entropy

MDPI

Article

Limitations to Estimating Mutual Information in Large Neural Populations

Jan Mölter and Geoffrey J. Goodhill *

Queensland Brain Institute & School of Mathematics and Physics, The University of Queensland, St. Lucia, QLD 4072, Australia; jan.moelter@uq.net.au

* Correspondence: g.goodhill@uq.edu.au

Received: 2 March 2020; Accepted: 22 April 2020; Published: 24 April 2020

Abstract: Information theory provides a powerful framework to analyse the representation of sensory stimuli in neural population activity. However, estimating the quantities involved such as entropy and mutual information from finite samples is notoriously hard and any direct estimate is known to be heavily biased. This is especially true when considering large neural populations. We study a simple model of sensory processing and show through a combinatorial argument that, with high probability, for large neural populations any finite number of samples of neural activity in response to a set of stimuli is mutually distinct. As a consequence, the mutual information when estimated directly from empirical histograms will be equal to the stimulus entropy. Importantly, this is the case irrespective of the precise relation between stimulus and neural activity and corresponds to a maximal bias. This argument is general and applies to any application of information theory, where the state space is large and one relies on empirical histograms. Overall, this work highlights the need for alternative approaches for an information theoretic analysis when dealing with large neural populations.

Keywords: sensory coding; information theory; entropy; sampling bias

1. Introduction

The neural code is inherently stochastic, so that even when the same stimulus is repeated, we observe a considerable variability in the neural response. Thus, an important question when considering the interplay between sensory stimuli and neural activity, and when trying to understand how information is represented and processed in neural systems, is how much an observer can actually infer about the stimulus from only looking at its representation in the neural activity. This problem can be formulated and quantitatively studied in the general framework of information theory.

With neural information processing happening at the level of populations, any such analysis necessarily has to consider the activity of an entire population [1,2]. The number of neurons which can be recorded simultaneously has increased in recent years, driven by technologies such as calcium fluorescence imaging [3] and high-density electrode arrays [4]. These methods allow recording of population activity from thousands of neurons simultaneously. However, at these scales, information theoretic analyses of neural activity become considerably harder. Critically, any information theoretic analysis depends at some point on the precise knowledge of the joint probability distribution of the states of the stimuli and the neural population.

However, estimating this distribution from data, and thus entropy and mutual information, is notoriously difficult. This problem is well known, in particular, when the state space for the neural activity becomes large, as is the case when considering large neural populations [5–9].

This problem is generally addressed by applying corrections derived from asymptotic considerations [10], or shuffling/bootstrap procedures together with extrapolations [11–17]. A crucial result in this context is the inconsistency theorem, which implies that under some circumstances estimators that are frequently employed will yield arbitrarily poor estimates [18]. More recently, and, in particular, with regard to entropy estimation, there have been advances towards new estimators or model-based approaches for estimating the aforementioned joint probability distributions [19–24].

In the following we revisit the direct (and model-independent) estimate of entropy and mutual information from experimental histograms. Rhee et al. [7] noted that estimated probability distributions from data samples appear "thinner" and that this leads to what has become known as the sampling bias. Moreover, according to these authors, ensuring the state space is adequately sampled becomes proportionally more difficult with the size of the neural population. The aim of this article is to make this intuition precise, and to show that under certain assumptions with increasing population size any such direct estimate of mutual information will with high probability only yield the stimulus entropy. This emphasises the problems of an information theoretic analysis on the basis of experimental histograms and highlights the importance of new approaches when it comes to estimating information theoretic quatities for large neural populations.

2. Results

In a typical experiment investigating sensory processing, one studies the activity from a neural population of interest in response to a defined set of stimuli. Due to the stochastic nature of neural activity, every stimulus is presented separately multiple times, yielding a finite set of samples of neural activity for every stimulus [25]. From these samples across all stimuli, one can then estimate the relative frequency of patterns of neural activity in response to different stimuli in order to, for example, calculate the dependence in terms of mutual information between the stimuli and the neural activity [26,27].

To make this paradigm more concrete, suppose we are interested in the representation of the visual field in the primary visual cortex. In order to probe this representation, we simultaneously record the spiking activity from a neural population in the respective brain area when exposed to a visual stimulus whose position we vary in the course of the experiment. We can sample the neural activity for every presentation of the stimulus, by measuring the neural activity, for example, in terms of the number of spikes. In order then to quantify the representation of the different stimulus values in their entirety, we can collect the samples of neural activity that we observed, compute empirical histograms and perform an information theoretic analysis.

The central objects of interest in information theory are random variables and their distributions. At its core one defines a measure of their uncertainty, the entropy [28]. If X is a discrete random variable taking values in an at most countable set, its entropy $H(X)$ is defined as

$$H(X) := -\sum_x \mathbb{P}\left[X = x\right] \log \mathbb{P}\left[X = x\right],$$

where $\mathbb{P}\left[X = \cdot\right]$ denotes the probability distribution of X, and therefore $\mathbb{P}\left[X = x\right]$ the probability the X attains the value x. In this definition, we adopt the customary convention that $0 \log 0 := 0$. For the purpose of this paper, we leave the base of the logarithm unspecified; however, we note that in the context of information theory the base is generally chosen to be 2 so that entropy will be measured in units of bits.

If X' is a second random variable, the conditional entropy of X given X', $H(X|X')$, is defined as

$$H\big(X|X'\big) := -\sum_{x,x'} \mathbb{P}\left[X = x \wedge X' = x'\right] \log \mathbb{P}\left[X = x \middle| X' = x'\right],$$

where $\mathbb{P}[X = \cdot \wedge X' = \cdot]$ and $\mathbb{P}[X = \cdot \,|\, X' = \cdot]$ denote the joint probability distribution of X and X' and conditional probability distribution of X given X', respectively. The conditional entropy is a measure for the residual uncertainty in a random variable given observation of another random variable. Specifically, we have that $H(X|X') \leq H(X)$ with equality if X and X' are independent.

