)*

### Abstract

The capacity defines the ultimate fidelity limits of information transmission by any system. We derive the capacity of parallel Poisson process channels to judge the relative effectiveness of neural population structures. Because the Poisson process is equivalent to a Bernoulli process having small event probabilities, we infer the capacity of multi-channel Poisson models from their Bernoulli surrogates. For neural populations wherein each neuron has individual innervation, inter-neuron dependencies increase capacity, the opposite behavior of populations that share a single input. We use Shannon's rate-distortion theory to show that for Gaussian stimuli, the mean-squared error of the decoded stimulus decreases exponentially in both the population size and the maximal discharge rate. Detailed analysis shows that population coding is essential for accurate stimulus reconstruction. By modeling multi-neuron recordings as a sum of a neural population, we show that the resulting capacity is much less than the population's, reducing it to a level that can be less than provided with two separated neural responses. This result suggests that attempting neural control without spike sorting greatly reduces the achievable fidelity. In contrast, single-electrode neural stimulation does not incur any capacity deficit in comparison to stimulating individual neurons.

**Keywords:** *information capacity, point process, neural populations, neural prosthetics*

### Introduction

From an information theoretic viewpoint, a neuron can be seen as a communications channel: a neuron decodes the information expressed by its inputs, processes it and somehow represents that information in its spike-train output, just like the

---

Correspondence: Don H. Johnson, Electrical & Computer Engineering Department, Rice University, 6100 Main Street, Houston 77005-1892, USA, . E-mail: dhj@rice.edu. Tel: (713) 348-4986. Fax: (713) 348-5687

ubiquitous communication system described in Shannon's (1948) paper. Information theory has been used in a variety of ways to evaluate neural coding schemes (Bialek et al. 1993; Rieke et al. 1993, 1999; Strong et al. 1998; Johnson et al. 2001; Schneidman et al. 2001) by quantifying how much information can be transferred with specific neural codes. While this approach may provide insight into neural coding strategies, Shannon's information theory tells us that channel capacity expresses the ultimate ability of a communication channel to transfer information. Capacity determines the greatest possible fidelity to which information, digital or not, can be extracted by any means, regardless of the transmission scheme, of the signal being conveyed and how we define fidelity. The central issue thus becomes calculating the capacity, a frequently difficult, if not impossible, task.

To study the information transfer properties of neural systems, we model stochastic spike trains as point processes (Snyder 1975; Johnson 1996). Kabanov (1978) derived the capacity of any single point process channel with constrained minimal and maximal instantaneous rates. Kabanov's derivation showed that the capacity of *any* point process cannot exceed the Poisson process's capacity. Johnson (2007) presents an argument why all other point processes have a strictly smaller capacity and describes how the non-Poisson capacity can be computed from the Poisson capacity. Thus, despite being a relatively noisy point process (Johnson 1996), the Poisson process represents the maximally effective point process channel. Though the Poisson process is at best an approximate model for neural spike trains, the capacity results for the Poisson channel provide an upper bound on how effectively any point process model represents information.

While Kabanov's single point process capacity result is of some interest in neuroscience applications, what would be much more relevant is to have some idea of the effectiveness of population codes by determining how much population coding boosts capacity beyond that provided by a single neuron. Unfortunately, extending the single point process capacity results to several point processes, the so-called vector point process case, is very difficult, especially when exploring the effect on capacity of dependencies that would arise from lateral connections among neurons, what we call *connection-induced dependencies*. For most point processes, including the Poisson, the joint probability function cannot easily be written, which presents a major stumbling block to calculating information theoretic quantities. Our approach circumvents this problem by noticing that, in the limit of small binwidths, a discrete-time Bernoulli process becomes statistically equivalent to a Poisson process: at each moment, either one event or no event occurs, statistically independent of what occurred at other times, with the probability of an event occurring being very small. The same limiting approach applies to vector Bernoulli processes as well, yielding in the limit the well-defined *infinitely divisible* vector Poisson process (Johnson and Goodman 2007). In this paper, we calculate the capacity of the vector Bernoulli process, even those incorporating interprocess dependencies, and evaluate the result in the small binwidth limit; in this way, we find the vector Poisson process capacity without needing the joint distribution function.

Using this approach, we can compute the capacity for several population models depicted in Figure 1. The only special case that yields a general answer is the simplest, shown in panel (a):  $M$  statistically independent point processes, each of

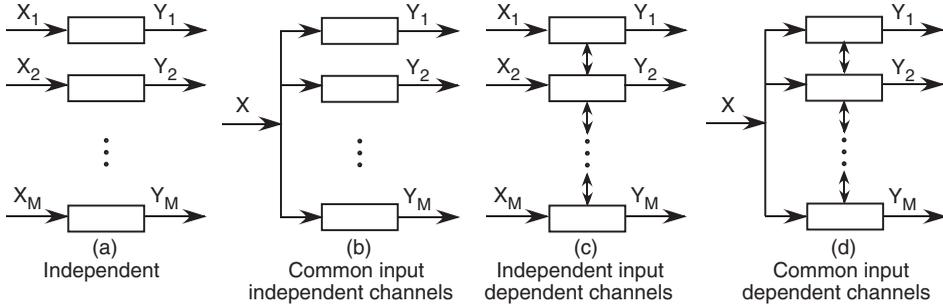


Figure 1. Several configurations of parallel channels for vector and common input cases are shown. Each block generates a Poisson process having an instantaneous rate equal to its input. We model these generators here as small-probability limits of Bernoulli process generators. The arrows between blocks symbolically represent the presence of statistical dependencies among the generators, creating a connection-induced dependence among the outputs. Each Poisson generator in (c, d) is assumed to interact with all others, not just between adjacent pairs as the simple graphics might suggest.

which has its own input signal statistically independent from the others. Simple manipulations using only the properties of mutual information show that in this case, the total capacity  $C^{(M)}$  equals to the sum of the component capacities:  $C^{(M)} = \sum_m C_m^{(1)}$ , where  $C^{(1)}$  is the single-channel capacity found by Kabanov. This result serves as the baseline against, which we compare the capacity for the other cases. Case (b) shows a population that has a common input, which creates a conditionally independent population response: only if the stimulus is provided are the individual responses independent of each other. Case (c and d) shows populations that have connection-induced dependencies, which makes the responses conditionally and unconditionally dependent. The goal is to investigate the information-theoretic consequences of these simple population structures.

We begin by describing known capacity results for the single point process channel and the probability model for jointly defined Poisson processes. Next, we define a model for jointly defined Bernoulli processes and show its equivalence in the small-binwidth limit to an important class of jointly defined Poisson processes. We use this model to derive the information capacity of the population structures shown in Figure 1, and furthermore, show how aggregating (summing) the outputs of a population affects the capacity. Finally, we show how to interpret capacity results in terms of the ultimate fidelity to, which information can be extracted from a population response. In particular, our capacity calculations for simple models of neural prostheses have important implications for the ultimate capabilities of these devices.

