entropy

MDPI

Article

Assessing the Relevance of Specific Response Features in the Neural Code

Hugo Gabriel Eyherabide [1,*] and Inés Samengo [2]

[1] Department of Computer Science and Helsinki Institute for Information Technology, University of Helsinki Gustaf Hällströmin katu 2b, FI00560 Helsinki, Finland
[2] Department of Medical Physics, Centro Atómico Bariloche and Instituto Balseiro, 8400 San Carlos de Bariloche, Argentina; samengo@cab.cnea.gov.ar
* Correspondence: neuralinfo@eyherabidehg.com; Tel.: +358-050-368-1758

Received: 1 October 2018; Accepted: 13 November 2018; Published: 15 November 2018

Abstract: The study of the neural code aims at deciphering how the nervous system maps external stimuli into neural activity—the encoding phase—and subsequently transforms such activity into adequate responses to the original stimuli—the decoding phase. Several information-theoretical methods have been proposed to assess the relevance of individual response features, as for example, the spike count of a given neuron, or the amount of correlation in the activity of two cells. These methods work under the premise that the relevance of a feature is reflected in the information loss that is induced by eliminating the feature from the response. The alternative methods differ in the procedure by which the tested feature is removed, and the algorithm with which the lost information is calculated. Here we compare these methods, and show that more often than not, each method assigns a different relevance to the tested feature. We demonstrate that the differences are both quantitative and qualitative, and connect them with the method employed to remove the tested feature, as well as the procedure to calculate the lost information. By studying a collection of carefully designed examples, and working on analytic derivations, we identify the conditions under which the relevance of features diagnosed by different methods can be ranked, or sometimes even equated. The condition for equality involves both the amount and the type of information contributed by the tested feature. We conclude that the quest for relevant response features is more delicate than previously thought, and may yield to multiple answers depending on methodological subtleties.

Keywords: neural code; representation; decoding; spike-time precision; discrimination; noise correlations; information theory; mismatched decoding

1. Introduction

Understanding the neural code involves, among other things, identifying the relevant response features that participate in the representation of information. Different studies have proposed several candidates, for example, the spiking rate [1,2], the response latency [3], the temporal organisation of spikes [4], the amount of synchrony in a given brain area [5], the amount of correlation between the activity of different neurons [6], or the phase of the local field potential at the time of spiking [7], to cite a few. One way of evaluating the relevance of each candidate feature is to assess how much information is lost by ignoring that feature. This strategy involves the comparison of the mutual information between the stimulus and the so-called *full response* (a collection of response features including the tested one) and the same information calculated with a *reduced response*, obtained by dropping the tested feature from the full response. If the tested feature is relevant, the information encoded by the reduced response should be smaller than that of the full response.

The procedure is fairly straightforward when the response features are defined in terms of variables that take definite values in each stimulus presentation, as for example, the spike count C fired in a fixed time window, or the latency L between the stimulus and the first spike. The full response in this case is a two-component vector $[C, L]$, the value of which is uniquely defined for each stimulus presentation—let us assume that in this example, C is never equal to 0, so L is always well defined. The reduced response is a one-component vector, either C or L, depending whether we are evaluating the relevance of the latency or the spike count, respectively. If the latency or the spike count are relevant, then the information encoded by C or L, respectively, should be smaller than that of the pair $[C, L]$. Throughout this paper, we often use C and L as examples of response features that take a precise value in each trial, to contrast with other features that are only defined in the whole collection of trials, as discussed below.

The method becomes more controversial when applied to response properties that can only be defined in multiple stimulus presentations, as for example, the amount of correlation in the activity of two or more neurons, or the temporal precision of the elicited spikes. These properties cannot be calculated from single responses, so more sophisticated methods are required to delete the tested feature. There are several alternative procedures to perform such deletion, and several are also the ways in which the lost information can be calculated. Interestingly, the lost information depends markedly on the chosen method, implying that the so-called *relevance* of a given feature is a subtle concept, that needs to be specified precisely. When assessing the relevance of noise correlations, two different sets of strategies have been proposed by the seminal works of Nirenberg et al. [8] and Schneidman et al. [9]. The first proposal evaluated the role of noise correlations in *decoding* the information represented in neural activity, whereas the second, in the amount of *encoded* information. Quite surprisingly, the contribution of correlations to the decoded information was shown to sometimes exceed the amount of encoded information [9], seemingly contradicting the intuitive idea that the encoded information constitutes an upper bound to the decoded information. The apparent inconsistency between the two measures has not been observed in later extensions of the technique, where the relevance of other response aspects was evaluated, such as spike-time precision, spike-counts or spike-onsets. Moreover, it has even been argued that the inconsistency was exclusively observed when assessing the role of noise correlations [10–13].