An important quantity in information theory is the mutual information between two random variables, X and X', which is defined as

$$\mathrm{MI}(X; X') = H(X) - H(X|X').$$

Mutual information quantifies the amount of entropy of X that knowledge of X' annihilates or, in other words, the information that X' holds about X. Furthermore, as the mutual information is symmetric in its arguments it likewise also quantifies the information X holds about X'. Mutual information is non-negative and vanishes if and only if X and X' are independent. Because of this, mutual information is frequently used as a measure of the independence of two random variables. The mutual information between X and X' can be written in terms of the relative entropy (Kullback–Leibler divergence) between their joint probability distributions and the product of the corresponding marginal distributions, which defines a premetric on the probability distribution.

Importantly, we note that, in contrast to what the notation suggests, both the entropy and the mutual information do not depend on the random variables and their values themselves, but are rather functionals of their distributions [29]. As a consequence of that, we refrain from specifying the random variables' codomains in the following and if a random variable X attains the value x, x may simply be regarded as a label in an abstract alphabet.

2.1. Analysis of a Computational Model of Sensory Processing Regarding Estimating Information Theoretic Quantities

Building on the experimental paradigm we outlined above, we devise and analyse a simple computational model. We assume the stimulus to be modelled by a discrete, almost surely non-constant random variable S and the neural population's activity by the discrete vector-valued random variable $X \equiv \bigotimes_{n=1}^{N} X^{(n)}$, where N is the size of the population. As the stimulus is determined by the experimental protocol, we assume perfect knowledge of its statistics. In addition, we assume that for every stimulus value s there exist a subset of the whole neural population $U_{\perp\!\!\!\perp s} \subseteq \{1, \dots N\}$ such that $N_{s,s'} := |U_{\perp\!\!\!\perp s} \cap U_{\perp\!\!\!\perp s'}| = \omega(\ln N)$ as $N \to \infty$, i.e., $N_{s,s'}$ asymptotically grows faster than the logarithm, for all stimulus values s and s' such that the components $X^{(n)} \,|\, S \in \{s, s'\}$ for $n \in U_{\perp\!\!\!\perp s} \cap U_{\perp\!\!\!\perp s'}$ are independent and identically distributed and almost surely non-constant (Figure 1). In an experimental setting, one might think of $U_{\perp\!\!\!\perp s}$ as the set of neurons that are not receptive to a stimulus value s and therefore activated independently according to a common noise profile. Importantly, these sets differ in general from stimulus to stimulus. However for every two stimulus values, they overlap on a sufficiently large common set. Moreover, in contrast to what one might initially think, these sets cannot be excluded since one can imagine a scenario where every neuron in the population is receptive to at least one of the stimulus values and is simultaneously an element of at least one independent set for other stimulus values.

Next, suppose that for every stimulus value s we are given K_s independent samples from $\mathbb{P}[X = \cdot \,|\, S = s]$, which we denoted as $x^{(s)} := \{x_k^{(s)}\}_{k=1,\dots K_s}$. Based on these samples the empirical estimate for $\mathbb{P}[X = \cdot \,|\, S = \cdot]$, $\hat{\mathbb{P}}[X = \cdot \,|\, S = \cdot]$, is

$$\hat{\mathbb{P}}[X = x | S = s] = \frac{1}{K_s} \sum_{k=1}^{K_s} \delta_{x, x_k^{(s)}}$$

for any sample x and stimulus value s. The Kronecker delta here attains the value 1 whenever its arguments coincide, and otherwise vanishes. Again, in the experimental setting, one might think of $x^{(s)}$ as the samples of neural activity that was recorded in response to a stimulus value s.

Consistent with the intuition of Rhee et al. [7], with high probability, the samples for every stimulus value are mutually different, as we are considering larger and larger neural populations and, moreover, these samples even become uniquely associated with one of the stimulus values, i.e., $x^{(s)} \cap x^{(s')} = \varnothing$ for $s \neq s'$. This simplifies the empirical estimate for $\mathbb{P}[X = \cdot \,|\, S = \cdot\,]$ from above so that, for a sample $x_k^{(s)}$ and a stimulus value s', $\hat{\mathbb{P}}\left[X = x_k^{(s)} \middle| S = s'\right] = \frac{1}{K_{s'}}\delta_{s,s'}$, using that every sample is uniquely associated with a stimulus value and occured only once for those stimulus value. As we will show, the probability for this event is at least $1 - \mathcal{O}(N^{-\infty})$ as $N \to \infty$. Note, that for a sequence $(z_N)_{N\in\mathbb{N}}$ we say $z_N = \mathcal{O}(N^{-\infty})$ if $z_N = \mathcal{o}(N^{-m})$ for every $m \geq 0$, i.e., $\lim_{N\to\infty} N^m z_N = 0$. This not only implies that in the limit of an infinitely large population, the probability is 1, but also makes a statement about the dependence on the size of population. In fact, the probability approaches 1 eventually faster than any polynomial, so that it will be close to 1 even for moderately sized populations.