## Background

The point process channel produces a sequence of events  $N_t$  that encodes an input signal  $X_t$  with its intensity  $\mu(t; \mathcal{H}_t)$ . The quantity  $N_t$  denotes the number of events that have occurred since observations began, usually taken to be the time origin  $t=0$ . The intensity  $\mu(t; \mathcal{H}_t)$  represents how the instantaneous event rate depends on

the input  $X_t$  and on the process's history  $\mathcal{H}_t$ : when and how many events occurred before time  $t$ . For a regular point process, the intensity controls the probability of events occurring in a small time interval  $\Delta t$ :

$$\begin{aligned}\Pr[N_{t,t+\Delta t} = 1|\mathcal{H}_t] &= \mu(t; \mathcal{H}_t)\Delta t, \\ \Pr[N_{t,t+\Delta t} > 1|\mathcal{H}_t] &= o(\Delta t).\end{aligned}\tag{1}$$

The notation  $N_{t_1, t_2}$  denotes the number of events that occur in the time interval  $[t_1, t_2]$ . In the latter expression,  $\lim_{\Delta t \rightarrow 0} o(\Delta t)/\Delta t = 0$ , meaning that the probability of more than one event in a small interval decreases more rapidly than linearly. For a Poisson process, the intensity does not depend on previous events, and we write the intensity as an instantaneous rate function,  $\mu(t; \mathcal{H}_t) = \lambda(t)$ . For more statistically complex point processes, the event rate depends on when previous events occurred. For example, in neural discharge models that incorporate refractory effects, the probability of a spike occurring at any time depends on when the previous one occurred, and possibly on earlier ones as well (Johnson 1996).

We define the capacity of the point process channel as

$$C = \lim_{T \rightarrow \infty} \max_{p(X_{0 \leq t < T}); \mu(t; \mathcal{H}_t) \in \mathcal{C}} \frac{1}{T} I(X_{0 \leq t < T}; N_{0 \leq t < T}).\tag{2}$$

In words, capacity is the maximal asymptotic time-averaged mutual information between the input signal  $X_t$ , which represents a stimulus encoded into the point process's intensity, and the point process  $N_t$  itself. Here, the input is assumed to be a stationary process. The maximization occurs with respect to all stimulus probability distributions that allow the resulting intensity to lie in the constraint class  $\mathcal{C}$ . In this way, the intensity constraint class implicitly places constraints on the input. The constraints reflect the characteristics of the point process channel under study and usually strongly affect capacity results. In this paper we focus on constraints on the maximal and average intensities of the point process, defined as  $\max_t \mu(t; \mathcal{H}_t)$  and  $E[\mu(t; \mathcal{H}_t)]$ , respectively. For a stationary Poisson process, wherein  $\lambda(t)$  is a constant, these quantities are equal; for stationary non-Poisson processes, they can differ substantially. Poisson or not, a recording that expresses precise spike timing means it has a large maximal rate and vice versa. In all point-process capacity calculations, the minimal and maximal rates are constrained; if the maximal rate were not constrained, the capacity would be infinite. Other constraints, such as a constraint on the average rate, can be added if the channel limits average rate in addition to maximal rate. Note that neuron models containing no inherent variability when the input is deterministic, like integrate-and-fire models, have an infinite maximal rate, and consequently infinite capacity.

Kabanov (1978; also see Brémaud 1981; Davis 1980) derived the capacity of the single point process channel when the minimal and maximal intensities are constrained according to  $\lambda_{\min} \leq \mu(t; \mathcal{H}_t) \leq \lambda_{\max}$ ,

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

The division by  $\ln 2$  leaves the capacity with units of bits/s, and we adopt this convention throughout this paper. The capacity is achieved by a Poisson process

whose instantaneous rate is given by a random telegraph wave – the rate randomly switches between its minimal and maximal values – with the probability of being at the maximal rate at any given time equaling  $1/e$ . In most cases of interest in neuroscience, the appropriate minimal-rate constraint is zero, which greatly simplifies the capacity formula,

$$\lim_{\lambda_{\min} \rightarrow 0} C^{(1)} = \frac{\lambda_{\max}}{e \ln 2}, \quad (3)$$

which means the capacity-achieving signal has an average rate of  $\lambda_{\max}/e$ .

Kabanov's derivation showed that the capacity of *any* point process satisfying the minimal and maximal intensity constraints cannot exceed the Poisson process's capacity. For example, a dead time-modified Poisson process has a capacity equal to  $1/(1 + \lambda_{\max} \Delta) \cdot \lambda_{\max}/e \ln 2$  (Johnson, 2007). Imposing additional constraints cannot increase capacity. For example, if in addition to maximal rate, we constrain the average rate to be less than  $\bar{\lambda}$ , the capacity  $\bar{C}^{(1)}$  is smaller unless  $\bar{\lambda} \geq \lambda_{\max}/e$  (Davis 1980).

$$\lim_{\lambda_{\min} \rightarrow 0} \bar{C}^{(1)} = \begin{cases} \frac{\bar{\lambda}}{\ln 2} \ln \frac{\lambda_{\max}}{\bar{\lambda}}, & \bar{\lambda} \leq \frac{\lambda_{\max}}{e} \\ \frac{\lambda_{\max}}{e \ln 2}, & \bar{\lambda} > \frac{\lambda_{\max}}{e} \end{cases} \quad (4)$$

This result illustrates that capacity expressions and computed values can change substantially when the constraint class is modified.

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 the 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 statistical property known as *infinite divisibility* (Daley and Vere-Jones 1988), meaning it can be decomposed continually into sums of independent vector Poisson processes. This property puts the joint Poisson process distribution function on a par with the jointly Gaussian distribution. Using a method that generalizes Holgate's (1964) way of creating a bivariate Poisson distribution, forming  $M$  jointly infinitely divisible Poisson processes requires the superposition of collections of no more than  $2^M - 1$  statistically independent *building-block* Poisson processes (Johnson and Goodman 2007). For example, to construct a pair of dependent Poisson processes, we need three building block processes, which we denote by  $B_{1,t}$ ,  $B_{2,t}$ ,  $B_{3,t}$  that have instantaneous rate functions  $v_1(t)$ ,  $v_2(t)$ ,  $v_3(t)$ , respectively. We form the pair according to the superposition

$$\begin{aligned} N_{1,t} &= B_{1,t} + B_{3,t} \\ N_{2,t} &= B_{2,t} + B_{3,t}. \end{aligned}$$

All of the dependence between the constructed processes is expressed by the events produced by  $B_{3,t}$ , which occur in both of the constructed processes. Consequently, correlations between the processes are non-negative, and only occur instantaneously when the common process produces an event. More detail on this construction can be found in a companion paper (Johnson and Goodman 2007).

Although constructing vector Poisson processes having dependencies is straightforward using this method, the joint probability function still cannot easily be written, which presents a major stumbling block to calculating information theoretic quantities. We circumvent this problem by noticing that the definition of a Poisson process represents a special case of Equation 1, wherein the probability of an event occurring in the interval  $[t, t + \Delta t]$  equals  $\lambda(t)\Delta t$  for sufficiently small binwidth  $\Delta t$ . “Sufficiently small” means that  $\lambda(t)\Delta t$  is not only less than one (so that it is a meaningful probability), but much smaller than one. The fact that a Poisson process’s intensity does not depend on its history means that event probabilities are statistically independent from bin to bin. Furthermore, the probability of multiple events within a single bin is of second-order and can be ignored. All in all, in the limit of small binwidth, the discrete-time Bernoulli process becomes statistically equivalent to a Poisson process: at each moment, either one event or no event occurs, statistically independent of what occurred at other times, with the probability of an event occurring being very small. The same limiting procedure applies to jointly defined Bernoulli processes as well, yielding an infinitely divisible vector Poisson process in the limit (Johnson and Goodman 2007). As we shall show, it is much easier to compute capacities for the vector Bernoulli process, even when interprocess dependencies are incorporated. Our approach amounts to computing the capacity first for the Bernoulli process model, then evaluating the vanishingly small binwidth limit to infer the capacity of the vector Poisson channel. This approach is justified because of the smoothness properties of mutual information, which essentially allows us to evaluate well-behaved limits in any order.