In this paper, for the first time, the different methods used in the literature to delete a given response feature are distinguished, and the implications of each method are discussed and compared. We show that the data processing inequality, stating that the decoded information cannot surpass the encoded information, can only be invoked with some - and not all - deletion procedures. The distinction between such procedures allows us to identify the conditions in which the decoded information can exceed the encoded information, and to demonstrate that there was no logical inconsistency in previous studies. We also show explicit examples where the decoded information surpasses the encoded information also when assessing the role of other response aspects different from noise correlations. In order to explain why such behaviours have not been identified until now, we scrutinise the arguments given in the literature to claim that only noise correlations could exhibit such syndrome. We conclude that although the measures employed to assess the relevance of individual response features initially distinguished clearly between the relevance for encoding and the relevance for decoding, this distinction was eventually lost in later modifications of the measures. By diagnosing the confusion, we prove that indeed, the response features for which the decoded information can surpass the encoded information are not restricted to noise correlations.

More generally, we discuss a wide collection of strategies employed to assess the relevance of individual response features, ranging from those encoded-oriented to those decoded-oriented. This distinction is related to the way the tested feature contributes to the performance of decoders, which can be mismatched or not. The relevance of the tested feature obtained with some of the measures is always bounded by the relevance of another measure. Yet, not all measures can be ordered hierarchically. There are examples where the relevance of a feature obtained with one method may

surpass or be surpassed by the relevance of another, depending on the specific values taken by the prior stimulus probability and the conditional response probabilities. We analyse a collection of carefully chosen examples to identify the cases where this is so. In certain restricted conditions, however, the hierarchy, or even the equality, can be ensured. Here we establish these conditions by means of analytic reasoning, and discuss their implications in terms of the amount and type of information encoded by the tested feature.

We also present examples in which the measures to assess the relevance of a given feature can be used to extract qualitative knowledge about the type of information encoded by the feature. In other words, we assess not only *how much* information is encoded by an individual feature, but also *what kind* of information is provided, with respect to individual stimulus attributes. Again, we prove that the type of encoded information depends on the method employed to assess it.

Finally, given that one important property of measures of relevance hinges on whether they represent the operation of matched or mismatched decoders, we also explore the consequences of operating mismatched decoders on noisy responses, instead of real responses. We conclude that it may be possible to improve the performance of a mismatched decoder by adding noise. From the theoretical point of view, this observation underscores the fact that the conditions for optimality for matched decoders need not hold for mismatched decoders. From the practical perspective, our results open new opportunities for potentially simpler, more efficient and more resilient decoding algorithms.

In Section 2.1, we establish the notation, and we introduce some of the key concepts that will be used throughout the paper. These concepts are employed in Section 2.2 to determine the cases where the data-processing inequality can be ensured. In Section 2.3 we introduce 9 measures of feature relevance that were previously defined in the literature, and briefly discuss their meaning, similarities and discrepancies. A numeric exploration of a set of carefully chosen examples is employed in Section 2.4 to detect the pairs of measures for which no general hierarchical order exists. In Section 2.5 we discuss the consequences of employing measures that are conceptually linked to matched or mismatched decoders. Later, in Section 2.6, we explore the way in which different measures of feature relevance arrogate different qualitative meaning to the type of information encoded by the tested feature. In Section 2.7 we discuss the conditions under which encoding-oriented measures provide the same amount of information as their decoding-oriented counterparts, and also the conditions under which the equality extends also to the content of that information. Then, in Section 2.8, we observe that sometimes, mismatched decoders may improve their performance when operating upon noisy responses. We discuss some relations of our work with other approaches and to the limiting sampling problem in Section 3, and we close with a summary of the main results of the paper in Section 4.

2. Results

2.1. Definitions

2.1.1. Statistical Notation

When no risk of ambiguity arises, we here employ the standard abbreviated notation of statistical inference [14], denoting random variables with letters in upper case, and their values, in lower case. For example, the symbol $P(x|y)$ always denotes the conditional probability of the random variable X taking the value x given that the random variable Y takes the value y. This notation may lead to confusion or be inappropriate, for example, when the random variable X takes the value u given that the random variable Y takes the value v. In those cases, we explicitly indicate the random variables and their values, as for example $P(X = u|Y = v)$.