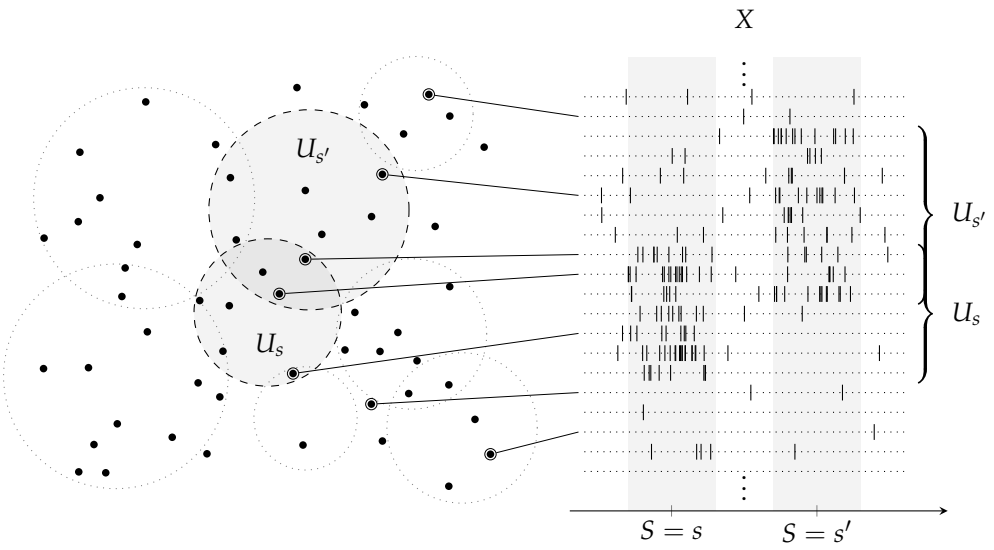

Figure 1. A computational model of sensory processing. We consider a population of N neurons, illustrated irrespective of the physical location as points in the plane on the left-hand side that is exposed to the presentation of a stimulus, which is modelled by a random variable S. The (measured) activity of this population, illustrated as spiking activity on the right-hand side over the presentation of two different values of the stimulus, is modelled by a vector-valued random variable $X \equiv \bigotimes_{n=1}^{N} X^{(n)}$. For every stimulus value s, we assume there exists a subset of the population, U_s, depicted as circular regions on the plane, such that, intuitively, the neurons within this subpopulation are receptive to the particular stimulus value. Conversely, the neurons in the complement of that set in $\{1, \ldots N\}$, $U_{\perp\!\!\!\perp s}$, are assumed to be not receptive to that stimulus value and to activate independently according to a common noise profile. As an example, the sets for stimulus values s and s' are highlighted. Neurons from any of the two sets are shown to have an increased spiking activity during the presentation of the corresponding stimulus value. In contrast to this activity, others are shown to be rather sporadically active.

For any two stimuli there exists by assumption a subset of the components of X which is independent and identically distributed when conditioned on either of the two stimuli. As it suffices that the samples differ in these components for them to be mutually different, the probability for this event is a lower

bound on the probability that the samples are mutually different. Therefore, the probability for all samples corresponding to stimulus s and s' to be mutually different is at least $1 - \mathcal{O}(N^{-\infty})$, for N large. As we show in the next section, this is the probability that a finite number of independent samples from a random vector of length $N_{s,s'}$ with $N_{s,s'} = \omega(\ln N)$ are mutually different, provided the components are independent and identically distributed and almost surely non-constant, which is the case by assumption.

From that the probability that this event occurs for all stimulus values can be estimated to be at least $1 - \sum_{s,s'} (1 - (1 - \mathcal{O}(N^{-\infty}))) = 1 - \mathcal{O}(N^{-\infty})$.

In that event, when all the samples are mutually different, we compute for the entropy and conditional entropy, $\hat{H}(X)$ and $\hat{H}(X|S)$, using the law of total probability to express both in terms of the distribution of S and the empirical estimate of the distribution of X given S, the fact that the latter vanishes away from the samples and that, as we have seen above, for a sample $x_k^{(s)}$ and a stimulus value s', it takes the form $\hat{\mathbb{P}}\left[X = x_k^{(s)} \middle| S = s'\right] = \frac{1}{K_{s'}} \delta_{s,s'}$,

$$\begin{aligned}
\hat{H}(X) &= -\sum_x \hat{\mathbb{P}}\left[X = x\right] \log \hat{\mathbb{P}}\left[X = x\right] \\
&= -\sum_s \mathbb{P}\left[S = s\right] \sum_x \hat{\mathbb{P}}\left[X = x \middle| S = s\right] \log \sum_{s'} \mathbb{P}\left[S = s'\right] \hat{\mathbb{P}}\left[X = x \middle| S = s'\right] \\
&= -\sum_s \mathbb{P}\left[S = s\right] \sum_{k=1}^{K_s} \hat{\mathbb{P}}\left[X = x_k^{(s)} \middle| S = s\right] \log \sum_{s'} \mathbb{P}\left[S = s'\right] \hat{\mathbb{P}}\left[X = x_k^{(s)} \middle| S = s'\right] \\
&= -\sum_s \mathbb{P}\left[S = s\right] \log \mathbb{P}\left[S = s\right] \frac{1}{K_s} \\
&= H(S) + \mathbb{E}\left[\log K_S\right]
\end{aligned}$$

and

$$\begin{aligned}
\hat{H}(X|S) &= -\sum_{s,x} \hat{\mathbb{P}}\left[X = x \wedge S = s\right] \log \hat{\mathbb{P}}\left[X = x \middle| S = s\right] \\
&= -\sum_s \mathbb{P}\left[S = s\right] \sum_x \hat{\mathbb{P}}\left[X = x \middle| S = s\right] \log \hat{\mathbb{P}}\left[X = x \middle| S = s\right] \\
&= -\sum_s \mathbb{P}\left[S = s\right] \sum_{k=1}^{K_s} \hat{\mathbb{P}}\left[X = x_k^{(s)} \middle| S = s\right] \log \hat{\mathbb{P}}\left[X = x_k^{(s)} \middle| S = s\right] \\
&= -\sum_s \mathbb{P}\left[S = s\right] \log \frac{1}{K_s} \\
&= \mathbb{E}\left[\log K_S\right].
\end{aligned}$$

Thus, $\hat{\mathrm{MI}}(X;S) = \hat{H}(X) - \hat{H}(X|S) = H(S)$, and in addition we also get from the above that $\hat{H}(X,S) = \hat{H}(X)$ and that $\hat{H}(S|X) = 0$. As for mutual information the classical bound $\mathrm{MI}(X;S) \leq \min(H(X), H(S))$ holds, mutual information is in fact estimated to be maximal.

Importantly, in this analysis we did not make any assumptions about the precise dependence between S and X apart from the existence of a sufficiently large subpopulation $U_{\perp\!\!\!\perp s}$ for every stimulus value s. Therefore, the conclusion holds also if S and X were, in fact, independent and mutual information should have vanished.