## Results

### *The Bernoulli channel*

A Bernoulli process  $Y(n)$  equals either zero or one at each bin index  $n$ , and the resulting value is statistically independent of what occurs at any other bins. We let  $X(n)$  denote the probability that the Bernoulli channel output equals one at bin  $n$ .

$$P[Y(n)|X(n)] = \begin{cases} X(n), & Y(n) = 1 \\ 1 - X(n), & Y(n) = 0 \end{cases}$$

We consider the Bernoulli probability  $X(n)$  to be stochastic as well, resulting in what can be called a doubly stochastic Bernoulli process. The unconditional output probability distribution has a simple expression.

$$P[Y(n)] = \begin{cases} \bar{X}, & Y(n) = 1 \\ 1 - \bar{X}, & Y(n) = 0 \end{cases} \quad (5)$$

Here,  $\bar{X} = E[X(n)]$ , the expected value of the input process, taken here to be stationary. We further assume that the input consists of statistically independent values from bin to bin. Consequently, 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 mutual information between a channel's input and its output can be written in several equivalent forms (Cover and Thomas 2006).

$$\begin{aligned} I(X; Y) &= \int p_{X,Y}(x,y) \log \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)} dx dy \\ &= \int p_{X,Y}(x,y) \log \frac{p_{Y|X}(y|x)}{p_Y(y)} dx dy \\ &= H(Y) - H(Y|X) \end{aligned} \quad (6)$$

Here,  $H(Y)$  denotes the entropy of the random variable appearing in its argument:  $H(Y) = -\int p_Y(y) \log p_Y(y) dy$ . Evaluating the mutual information expression (6) for a doubly stochastic Bernoulli channel yields

$$I(X; Y) = E[X \log X + (1 - X) \log(1 - X)] - \bar{X} \log \bar{X} - (1 - \bar{X}) \log(1 - \bar{X}). \quad (7)$$

Calculation of the required expected value can be analytically difficult. That said, under minimal and maximal probability constraints  $x_{\min} \leq X \leq x_{\max}$ , we found by using the Arimoto-Blahut algorithm (Cover and Thomas 2006) that the capacity-achieving input probability distribution consists of impulses (probability masses) situated at the extremes of the input's possible values. Assuming the input forces the Bernoulli probability to be within the interval  $[0, x_{\max}]$ , this capacity-achieving input probability distribution has the form

$$p_X(x) = q\delta(x - x_{\max}) + (1 - q)\delta(x), \quad (8)$$

with  $\delta(x)$  denoting a Dirac delta-function and  $q$  the probability the input is not zero. This input probability distribution allows easy evaluation of the expected value in (7). If we want to calculate the capacity only under the maximal probability constraint, we maximize this result with respect to the parameter  $q$ .

$$\begin{aligned} \tilde{C} &= \max_q \{ qx_{\max} \log x_{\max} + q(1 - x_{\max}) \log(1 - x_{\max}) \\ &\quad - qx_{\max} \log qx_{\max} - (1 - qx_{\max}) \log(1 - qx_{\max}) \} \end{aligned} \quad (9)$$

Here,  $\tilde{C}$  denotes the capacity for each bin. We will eventually divide this capacity by the binwidth to obtain a result having units of bits/s:  $C^{(1)} = \tilde{C}/\Delta t$ . The capacity-achieving probability  $q_C$  equals

$$q_C = \frac{(1 - x_{\max})^{1/x_{\max}}}{1 - x_{\max} + x_{\max}(1 - x_{\max})^{1/x_{\max}}},$$

and the resulting expression for the capacity is complicated but easily found. Since the Bernoulli process approaches the Poisson process only in the limit of small binwidths, we focus on the asymptotic case where  $x_{\max} \rightarrow 0$ . Calculations show that  $\lim_{x_{\max} \rightarrow 0} q_C = 1/e$  and that the capacity is

$$\tilde{C} = \frac{x_{\max}}{e \ln 2} + o(x_{\max}) \text{ bits},$$

where  $o(x_{\max})$  denotes terms of higher order than linear. If we take  $x_{\max} = \lambda_{\max} \Delta t$ , the capacity per unit time in the small-binwidth limit equals

$$C^{(1)} = \frac{\lambda_{\max}}{e \ln 2} \text{ bits/s,} \quad (10)$$

the same expression as (3) for the point-process channel capacity under rate-range constraints when the minimal rate equals zero.

If we want to impose an average rate constraint as well, note that the average value of the capacity-achieving input distribution is  $\bar{X} = qx_{\max}$ . 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 (9) to the value  $\bar{X}/x_{\max}$  so long as  $q < 1/e$ . Echoing the previous analysis, setting  $x_{\max} = \lambda_{\max} \Delta t$  and letting  $\Delta t \rightarrow 0$ , we find that with  $q = \bar{\lambda}/\lambda_{\max}$ , the average-rate-constrained point process capacity of (4) results.

In either case, the capacity-achieving probability distribution corresponds to a discrete-time telegraph wave switching randomly between zero and the maximal probability. With only a maximal rate constraint, the probability of being in the maximal-probability state equals  $1/e$ , mimicking the continuous-time answer for point processes. Because the Poisson process and the Bernoulli process become statistically indistinguishable as the binwidth and the Bernoulli probability approach zero, the agreement of capacity results justifies our claim that we indeed can use a Bernoulli model to find the Poisson channel's capacity.

### *Capacity of the vector Bernoulli channel*

We are now in a position to calculate the capacity of the population structures shown in Figure 1 using a vector Bernoulli process model.

*Conditionally independent outputs.* This special case, exemplified by Figure 1(a) and (b), has no dependencies among the outputs save for that induced by the input signals. Here,

$$P(\mathbf{Y}|\mathbf{X}) = \prod_{m=1}^M P(Y_m|\mathbf{X}), \quad (11)$$

where  $\mathbf{X}$  represents the vector of inputs. The unconditional joint output probability distribution equals

$$P(\mathbf{Y}) = \int \prod_{m=1}^M \underbrace{P(\mathbf{Y}_m|\mathbf{X} = \mathbf{x})}_{\text{blue}} \underbrace{p_{\mathbf{X}}(\mathbf{x})}_{\text{blue}} d\mathbf{x}. \quad (12)$$

Only when the inputs are statistically independent, as in Figure 1(a), does this expression factor, leaving the outputs statistically independent as well. Otherwise, the outputs are (unconditionally) statistically *dependent*. When each Bernoulli channel shares a common input, we replace  $\mathbf{X}$  by the scalar  $X$  in Equations (11) and (12).

To find the mutual information, we use (6). Assuming the outputs are conditionally independent as in (11), the conditional entropy term equals the expected value of each output's conditional entropy,

$$\begin{aligned} H(\mathbf{Y}|\mathbf{X}) &= \mathbb{E}_{\mathbf{X}} \left[ \sum_m H(Y_m|\mathbf{X}) \right] \\ &= \mathbb{E}_{\mathbf{X}} \left[ \sum_m \sum_{y_m} -P_{Y_m|\mathbf{X}}(y_m|\mathbf{X}) \log P_{Y_m|\mathbf{X}}(y_m|\mathbf{X}) \right]. \end{aligned}$$

The conditional entropy  $H(\mathbf{Y}|\mathbf{X})$  for the common-input case can be easily found, especially when the input has the probability distribution given in (8). Simple calculations reveal that

$$H(\mathbf{Y}|X) = -Mq \cdot (x_{\max} \log x_{\max} + (1 - x_{\max}) \log(1 - x_{\max})). \quad (13)$$

The first term in the expression for mutual information, the entropy of the joint output  $H(\mathbf{Y})$ , does *not* equal the sum of the component entropies because the common input makes the outputs statistically dependent on each other. However, the joint probability distribution is easily found when the input has the form of (8). Letting  $m_{\text{nz}} = \sum_m y_m$  be the number of non-zero outputs, this joint distribution is given by the following.

$$P(\underbrace{Y_1 = 1, Y_2 = 0, \dots, Y_M = 1}_{m_{\text{nz}} \text{ non-zero terms}}) = \begin{cases} q(1 - x_{\max})^M + (1 - q), & m_{\text{nz}} = 0 \\ qx_{\max}^{m_{\text{nz}}}(1 - x_{\max})^{M - m_{\text{nz}}}, & m_{\text{nz}} = 1, \dots, M \end{cases}$$

The joint entropy consequently equals

$$\begin{aligned} H(Y) &= -[q(1 - x_{\max})^M + (1 - q)] \log[q(1 - x_{\max})^M + (1 - q)] \\ &\quad - \sum_{m_{\text{nz}}=1}^M \binom{M}{m_{\text{nz}}} qx_{\max}^{m_{\text{nz}}}(1 - x_{\max})^{M - m_{\text{nz}}} \log [qx_{\max}^{m_{\text{nz}}}(1 - x_{\max})^{M - m_{\text{nz}}}], \quad (14) \end{aligned}$$

and the mutual information equals the difference of (14) and (13). Focusing on the asymptotic case  $x_{\max} \rightarrow 0$ , we find that

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

Ignoring the higher order term, the maximizing value of  $q$ ,  $q_C$ , equals  $1/e$ , the same value as in the single-output case. If an average rate constraint is imposed as well, the capacity is found by setting  $q = \bar{\lambda}/\lambda_{\max}$ . Under either constraint, we conclude that the capacity of conditionally independent parallel Poisson channels having common inputs equals that obtained when the outputs are statistically independent.

$$\begin{aligned} C^{(M)} &= M \frac{\lambda_{\max}}{e \ln 2} = MC^{(1)} \\ \bar{C}^{(M)} &= M \frac{\bar{\lambda}}{\ln 2} \ln \frac{\lambda_{\max}}{\lambda_{\min}} = M\bar{C}^{(1)}, \quad \bar{\lambda} < \frac{\lambda_{\max}}{e} \end{aligned}$$

The dependence induced by the common input does not affect the capacity of conditionally independent Bernoulli channels in the small event-probability limit and therefore, by inference, conditionally independent Poisson channels

have the same property. 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.

*Conditionally dependent outputs.* We have already shown how output dependence induced by a common input does not affect the capacity of a population. However, statistical dependencies among the outputs can also occur because of inter-channel interactions, depicted in Figure 1(c) and (d), which give rise to *connection-induced* dependencies. As before, we can use the limiting behavior of the vector Bernoulli channel to infer whether these dependencies reduce or increase channel capacity compared to the independent case.

The joint probability function for a Bernoulli random vector can be written in several equivalent forms. The most convenient is the Sarmanov-Lancaster expansion (Goodman 2004; Johnson and Goodman 2007). For example, when  $M=2$ , the conditional joint distribution can be written as

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

In this case, the coefficient  $\rho_{12}^{(2)}$  equals the simple correlation coefficient between the pair of Bernoulli random variables  $Y_1$  and  $Y_2$ . When  $M=3$ , the term inside the square brackets consists of three pairwise dependency terms with coefficients  $\rho_{12}^{(2)}$ ,  $\rho_{13}^{(2)}$  and  $\rho_{23}^{(2)}$  like the one above, along with a third-order dependency term that has the form

$$\rho_{123}^{(3)} \frac{(Y_1 - E[Y_1|X])(Y_2 - E[Y_2|X])(Y_3 - E[Y_3|X])}{\left(\sigma_{Y_1|X}^2 \sigma_{Y_2|X}^2 \sigma_{Y_3|X}^2\right)^{2/3}}.$$

The somewhat unusual denominator is required to make  $\rho_{123}^{(3)}$  equal the *cumulant correlation coefficient*, a quantity that arises from evaluating the cumulants of the probability generating functional for the vector Poisson process (Johnson and Goodman 2007). In this paper, we only consider the case of complete symmetry in the dependence structure: all second-order dependence parameters equal some value  $\rho^{(2)}$ , all third-order parameters equal some other value  $\rho^{(3)}$ , etc. In that case, the cumulant correlation coefficients obey the following inequalities:

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

$$\sum_{k=m}^M \rho^{(k)} (-1)^{k+m} \binom{M-m}{k-m} \geq 0, \quad m = 2, \dots, M. \tag{15}$$

In particular, the above inequalities imply that cumulant correlation coefficients of all orders are non-negative, less than one and smaller than all lower-order cumulant correlation coefficients:  $0 \leq \rho^{(k+1)} \leq \rho^{(k)} \leq 1$ ,  $k = 2, \dots, M-1$ . Furthermore,

if no building block processes of order higher than two are present, which makes  $\rho^{(3)} = 0, \rho^{(4)} = 0, \dots$ , the second-order correlation coefficient cannot be bigger than  $1/(M-1)$ .

The resulting mutual information expressions are complex, but we were able to compute them with the aid of the symbolic manipulation software MATHEMATICA. Consider first the case of two dependent Bernoulli channels (cumulant correlation coefficient  $\rho^{(2)}$ ) [having a common input. Just as before, assuming the input has the probability distribution given in (8) and evaluating the asymptotic behavior as  $x_{\max} \rightarrow 0$ , we obtain

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

Consequently, we infer that the capacity of the two-component, common-input, vector Poisson channel equals

$$C^{(2)} = (2 - \rho^{(2)}) \frac{\lambda_{\max}}{e \ln 2} = (2 - \rho^{(2)}) C^{(1)},$$

a quantity decreasing linearly with increasing correlation. A similar result applies to the average-rate constrained capacity as well. 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 optimizing input has  $q_C = 1/e$  and equals

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

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, capacity decreases as the population's correlation coefficients increase. 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, small pairwise correlations can dramatically reduce capacity. Figure 2 illustrates how increasing correlation diminishes capacity of Poisson channels having common inputs.