2.2. Computing the Probability of a Finite Number of Independent and Identically Distributed Random Variables being Mutually Different

In the previous section, we were interested in the probability that a finite set of independent and identically distributed (discrete) random variables yields mutually different realisations. This probability is trivially 0 by the pigeonhole principle if there are more random variables in any such set than the number of values these random variables attain with non-vanishing probability. However, once the number of values that are attained exceeds the number of random variables this is not the case anymore, and for any arbitrary distribution of the random variable it is not obvious how to obtain the exact probabilities.

We first take a step back and consider generally series of the form

$$\sum_{\substack{n_1,\dots n_K\in\mathbb{N}\\ n_1\neq n_2\neq\dots n_K}} a_{n_1}\cdots a_{n_K}$$

for sequences $(a_n)_{n\in\mathbb{N}}$ such that the corresponding series is absolutely convergent. This specifically guarantees that the series in question will also converge. We will show how to evaluate such a series and from a closed-form expression derive the probabilities for any finite set of random variables to yield mutually different realisations. From this it will follow, in particular, that large random vectors yield mutually different values with high probability.

Let $(a_n)_{n\in\mathbb{N}}$ be a sequence with $a_n\in\mathbb{R}$ for every $n\in\mathbb{N}$ such that $\sum_{n\in\mathbb{N}} a_n$ is absolutely convergent and define $Q_m := \sum_{n\in\mathbb{N}} a_n^m$ for $m\in\mathbb{N}$. Then, for any $K\in\mathbb{N}$ we have that

$$\sum_{\substack{n_1,\dots n_K\in\mathbb{N}\\ n_1\neq n_2\neq\dots n_K}}\prod_{k=1}^{K} a_{n_k} = \sum_{\alpha\in\mathbb{N}_0^K:\langle\alpha\rangle_K=K} (-1)^{K-|\alpha|}\,\frac{K!}{\alpha!}\prod_{m=1}^{K}\left(\frac{Q_m}{m}\right)^{\alpha_m}.$$

Here and in the following we use the conventional notation for multi-indices $\alpha\in\mathbb{N}_0^K$, so that, e.g., $|\alpha| = \sum_{k=1}^K \alpha_k$ and $\alpha! = \prod_{k=1}^K \alpha_k!$. In addition, we set $\langle\alpha\rangle_k = \sum_{m=1}^k \alpha_m m$ for $1\leq k\leq K$.

We will prove this assertion in two steps. We will first derive a recursion relation for the series in question and then in a second step conclude with an induction argument.

We consider a family of operators $\Gamma_1,\dots\Gamma_K$, and for $1\leq k\leq K$, formally define the corresponding operator Γ_k on monomials of a_{n_k} via

$$\Gamma_k : a_{n_k}^m \mapsto \sum_{n_k\in\mathbb{N}\setminus\{n_1,\dots n_{k-1}\}} a_{n_k} a_{n_k}^m = Q_{m+1} - \sum_{r=1}^{k-1} a_{n_r}^{m+1}$$

and extend it linearly to span $\left\{1, a_{n_k}, a_{n_k}^2,\dots\right\}$. Using these operators we find that

$$A_K := \sum_{\substack{n_1,\dots n_K\in\mathbb{N}\\ n_1\neq n_2\neq\dots n_K}}\prod_{k=1}^{K} a_{n_k} = \sum_{n_1\in\mathbb{N}} a_{n_1}\sum_{n_k\in\mathbb{N}\setminus\{n_1\}} a_{n_2}\cdots\sum_{n_K\in\mathbb{N}\setminus\{n_1,\dots n_{K-1}\}} a_{n_K} = (\Gamma_1\circ\Gamma_2\circ\cdots\Gamma_K)(1)\,.$$

In particular, we have for every $1\leq k'\leq k\leq K$ and $m\in\mathbb{N}_0$ that $(\Gamma_1\circ\Gamma_2\circ\cdots\Gamma_k)\left(a_{n_{k'}}^m\right) = \sum_{\substack{n_1,\dots n_k\in\mathbb{N}\\ n_1\neq n_2\neq\dots n_k}} a_{n_1}\cdots a_{n_{k'}}^{m+1}\cdots a_{n_k} = \sum_{\substack{n_1,\dots n_k\in\mathbb{N}\\ n_1\neq n_2\neq\dots n_k}} a_{n_1}\cdots a_{n_k}^{m+1} = (\Gamma_1\circ\Gamma_2\circ\cdots\Gamma_k)\left(a_{n_k}^m\right)$, where we have relabeled the indices and the interchangability of the sums is guaranteed by the absolute convergence of every individual series. Therefore, $(\Gamma_1\circ\Gamma_2\circ\cdots\Gamma_k)\left(a_{n_{k'}}^m\right) = (\Gamma_1\circ\Gamma_2\circ\cdots\Gamma_k)\left(a_{n_k}^m\right)$.

Now, using the linearity of the operators $\Gamma_1, \dots \Gamma_K$ we have that

$$\begin{aligned}(\Gamma_1 \circ \Gamma_2 \circ \cdots \Gamma_K)\big(a_{n_K}^m\big) &= (\Gamma_1 \circ \Gamma_2 \circ \cdots \Gamma_{K-1})\left(Q_{m+1} - \sum_{r=1}^{K-1} a_{n_r}^{m+1}\right) \\ &= Q_{m+1}\,(\Gamma_1 \circ \Gamma_2 \circ \cdots \Gamma_{K-1})(1) - \sum_{r=1}^{K-1} (\Gamma_1 \circ \Gamma_2 \circ \cdots \Gamma_{K-1})\left(a_{n_r}^{m+1}\right) \\ &= Q_{m+1} A_{K-1} - (K-1)\,(\Gamma_1 \circ \Gamma_2 \circ \cdots \Gamma_{K-1})\left(a_{n_{K-1}}^{m+1}\right) \\ &= \sum_{k=1}^{K} (-1)^{k+1}\,\frac{(K-1)!}{(K-k)!} Q_{m+k} A_{K-k}.\end{aligned}$$

In the last step we explicitly resolved the recursion, which can be verified by a simple inductive argument. In particular, for $m = 0$ we obtain an expression for A_K so that we conclude after reordering the sum and rearranging the terms that

$$\frac{A_K}{(-1)^K K!} = -\frac{1}{K} \sum_{k=0}^{K-1} Q_{K-k} \frac{A_k}{(-1)^k k!}.$$

As A_K is precisely the sum we intend to compute, the last expression yields a recursion relation for it with initial datum $A_0 = 1$. In particular, we note that this equation is reminiscent of a discrete Volterra equation of convolution type. Thus, this concludes the first step of the argument and in the second step we will show that the expression stated in the beginning actually solves this recursion relation.

For the induction argument we assume now that the assertion holds for $1 \leq k \leq K-1$ in order to perform the step $K-1 \to K$ and we use the recursion relation derived above to relate A_K to $A_0, A_1, \dots A_{K-1}$. In addition, we denote with e_k the multi-index that is 0 in every component except the k^{th} one, where it is 1.

$$\begin{aligned}\frac{A_K}{(-1)^K K!} &= -\frac{1}{K} \sum_{k=0}^{K-1} Q_{K-k} \frac{1}{(-1)^k k!} \left\{ \sum_{\alpha \in \mathbb{N}_0^k : \langle \alpha \rangle_k = k} (-1)^{k-|\alpha|}\, \frac{k!}{\alpha!} \prod_{m=1}^{k} \left(\frac{Q_m}{m}\right)^{\alpha_m} \right\} \\ &= -\frac{1}{K} \sum_{k=0}^{K-1} \sum_{\alpha \in \mathbb{N}_0^K : \langle \alpha \rangle_K = k} \frac{(-1)^{-|\alpha|}}{\alpha!} Q_{K-k} \prod_{m=1}^{K} \left(\frac{Q_m}{m}\right)^{\alpha_m} \\ &= -\frac{1}{K} \sum_{k=1}^{K} \sum_{\alpha \in \mathbb{N}_0^K : \langle \alpha \rangle_K = K-k} \frac{(-1)^{-|\alpha|}}{\alpha!} Q_k \prod_{m=1}^{K} \left(\frac{Q_m}{m}\right)^{\alpha_m} \\ &= \sum_{k=1}^{K} \sum_{\alpha \in \mathbb{N}_0^K : \langle \alpha \rangle_K = K-k} \frac{(\alpha_k + 1)\,k}{K} \frac{(-1)^{-|\alpha+e_k|}}{(\alpha + e_k)!} \prod_{m=1}^{K} \left(\frac{Q_m}{m}\right)^{(\alpha+e_k)_m} \\ &= \sum_{\alpha \in \mathbb{N}_0^K : \langle \alpha \rangle_K = K} \frac{(-1)^{-|\alpha|}}{\alpha!} \prod_{m=1}^{K} \left(\frac{Q_m}{m}\right)^{\alpha_m}\end{aligned}$$

In the last step, we used that, as we will show, for any multi-index functional Φ,

$$\sum_{k=1}^{K} \sum_{\alpha \in \mathbb{N}_0^K : \langle \alpha \rangle_K = K-k} \frac{(\alpha_k + 1)\,k}{K} \Phi(\alpha + e_k) = \sum_{\alpha \in \mathbb{N}_0^K : \langle \alpha \rangle_K = K} \Phi(\alpha)\,.$$

Indeed, we first observe that the terms of Φ that appear in both sums are identical. Therefore, we only have to carefully evaluate the coefficients.

$$\begin{aligned}
&\sum_{k=1}^{K} \sum_{\alpha\in\mathbb{N}_0^K:\langle\alpha\rangle_K=K-k} \frac{(\alpha_k+1)\,k}{K}\Phi(\alpha+e_k)\\
&= \sum_{\alpha\in\mathbb{N}_0^K:\langle\alpha\rangle_K=K} \sum_{k=1}^{K} \sum_{\alpha'\in\mathbb{N}_0^K:\langle\alpha'\rangle_K=K-k} \frac{(\alpha'_k+1)\,k}{K}\Phi\left(\alpha'+e_k\right)\delta_{\alpha,\alpha'+e_k}\\
&= \sum_{\alpha\in\mathbb{N}_0^K:\langle\alpha\rangle_K=K} \Phi(\alpha) \sum_{k=1}^{K} \frac{\alpha_k k}{K} \sum_{\alpha'\in\mathbb{N}_0^K:\langle\alpha'\rangle_K=K-k} \delta_{\alpha,\alpha'+e_k}\\
&= \sum_{\alpha\in\mathbb{N}_0^K:\langle\alpha\rangle_K=K} \Phi(\alpha) \sum_{k=1}^{K} \frac{\alpha_k k}{K}\\
&= \sum_{\alpha\in\mathbb{N}_0^K:\langle\alpha\rangle_K=K} \Phi(\alpha)
\end{aligned}$$

This concludes the argument, so that as claimed

$$\sum_{\substack{n_1,\dots n_K\in\mathbb{N}\\ n_1\neq n_2\neq\dots n_K}} \prod_{k=1}^{K} a_{n_k} = \sum_{\alpha\in\mathbb{N}_0^K:\langle\alpha\rangle_K=K} (-1)^{K-|\alpha|}\,\frac{K!}{\alpha!}\prod_{m=1}^{K}\left(\frac{Q_m}{m}\right)^{\alpha_m}.$$

In order to apply this result now in a probabilistic context to answer the question about the probability that any finite set of (discrete) random variables yields mutually different realisations, consider a discrete random variable X that takes values in the set $\{x_1, x_2, \dots\}$, which we assume without loss of generality to be countably infinite. Then, applying the above result to the sequence $(\mathbb{P}[X=x_n])_{n\in\mathbb{N}}$ we can compute the probability that K independent copies of this random variable, $X_1, \dots X_K$, all yield mutually different realisations. Specifically, we have that

$$\begin{aligned}
\mathbb{P}[X_1\neq X_2\neq\cdots X_K] = \mathbb{E}\left[\mathbb{1}_{X_1\neq X_2\neq\cdots X_K}\right] &= \sum_{x_1\neq x_2\neq\cdots x_K}\prod_{k=1}^{K}\mathbb{P}[X=x_k]\\
&= \sum_{\alpha\in\mathbb{N}_0^K:\langle\alpha\rangle_K=K} (-1)^{K-|\alpha|}\,\frac{K!}{\alpha!}\prod_{m=1}^{K}\left(\frac{Q_m}{m}\right)^{\alpha_m}.
\end{aligned}$$

with $Q_m = \sum_{n\in\mathbb{N}}\mathbb{P}[X=x_n]^m$.