We can elaborate the common-input model to allow different maximal event probabilities, akin to considering component Poisson processes that have different maximal rates, and different cumulant correlation coefficients. We modify the common-input model by inserting an attenuation factor  $a_m$  for each channel's event probability.

$$P[Y_m|X] = \begin{cases} a_m X, & Y_m = 1 \\ 1 - a_m X, & Y_m = 0 \end{cases} \quad 0 \leq a_m \leq 1$$

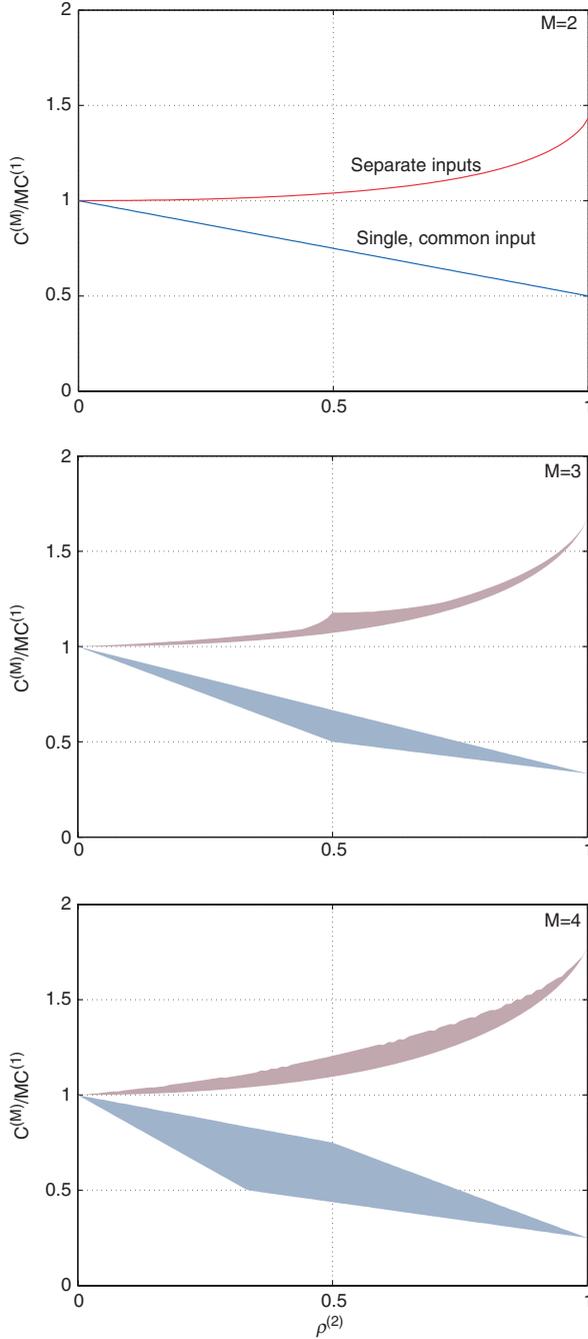


Figure 2. Capacity varies with population size  $M$  and with the dependence parameters. The capacity for  $M = 2, 3, 4$  are plotted as a function of  $\rho^{(2)}$  for two scenarios: the common-input case and the separate-input case. The vertical axis is the capacity normalized by its zero-dependence value ( $MC^{(1)}$ ). The broad spread for  $M = 3$  and  $M = 4$  occur because cumulant correlation coefficients 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 cumulant correlation coefficients range over their allowable values.

As before, the input's values range over the interval  $[0, x_{\max}]$ , but each channel now has a smaller event probability determined by its attenuation. As with our previous results, the capacity-achieving input distribution is again bi-valued as expressed by (8), with  $q_C = 1/e$  regardless of the values for the cumulant correlation coefficients and the attenuations. The capacity equals

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

where, the cumulant correlation coefficients must be non-negative but obey a much more complicated version of the inequalities in (15). The combination of correlation coefficients 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 event generators always decreases capacity for vector Poisson processes having a common input.

Now consider the case of dependent channels with separate inputs. Again, for initial simplicity we consider the special case of two channels ( $M=2$ ). Symmetry considerations suggest that the mutual information is maximized by identically distributed inputs, which we take to equal the bi-valued distribution expressed by (8). Calculations show that independent inputs maximize mutual information regardless of the value of  $\rho^{(2)}$ . The resulting expression for mutual information as  $x_{\max} \rightarrow 0$  is

$$I(X_1, X_2; Y_1, Y_2) = 2q[q(1 - \rho^{(2)}) \log(1 - \rho^{(2)}) - (1 - q\rho^{(2)}) \log(1 - q\rho^{(2)}) - \log q] \\ \times x_{\max} + o(x_{\max})$$

To calculate the capacity, we need to maximize with respect to  $q$ , the sole remaining parameter of the input distribution. 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. Numeric optimization shows that  $q_C$  depends on  $\rho^{(2)}$ , monotonically increasing from  $1/e$  for  $\rho^{(2)} = 0$  to 0.575 as  $\rho^{(2)} \rightarrow 1$ . Consequently, we could not find an expression for the capacity, but could numerically evaluate it. Surprisingly, the capacity *increases* with  $\rho^{(2)}$ , equaling 1.43 times its  $\rho^{(2)} = 0$  value when  $\rho^{(2)} = 1$  (Figure 2).<sup>1</sup> Similar results occur in the average-rate constrained case when we set  $q = \bar{\lambda}/\lambda_{\max}$ .

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 found that values for  $q_C$  and capacity depend 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 channels. This percentage increase seems to saturate at some value less than 100%: the increases for  $M=2, 3, 4, 5$  are 43.3, 67.5, 78.2, and 83.2%, respectively.

Aggregate population behavior

Another interesting case arises when we sum the outputs of the separate-input or common-input systems shown in Figure 1(c) and (d) to produce a single output:  $Y = \sum_m Y_m$ . The summation, which amounts to a superposition of the component Bernoulli processes, models unsorted extracellular recordings made with a single-electrode and EEG recordings that represent the superposition of many neural signals.

Intuitively, aggregating the outputs should decrease the mutual information; indeed, by the Data processing inequality,  $I(\mathbf{X}; \mathbf{Y}) \geq I(\mathbf{X}; Y)$  (Cover and Thomas 2006). Using our Bernoulli approximation approach, we can determine to what extent considering only aggregate behavior reduces the capacity of a population. Note that the capacity achieving input does not necessarily consist of statistically independent inputs when the outputs are aggregated. However, we constrain the inputs to be independent to determine the effect of aggregation with all other things being equal. We make the simplifying assumption that the probability distribution of each input has the form of (8), which means that some inputs are identically zero and others equal  $x_{\max}$ . Letting  $m_{\text{nz}}$  represent the number of non-zero inputs,  $P(m_{\text{nz}}) = \binom{M}{m_{\text{nz}}} \times q^{m_{\text{nz}}} (1 - q)^{M - m_{\text{nz}}}$ . The probability distribution of the summed output conditioned on  $m_{\text{nz}}$  non-zero input probabilities equals

$$P(Y|m_{\text{nz}}) = \binom{m_{\text{nz}}}{Y} x_{\max}^Y (1 - x_{\max})^{m_{\text{nz}} - Y} \left[ 1 + \sum_{k=2}^{m_{\text{nz}}} \rho^{(k)} \sum_{l=0}^k \binom{m_{\text{nz}} - Y}{k - l} \binom{Y}{l} \times (-1)^{k-l} x_{\max}^{1-l} (1 - x_{\max})^{l-k+1} \right], \quad Y = 0, \dots, m_{\text{nz}}.$$