Now, consider random vectors of increasing length, $X \equiv \bigotimes_{n=1}^{f(N)} X^{(n)}$, with independent components and assume that there exists $\epsilon > 0$ such that $Q_2^{(n)} = \sum_x \mathbb{P}[X^{(n)}=x]^2 \leq 1-\epsilon$ for every $n\in\mathbb{N}$. For the latter a necessary condition is that $X^{(n)}$ is almost surely non-constant and in particular it is satisfied if the components are identically distributed and almost surely non-constant. Then, if $f(N) = \omega(\ln N)$ increasing, we will show that this implies that in the limit $N\to\infty$

$$\mathbb{P}[X_1\neq X_2\neq\cdots X_K] = 1-\mathcal{O}\left(N^{-\infty}\right),$$

where we recall that for some sequence $(z_N)_{N\in\mathbb{N}}$ we have that $z_N = \mathcal{O}(N^{-\infty})$ if for every $m\geq 0$ $\lim_{N\to\infty} N^m z_N = 0$.

Indeed, $0 < Q_m = \prod_{n=1}^{f(N)} Q_m^{(n)} \leq \left(\sup_{n\in\mathbb{N}} Q_m^{(n)}\right)^{f(N)} \leq \left(\sup_{n\in\mathbb{N}} Q_2^{(n)}\right)^{f(N)} \leq (1-\epsilon)^{f(N)}$ for every $m \geq 2$ so that we conclude that $Q_m \leq \mathrm{e}^{-|\ln(1-\epsilon)|f(N)} = \mathcal{O}(N^{-\infty})$. This implies the claim since $\mathbb{P}[X_1 \neq X_2 \neq \cdots X_K]$ is a polynomial in terms of $Q_1, \ldots Q_K$ with the only asymptotically non-vanishing term being $Q_1^K = 1$.

Besides the immediate application to random vectors that we required in the last section, we remark that the expression for the probability of random variables being mutually different can also be used to derive a closed-form expression for the Stirling numbers of the first kind. Briefly, let X be a uniform random variable on $\{1, \ldots L\}$ for some $L \in \mathbb{N}$, so that $Q_m = L^{1-m}$ in the above statement. The probability that $K \leq L$ independent copies of this random variable are all mutually different is then given by the above expression. On the other hand, this probability can also be computed purely combinatorially to be $\frac{(L)_K}{L^K}$, where $(\,\cdot\,)_K$ denotes the falling factorial [30]. By definition, $(L)_K = \sum_{n=0}^{K} s(K,n)\, L^n$ with $s(K,n)$ the (signed) Stirling numbers of the first kind. Therefore, comparing the two expressions we find the Stirling numbers to be given as (cf. [31])

$$s(K,n) = (-1)^{K-n} \begin{cases} \sum_{\alpha\in\mathbb{N}_0^K : \langle\alpha\rangle_K = K} \delta_{|\alpha|,n} \frac{K!}{\alpha! \prod_{m=1}^{K} m^{\alpha_m}} & 1 \leq n \leq K \\ 0 & \text{otherwise} \end{cases}.$$

3. Discussion

In this work, we have analysed a concrete, but general, model of sensory processing and demonstrated the extent of the sampling bias when directly estimating information theoretic quantities from experimental histograms. We have shown that as we consider larger and larger neural populations, with high probability any estimate of entropy or mutual information will only depend on the stimulus entropy.

We found that the issue lies in the fact that the samples of neural activity in response to the presentation of different stimulus values turn out to be mutually distinct. One way this can happen is through the existence of a subpopulation of neurons that only contributes independent noise to the population's activity. Importantly, the composition of the subpopulation can be stimulus-dependent, so that it is impossible to exclude this subpopulation from the beginning in the analysis. A plausible origin for these noisy subpopulations are sufficiently localised, compactly supported receptive fields. In a simplified scenario, imagine a neural population, whose neurons are receptive to any one of three stimuli. For each of the stimuli, the neurons receptive to one of the other two stimuli constitute such a noisy subpopulation. Now, as we have mentioned also earlier, none of the three subpopulations can be excluded on the grounds that it contributes only noise across all stimuli, yet their activity lets samples recorded from the population appear more distinctive than they are. While for small neural populations the effect of noisy subpopulations can be counteracted by increasing the number of samples that one considers, this becomes infeasible for even moderately sized populations due to the combinatorially large repertoire of population activity. As we have shown, as we consider larger and larger populations and therefore also larger noisy subpopulations the possibility to sample even one activity pattern twice is exponentially suppressed. In reality, receptive fields are localised, yet at least in computational models their range often extends indefinitely. This is the case, for example, for Gaussian receptive fields and the activity of a neuron then depends strongly on some stimulus values while for others the dependence in vanishingly small. While the assumptions of our model are clearly not met, we argue that for all practical purposes the consequence is the same, although weakened. This is because one will only ever consider a finite number of samples. The smaller the dependence between the stimulus and the activity, the weaker it manifests itself in particular when only considering those finitely many samples, so that corresponding neurons appear essentially independent.

In the model we have proposed, we assumed that for any two stimulus values there exist sufficiently large subpopulations such that their components contribute independent and identically distributed noise when exposed to either of the two stimulus values. Now, this is certainly a simplification and a more realistic assumption would be that in the absence of sensory stimulation the activity in this subpopulations is governed by low-dimensional dynamics in addition to some individual noise [32,33]. In principle though, our conclusions should hold true irrespectively following the same general argument that we formalised in this work: In a large neural population, stochastic variability, even if it is only in a relatively small subpopulation, is sufficient to produce unique samples of neural activity in an experiment. In turn, this then manifests itself in a maximal bias when estimating mutual information and other information theoretic quantities.