Using this expression, we can find the capacity for the output-summed channel in the separate-input case. Again, the optimizing value  $q_C$  can only be found numerically and depends on the cumulant correlation coefficients. As expected, Figure 4 shows that dependence results in higher capacity than when the coefficients are zero. As the size of the population grows, the aggregated-output capacity differs more and more from unaggregated values. When statistically independent channels

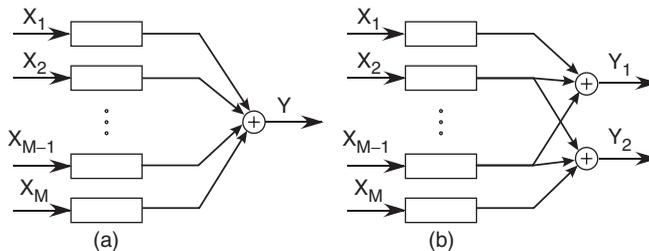


Figure 3. The simplest aggregation model shown on the left sums the population response into a single signal. More complicated models use multiple recordings from subpopulations. The diagram on the right shows two recordings made from subpopulations, each of size  $M - 1$ . Our capacity results show that the most effective subpopulation grouping for two recordings sums  $2M/3$  neural responses.

are separately received, we have shown the capacity to be  $C^{(M)} = MC^{(1)}$ , which grows without bound as the population increases in size. When aggregated, the capacity cannot be larger than  $1.58C^{(1)}$  when the channels are independent. Consequently, not separating an aggregate recording into its constituents greatly reduces the information that can be gleaned, with the aggregate information flow in the conditionally independent case being substantially less than what two component channels can sustain.

This sharp capacity decrease can be mitigated by forming multiple aggregations (Figure 3), 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. Calculations similar to those leading to the single-aggregation result 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 subpopulations must overlap substantially to maximize capacity. For large populations with no connection-induced dependencies, the capacity becomes  $(2L - 1) 1.58C^{(1)}$ , indicating that multiple aggregated recordings can greatly increase capacity, equaling the single-aggregation capacity multiplied by a factor of about twice the number of recordings. This asymptotic result breaks down when the factor  $(2L - 1) 1.58$  approaches  $M$ .

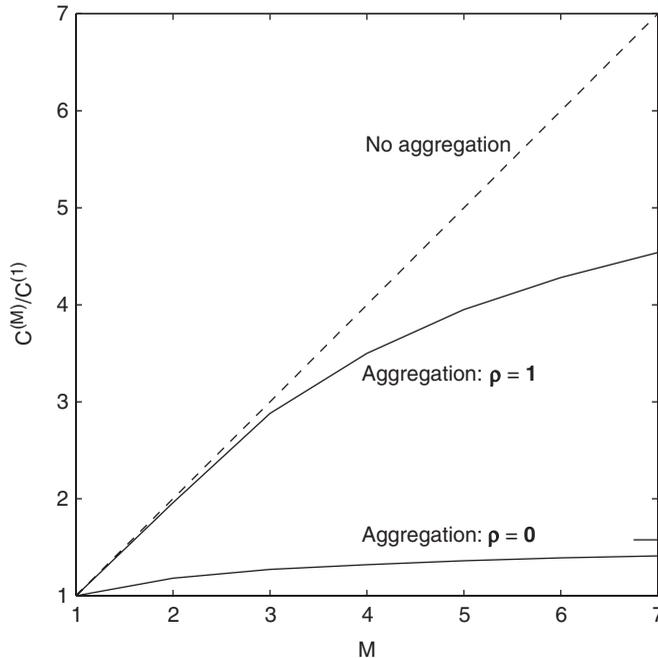


Figure 4. The normalized capacity  $C^{(M)}/C^{(1)}$  of the aggregated population grows slowly with population size  $M$  regardless of whether the population is conditionally independent ( $\rho = 0$ ) or completely dependent ( $\rho = 1$ ). The long tick mark on the right indicates the asymptotic value (1.577) of the normalized capacity in the conditionally independent case. Here,  $\rho$  is the vector of cumulant correlation coefficients. The unaggregated normalized capacities for these cases are  $M$  (indicated by the dashed line) and  $\alpha(\rho)M$ ,  $\alpha(\rho) < 2$ , respectively.

## Discussion

### *Capacity and information processing in neurons*

As we noted in the Introduction, capacity determines the maximal fidelity of information about a channel's inputs that can be extracted from its outputs. Most often, Shannon's channel capacity is used to characterize *digital* communication systems; using capacity to study neural channels, which are not digital, requires careful consideration of exactly how Shannon's results apply in this context.

Figure 5 shows a communications model of the prototypical neural processing schema. A stimulus signal  $S$  is encoded into the signal  $X$ , which could be the culmination of multiple layers of neural processing. The encoded signal  $X$  serves as the input to a system for further processing and/or relaying. This system could be a single neuron, a population of neurons, or several layers of neural systems. It has two goals: (1) process its inputs to extract salient features; and (2) jointly represent the result in the discharge pattern occurring in its outputs. Because neural patterns are inherently noisy, the representational stage introduces randomness into the processed result. Using communications theory jargon, the system serves as a channel. The channel's output is a spike-train signal  $Y$ , which serves as the input to a decoder that infers as best it can what the stimulus was and produces a stimulus estimate  $\hat{S}$ . Although a stimulus estimate may not explicitly be produced in an actual neural system, the ability of the channel to transmit information about the stimulus is characterized by how well the stimulus *could* be estimated from  $Y$ . The key to this analysis is the capacity of the channel relating  $X$  and  $Y$ . When  $X$  and  $Y$  have discrete values, Shannon's Noisy Channel Coding Theorem states that the channel's input can be decoded without error if the information rate is less than capacity. Thus, for digital communication problems, capacity defines a sharp boundary: communication rates less than capacity offer perfect fidelity while higher rates must suffer the smallest possible fidelity.

Spike trains, modeled as point processes, contain both discrete- (how many spikes occur) and continuous- (when the spikes occur) valued quantities. Consequently, to examine neural coding schemes in their full generality by including spike timing, the Noisy Channel Coding Theorem does not apply. Instead, the significance of capacity becomes evident only when viewed in the context of another of Shannon's classic results: rate-distortion theory. Shannon defined the *rate-distortion function*, which measures how accurately a source must be encoded to achieve a specified degree of error (Berger 1971). In the model shown in Figure 5, the encoded signal

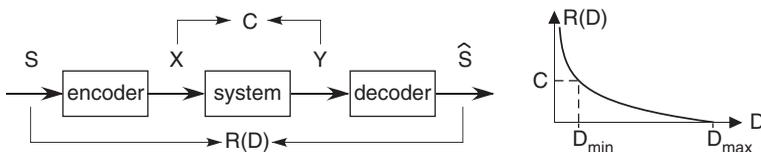


Figure 5. The communication system model is shown, along with the rate-distortion curve for a single source. Capacity is derived from the mutual information between channel input and output; the rate-distortion function arises from the mutual information between the stimulus and its estimate. The channel capacity dictates the minimal distortion achievable by any estimator.

$S$  is transmitted and/or processed to produce the estimate  $\widehat{S}$ . The quality of  $\widehat{S}$  is characterized by the error or *distortion*  $d(S, \widehat{S})$  between the signal and its estimate. The theory is quite general, as the signals can be defined in continuous- or discrete-time and can have continuous- and/or discrete-valued components. In addition, rate-distortion theory subsumes *any* choice for the distortion measure; thus, for example, a particular distortion measure could account for how well the salient features have been extracted and represented. The only restriction on the choice of  $d(S, \widehat{S})$  is that  $\overline{D} = E[d(S, \widehat{S})]$ , the average distortion, is well-defined. The rate-distortion function  $\mathcal{R}(D)$  is defined to be the minimal mutual information between the signal and its estimate when the average distortion is no greater than  $D$ . To characterize how well a specific source  $S$  (as described by its probability distribution) can be encoded, the minimization is calculated with respect to all conditional distributions  $p(\widehat{S}|S)$ , regardless of how the signal is encoded, corrupted, and processed.

$$\mathcal{R}(D) = \min_{p(\widehat{S}|S): \overline{D} \leq D} I(S; \widehat{S}).$$

The value of  $\mathcal{R}(D)$ , known as the rate, measures the quality of the encoding system: the greater the rate, the more information about the signal is conveyed and the smaller the distortion. Although rate has units of bits (or bits/s if a time-average is used in the definition), rate-distortion functions can be meaningfully calculated for all situations, even those where physical bits play no role. For example, if the stimulus were a bandlimited Gaussian random process having power  $\rho$  with highest frequency  $W$  and if the average distortion were mean-squared error, the rate-distortion function is given by (Berger 1971, Chap. 4):

$$\mathcal{R}(D) = W \log \frac{P}{D}, \quad D \leq P. \quad (16)$$

This example is depicted in Figure 5 and illustrates the behavior shared by all rate-distortion functions: they are always strictly convex and strictly decreasing, equaling zero at some maximal distortion  $D_{\max}$  (Berger 1971). In this Gaussian example,  $D_{\max} = P$ . Zero rate means that nothing about the signal is encoded, leaving the estimation system to make an intelligent guess based on the input's properties. For example, if the signal is a random variable and the distortion measures mean-squared error, using the signal's expected value as the estimate minimizes the data-ignorant estimator's distortion. No rational system should yield a distortion larger than  $D_{\max}$ .

Figure 5 illustrates the importance of capacity in the context of rate-distortion theory. More general than the Noisy Channel Coding Theorem, Shannon's rate-distortion theory states that if a communication channel or a processing system having capacity  $C$  intervenes between an encoder and decoder, the point at which capacity equals the rate-distortion function defines the smallest achievable distortion  $D_{\min} : C = \mathcal{R}(D_{\min})$ . Since this result holds for *any* source and *any* distortion function, capacity dictates the ultimate limit to which a communication system can convey information about any source signal regardless of how error is quantified. And, as *all* rate-distortion functions are smooth and decreasing, a smaller capacity *must* lead to larger distortions (i.e., reduced fidelity), regardless of how distortion is defined. Said another way, capacity, in concert with the rate-distortion function,

determines the ultimate capabilities that any given encoding and processing system can have.

To determine what effect the population capacities might have on perceptual accuracy, we would need the rate-distortion function that incorporates perceptual distortion measures and makes use of the properties of naturally occurring stimuli. Rate-distortion calculations are notoriously difficult to make; the Gaussian stimulus result (16) for mean-squared error distortion is one of the few known. For this case, Equation (16) says that the smallest achievable distortion decreases exponentially in the capacity (Figure 6) regardless of the type of channel that intervenes:

$$D_{\min} = D_{\max} e^{-C/W}.$$

Substituting our capacity results for the vector Poisson channel, we obtain

$$D_{\min} = D_{\max} \exp\left\{-\frac{\alpha(\rho)M\lambda_{\max}}{e \cdot W}\right\}. \quad (17)$$

As Figure 6 shows, if the total 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 Poisson process generator, capacity (in bits/s) equals  $\lambda_{\max}/1.88$ , 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. For example, visual signals having a bandwidth of about 30 Hz would require a maximal firing rate well over 100 Hz for a single neuron to represent accurately temporal stimulus changes. In the auditory system, the situation is much worse. Auditory-nerve fibers having a center frequency of 1 kHz have a bandwidth of about 500 Hz; thus, a single neuron would need to be capable of a maximal discharge rate of several thousand spikes/s.

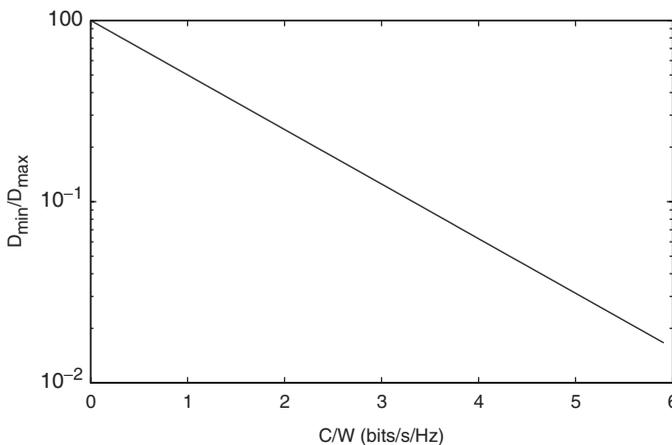


Figure 6. The smallest possible mean-squared distortion (normalized by the maximal distortion) decreases exponentially with the ratio of the capacity and the Gaussian stimulus's bandwidth. The plot shows how large the capacity-to-signal-bandwidth ratio needs to be to produce small distortions. For example, to reduce distortion by a factor of ten,  $C/W$  needs to be about three.

The analysis indicates that population coding is *essential* for accurate stimulus reconstruction. For all the population structures we considered here, capacity is proportional to population size. The quantity  $\alpha(\rho)$ , plotted in Figure 2, multiplies the population size and captures how capacity depends on connection-induced dependence as summarized by the vector  $\rho = \{\rho^{(2)}, \dots, \rho^{(M)}\}$  of cumulant correlation coefficients. When the population's outputs are conditionally independent so that there is no connection-induced dependence,  $\rho = \mathbf{0}$  and  $\alpha(\mathbf{0}) = 1$ , regardless of the input innervation pattern. In this simple case, individually innervated populations do not necessarily encode information better than ones that have shared innervation. However, connection-induced dependence among the population's members changes the story. When the population has a common input,  $\alpha(\rho)$  decreases with increasing connection-induced dependence, becoming as small as  $1/M$ , which erases the capacity gain afforded by the population. In the separate-input case,  $\alpha(\rho) > 1$ , meaning connection-induced dependence effectively magnifies the population size. Somewhat surprisingly, the effect of  $\alpha(\rho)$  on the capacity (and therefore, the minimal achievable distortion) increases with population size: larger populations obtain a greater percentage increase in capacity at maximal correlation. Thus, depending on the input innervation pattern, connection-induced dependence can facilitate or hinder the ability of a population to express information accurately.