Without any further assumptions on the statistical relation between the different neurons within the population, this work shows that, despite being a powerful framework, in general, information theoretic analyses become essentially intractable when considering larger neural populations, because of the difficulties of accurately estimating the joint probability distribution between activity and stimulus states from experimental histograms. Therefore, this is where modelling approaches come into play. These approaches frequently employ maximum entropy Ising models that were fit initially only to low-order [34] and later also higher-order statistics of the data [35–37]. Alternative approaches include the cascaded logistic model which utilises Dirichlet processes at its core [38] or more recently the population tracking model [24]. However, for an information theoretic analysis it is in addition important to be able to compute the involved quantities in an efficient way. While most models include the possibility to sample states, this is not an option for large populations again because of the issues presented in this work. In the context of maximum entropy Ising models, thermodynamic considerations turned out to be useful to efficiently compute quantities such as the entropy [37,39,40]. In the population tracking model, on the other hand, one derives a reduced, low-dimensional model, from which those quantities can be computed likewise [24].

Altogether, while for many statistics of neural activity it might be sufficient to consider experimental histograms, this is not the case for information theoretic quantities when considering large neural populations. Given the many insights that information theoretic analyses can bring [41–43], this motivates new approaches for studying experimental recordings from ever larger neural populations. Furthermore, it also inspires consideration of how the brain handles the informational constraints we have identified.

Author Contributions: Conceptualization, J.M. and G.J.G.; Investigation, J.M.; Supervision, G.J.G.; Writing—original draft, J.M.; Writing—review & editing, J.M. and G.J.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by an Australian Government Research Training Program (RTP) Scholarship awarded to Jan Mölter and the Australian Research Council Discovery Grant DP170102263 awarded to Geoffrey J. Goodhill.

Acknowledgments: We thank Marcus A. Triplett for very helpful discussions and comments on the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Pouget, A.; Dayan, P.; Zemel, R. Information processing with population codes. *Nat. Rev. Neurosci.* **2000**, *1*, 125–132, doi:10.1038/35039062.
2. Sakurai, Y. Population coding by cell assemblies—What it really is in the brain. *Neurosci. Res.* **1996**, *26*, 1–16, doi:10.1016/0168-0102(96)01075-9.
3. Scanziani, M.; Häusser, M. Electrophysiology in the age of light. *Nature* **2009**, *461*, 930–939, doi:10.1038/nature08540.

4. Jun, J.J.; Steinmetz, N.A.; Siegle, J.H.; Denman, D.J.; Bauza, M.; Barbarits, B.; Lee, A.K.; Anastassiou, C.A.; Andrei, A.; Aydın, Ç.; et al. Fully integrated silicon probes for high-density recording of neural activity. *Nature* **2017**, *551*, 232–236, doi:10.1038/nature24636.
5. Quian Quiroga, R.; Panzeri, S. Extracting information from neuronal populations: Information theory and decoding approaches. *Nat. Rev. Neurosci.* **2009**, *10*, 173–185, doi:10.1038/nrn2578.
6. Kinney, J.B.; Atwal, G.S. Equitability, mutual information, and the maximal information coefficient. *Proc. Natl. Acad. Sci. USA* **2014**, *111*, 3354–3359, doi:10.1073/pnas.1309933111.
7. Rhee, A.; Cheong, R.; Levchenko, A. The application of information theory to biochemical signaling systems. *Phys. Biol.* **2012**, *9*, 045011, doi:10.1088/1478-3975/9/4/045011.
8. Dorval, A.D. Estimating Neuronal Information: Logarithmic Binning of Neuronal Inter-Spike Intervals. *Entropy* **2011**, *13*, 485–501, doi:10.3390/e13020485.
9. Macke, J.H.; Murray, I.; Latham, P.E. How biased are maximum entropy models? *Adv. Neural Inf. Process. Syst.* **2011**, *24*, 2034–2042.
10. Panzeri, S.; Senatore, R.; Montemurro, M.A.; Petersen, R.S. Correcting for the Sampling Bias Problem in Spike Train Information Measures. *J. Neurophysiol.* **2007**, *98*, 1064–1072, doi:10.1152/jn.00559.2007.
11. Treves, A.; Panzeri, S. The Upward Bias in Measures of Information Derived from Limited Data Samples. *Neural Comput.* **1995**, *7*, 399–407, doi:10.1162/neco.1995.7.2.399.
12. Panzeri, S.; Treves, A. Analytical estimates of limited sampling biases in different information measures. *Network* **1996**, *7*, 87–107, doi:10.1080/0954898X.1996.11978656.
13. Adibi, M.; McDonald, J.S.; Clifford, C.W.G.; Arabzadeh, E. Adaptation Improves Neural Coding Efficiency Despite Increasing Correlations in Variability. *J. Neurosci.* **2013**, *33*, 2108–2120, doi:10.1523/JNEUROSCI.3449-12.2013.
14. Takaguchi, T.; Nakamura, M.; Sato, N.; Yano, K.; Masuda, N. Predictability of Conversation Partners. *Phys. Rev. X* **2011**, *1*, 011008, doi:10.1103/PhysRevX.1.011008.
15. Pachitariu, M.; Lyamzin, D.R.; Sahani, M.; Lesica, N.A. State-Dependent Population Coding in Primary Auditory Cortex. *J. Neurosci.* **2015**, *35*, 2058–2073, doi:10.1523/JNEUROSCI.3318-14.2015.
16. Lopes-dos Santos, V.; Panzeri, S.; Kayser, C.; Diamond, M.E.; Quian Quiroga, R. Extracting information in spike time patterns with wavelets and information theory. *J. Neurophysiol.* **2015**, *113*, 1015–1033, doi:10.1152/jn.00380.2014.
17. Montgomery, N.; Wehr, M. Auditory Cortical Neurons Convey Maximal Stimulus-Specific Information at Their Best Frequency. *J. Neurosci.* **2010**, *30*, 13362–13366, doi:10.1523/JNEUROSCI.2899-10.2010.
18. Paninski, L. Estimation of Entropy and Mutual Information. *Neural Comput.* **2003**, *15*, 1191–1253, doi:10.1162/089976603321780272.
19. Zhang, Z. Entropy Estimation in Turing's Perspective. *Neural Comput.* **2012**, *24*, 1368–1389, doi:10.1162/NECO_a_00266.
20. Yu, Y.; Crumiller, M.; Knight, B.; Kaplan, E. Estimating the amount of information carried by a neuronal population. *Front. Comput. Neurosci.* **2010**, *4*, 10, doi:10.3389/fncom.2010.00010.
21. Archer, E.W.; Park, I.M.; Pillow, J.W. Bayesian entropy estimation for binary spike train data using parametric prior knowledge. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 1700–1708.
22. Vinck, M.; Battaglia, F.P.; Balakirsky, V.B.; Vinck, A.J.H.; Pennartz, C.M.A. Estimation of the entropy based on its polynomial representation. *Phys. Rev. E* **2012**, *85*, doi:10.1103/PhysRevE.85.051139.
23. Xiong, W.; Faes, L.; Ivanov, P.C. Entropy measures, entropy estimators, and their performance in quantifying complex dynamics: Effects of artifacts, nonstationarity, and long-range correlations. *Phys. Rev. E* **2017**, *95*, 062114, doi:10.1103/PhysRevE.95.062114.
24. O'Donnell, C.; Gonçalves, J.T.; Whiteley, N.; Portera-Cailliau, C.; Sejnowski, T.J. The Population Tracking Model: A Simple, Scalable Statistical Model for Neural Population Data. *Neural Comput.* **2017**, *29*, 50–93, doi:10.1162/NECO_a_00910.
25. Victor, J.D. Approaches to Information-Theoretic Analysis of Neural Activity. *Biol. Theory* **2006**, *1*, 302–316, doi:10.1162/biot.2006.1.3.302.