#### *Application to neural prostheses*

Interest in the aggregated-output case stems from classic work on analyzing EEG recordings and current work in neural-activity-driven control of prosthetic devices. In many cases, extracellular recordings are not teased apart into individual neural activities; instead, the summed activity is used as a surrogate for population activity (Andersen et al. 2004). Extracellular recordings can thus be modeled as a superposition of individual neural activity patterns followed by a filtering operation that converts pooled event streams into analog recordings. As long as the filtering operation satisfies mild technical conditions, mutual information will not decrease because of it. However, subsequently added noise would decrease the mutual information. Our capacity results for aggregated populations thus represent the largest possible capacities afforded by noise-free recordings.

Aggregating conditionally independent population responses into  $L$  recordings results in a greatly reduced capacity, and consequently the smallest-possible distortion must increase. In the large-population, Gaussian-source case,

$$D_{\min} = D_{\max} \exp \left\{ \frac{-1.58(2L - 1)\lambda_{\max}}{e \cdot W} \right\}.$$

If the population contains connection-induced dependencies, the argument of the exponential may be somewhat larger. Increasing the number of aggregated recordings made from a population increases the capacity proportionally, but only if they summarize well-chosen sub-populations. In any of these cases, the distortion no longer depends on the population size, but on the number of electrodes used to acquire summed extracellular signals. Thus, the usefulness of unsorted data, both for providing interpretative insights and achieving precise control operations, must be greatly reduced in comparison to sorted or intracellular recordings.

The aggregated-output capacity results sharply differ with the complementary situation, in which a single input is common to each population constituent. From a neural prosthetics viewpoint, this situation models an electrical stimulation device that drives many nearby neurons with a single current source. When there are no connection-induced dependencies between neurons, providing a single input to a population does not reduce capacity from its individual-input value.

### *Model limitations*

Strictly speaking, the capacity results derived here apply *only* to Poisson processes, not point processes in general. Since actual neural systems clearly do not obey Poisson statistics, this may appear to limit the applicability of our results to the understanding of actual neural coding. However, as Kabanov (1978) showed, the Poisson process's capacity applies to any other single point process restricted to the same minimal and maximal rates. In fact, we can show that single non-Poisson channels have a strictly smaller capacity than a Poisson channel having the same maximal instantaneous rate (Johnson 2007), and we conjecture that our results for vector Poisson channels apply to more general point process channels in the same way: whatever the capacities may be for jointly non-Poisson processes, they cannot exceed the values derived here. Thus we believe, but have not proved, that actual neural systems can only have capacities smaller than or equal to the values derived here.

The two basic input models we considered describe extremes of innervation, from a single input driving every neuron to individual innervation. If no connection-induced dependence is present, the capacities at these extremes are the same. It is tempting to infer that without connection-induced dependence, the innervation pattern does not affect the capacity. However, more complicated innervation patterns not captured by our model (for example, when each neuron can receive multiple inputs) may well result in different capacities. When connection-induced dependence is present, our analysis did reveal differences in how the innervation pattern affects capacity. Nevertheless, we still cannot speculate on the interplay between capacity and innervation for more complicated models of population innervation.

The capacity-achieving instantaneous rate variations that ubiquitously achieve capacity are random telegraph waves that swing between the minimal (zero) and the maximal rate extremes. Information theoretical results suggest that channel codes that provide the most effective interface to a channel should have empirical amplitude distributions that mimic the capacity-achieving distribution (Shamai (Shitz) and Verdú 1977). Recorded neural rates do not resemble random telegraph waves, but such observations do not necessarily mean that neural codes are informationally inefficient. Deviations from the simple situations used here – population innervation, varying individual and joint neural behaviors and the non-Poisson nature of discharge patterns – could easily require different capacity achieving inputs. However, if our speculation is correct, the results presented here still provide an upper bound on the capacity in any situation.

## Acknowledgments

The authors thank the anonymous reviewers for their suggestions, which served to improve the paper's clarity.

## Note

- [1] Although differing in detail, the capacity of additive Gaussian noise channels also increases with increasing correlation among the channels in the independent-input case.

## References

- Andersen RA, Musallam S, Pesaran B. 2004. Selecting the signals for a brain-machine interface. *Current Opinion in Neurobiology* 14:1–7.
- Berger T. 1971. *Rate distortion theory*. NJ: Prentice-Hall: Englewood Cliffs.
- Bialek W, DeWeese M, Rieke F, Warlan D. 1993. Bits and brains: Information flow in the nervous system. *Physica A* 200:581–593.
- Brémaud P. 1981. *Point processes and queues*. New York: Springer-Verlag.
- Cover TM, Thomas JA. 2006. *Elements of information theory*. 2nd ed. New York: Wiley.
- Daley DJ, Vere-Jones D. 1988. *An introduction to the theory of point processes*. New York: Springer-Verlag.
- Davis MHA. 1980. Capacity and cutoff rate for Poisson-type channels. *IEEE Transactions on Information Theory* 26:710–715.
- Goodman IN. 2004. Analyzing statistical dependencies in neural populations. Master's thesis, Dept. Electrical & Computer Engineering, Rice University, Houston, Texas.
- Holgate P. 1964. Estimation for the bivariate Poisson distribution. *Biometrika* 51:241–245.
- Johnson DH. 2007. The capacity of non-Poisson channels. In preparation.
- Johnson DH. 1996. Point process models of single-neuron discharges. *Journal of Computational Neuroscience* 3:275–299.
- Johnson DH, Goodman IN. 2007. Jointly Poisson processes. In preparation.
- Johnson DH, Gruner CM, Baggerly K, Seshagiri C. 2001. Information-theoretic analysis of neural coding. *Journal of Computational Neuroscience* 10:47–69.
- Kabanov YM. 1978. The capacity of a channel of the Poisson type. *Theory of Probability and its Applications* 23:143–147.
- Rieke F, Warland D, Bialek W. 1993. Coding efficiency and information rates in sensory neurons. *Europhysics Letters* 22:151–156.
- Rieke F, Warland D, de Ruyter van Steveninck R, Bialek W. 1999. *Spikes: Exploring the neural code*. Cambridge, Massachusetts: MIT Press.
- Schneidman E, Slomin N, Tishby N, de Ruyter van Steveninck RR, Bialek W. 2001. Analyzing neural codes using the information bottleneck method. In: Dietterich TG, Becker S, Ghahramani Z, editors. *Advances in neural information processing systems*. Vol. 14. Cambridge, MA: MIT Press.
- Shamai (Shitz) S, Verdú S. 1977. Empirical distribution for good codes. *IEEE Trans. Info. Th.* 43:836–846.
- Shannon CE. 1948. A mathematical theory of communication. *Bell System Technical Journal* 27:379–423, 623–656.
- Snyder DL. 1998. *Random point processes*. New York: Wiley.
- Strong SP, Koberle R, de Ruyter van Steveninck RR, Bialek W. 1998. Entropy and information in neural spike trains. *Physical Review Letters* 80:197–200.