26. Timme, N.M.; Lapish, C. A Tutorial for Information Theory in Neuroscience. *eNeuro* **2018**, *5*, doi:10.1523/ENEURO.0052-18.2018.
27. Pregowska, A.; Szczepanski, J.; Wajnryb, E. Mutual information against correlations in binary communication channels. *BMC Neurosci.* **2015**, *16*, 32, doi:10.1186/s12868-015-0168-0.
28. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423, doi:10.1002/j.1538-7305.1948.tb01338.x.
29. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2005, doi:10.1002/047174882X.
30. Stanley, R.P. *Enumerative Combinatorics*, 2nd ed.; Cambridge Studies in Advanced Mathematics; Cambridge University Press: Cambridge, UK, 2011, doi:10.1017/CBO9781139058520.
31. Malenfant, J. Finite, closed-form expressions for the partition function and for Euler, Bernoulli, and Stirling numbers. *arXiv* **2011**, arXiv:1103.1585.
32. Stringer, C.; Pachitariu, M.; Steinmetz, N.; Reddy, C.B.; Carandini, M.; Harris, K.D. Spontaneous behaviors drive multidimensional, brainwide activity. *Science* **2019**, *364*, eaav7893, doi:10.1126/science.aav7893.
33. Triplett, M.A.; Pujic, Z.; Sun, B.; Avitan, L.; Goodhill, G.J. Model-based decoupling of evoked and spontaneous neural activity in calcium imaging data. *bioRxiv* **2019**, doi:10.1101/691261.
34. Schneidman, E.; Berry, M.J., II; Segev, R.; Bialek, W. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* **2006**, *440*, 1007–1012, doi:10.1038/nature04701.
35. Granot-Atedgi, E.; Tkačik, G.; Segev, R.; Schneidman, E. Stimulus-dependent Maximum Entropy Models of Neural Population Codes. *PLoS Comput. Biol.* **2013**, *9*, e1002922, doi:10.1371/journal.pcbi.1002922.
36. Tkačik, G.; Marre, O.; Mora, T.; Amodei, D.; Berry, M.J., II; Bialek, W. The simplest maximum entropy model for collective behavior in a neural network. *J. Stat. Mech.* **2013**, *2013*, P03011, doi:10.1088/1742-5468/2013/03/P03011.
37. Tkačik, G.; Marre, O.; Amodei, D.; Schneidman, E.; Bialek, W.; Berry II, M.J. Searching for Collective Behavior in a Large Network of Sensory Neurons. *PLoS Comput. Biol.* **2014**, *10*, e1003408, doi:10.1371/journal.pcbi.1003408.
38. Park, I.M.; Archer, E.W.; Latimer, K.; Pillow, J.W. Universal models for binary spike patterns using centered Dirichlet processes. *Adv. Neural Inf. Process. Syst.* **2013**, *26*, 2463–2471.
39. Tkačik, G.; Schneidman, E.; Berry, M.J., II; Bialek, W. Ising models for networks of real neurons. *arXiv* **2006**, arXiv:q-bio/0611072.
40. Tkačik, G.; Schneidman, E.; Berry, M.J., II; Bialek, W. Spin glass models for a network of real neurons. *arXiv* **2009**, arXiv:0912.5409.
41. Stevens, C.F.; Zador, A.M. Information through a Spiking Neuron. *Adv. Neural Inf. Process. Syst.* **1996**, *8*, 75–81.
42. Strong, S.P.; Koberle, R.; de Ruyter van Steveninck, R.R.; Bialek, W. Entropy and Information in Neural Spike Trains. *Phys. Rev. Lett.* **1998**, *80*, 197–200, doi:10.1103/PhysRevLett.80.197.
43. Borst, A.; Theunissen, F.E. Information theory and neural coding. *Nat. Neurosci.* **1999**, *2*, 947–957, doi:10.1038/14731.