In the study of the neural code, the relevant random variables are the stimulus S and the response $\mathbf{R}$ generated by the nervous system. In this paper, we discuss the statistics of the true responses observed experimentally, and compare them with a theoretical model that describes how responses would be, if the encoding strategy were different. To differentiate these two situations, we employ the variable $\mathbf{R}_{\text{ex}}$ for the experimental responses (the real ones), and $\mathbf{R}_{\text{su}}$ for the surrogate responses (the fictitious ones). The associated conditional probability distributions are $P_{\text{ex}}(\mathbf{R}_{\text{ex}} = \mathbf{r}|S = s)$ and $P_{\text{su}}(\mathbf{R}_{\text{su}} = \mathbf{r}|S = s)$, which are often abbreviated as $P_{\text{ex}}(\mathbf{r}|s)$ and $P_{\text{su}}(\mathbf{r}|s)$, respectively. Once these distributions are known, and given the prior stimulus probabilities $P(s)$, the joint probabilities $P_{\text{ex}}(\mathbf{r}, s)$ and $P_{\text{su}}(\mathbf{r}, s)$ can be deduced, as well as the marginals $P_{\text{ex}}(\mathbf{r})$ and $P_{\text{su}}(\mathbf{r})$. When interpreting the abbreviated notation, readers should keep in mind that P_{ex} governs the variable $\mathbf{R}_{\text{ex}}$, and P_{su}, $\mathbf{R}_{\text{su}}$. If a statement is made about a distribution P or a response variable $\mathbf{R}$ that has no sub-index, the argument is intended for both the real and surrogate distributions or variables.

2.1.2. Encoding

The process of converting stimuli S into neural responses $\mathbf{R}$ (e.g., spike-trains, local-field potentials, electroencephalographic or other brain signals, etc.) is called "encoding" [9,15]. The encoding process is typically noisy, in the sense that repeated presentations of the same stimulus may yield different neural responses, and is characterised by the joint probability distribution $P(s, \mathbf{r})$. The associated marginal probabilities are

$$\begin{aligned} P(s) &= \sum_{\mathbf{r}} P(s, \mathbf{r}), \\ P(\mathbf{r}) &= \sum_{s} P(s, \mathbf{r}), \end{aligned}$$

from which the conditional response probability $P(\mathbf{r}|s) = P(s, \mathbf{r})/P(s)$, and the posterior stimulus probability $P(s|\mathbf{r}) = P(s, \mathbf{r})/P(\mathbf{r})$ can be defined.

The mutual information that $\mathbf{R}$ contains about S is

$$I(S; \mathbf{R}) = \sum_{s,\mathbf{r}} P(s, \mathbf{r}) \log_2 \frac{P(s|\mathbf{r})}{P(s)} . \tag{1}$$

More generally, the mutual information $I(S; X)$ about S contained in any random variable X, including but not limited to $\mathbf{R}$, can be computed using the above formula with $\mathbf{R}$ replaced by X. For compactness, we denote $I(S; X)$ as I_X unless ambiguity arises.

2.1.3. Data Processing Inequalities

When the response $\mathbf{R}_2$ is a post-processed version of the response $\mathbf{R}_1$, the joint probability distribution $P(s, \mathbf{r}_1, \mathbf{r}_2)$ can be written as $P(s, \mathbf{r}_1)\ P(\mathbf{r}_2|\mathbf{r}_1)$. This decomposition implies that $\mathbf{R}_2$ is conditionally independent of S. In these circumstances, the information about S contained in $\mathbf{R}_2$ cannot exceed the information about S contained in $\mathbf{R}_1$ [16]. In addition, the accuracy of the optimal decoder operating on $\mathbf{R}_2$ cannot exceed the accuracy of the optimal decoder operating on $\mathbf{R}_1$ [17]. These results constitute the data processing inequalities.

2.1.4. Decoding

The process of transforming responses $\mathbf{r}$ into estimated stimuli $\hat{s}$ is called "decoding" [9,15]. More precisely, a decoder is a mapping $\mathbf{r} \to \hat{s}$ defined by a function $\hat{s} = D(\mathbf{r})$. The inverse of this function is D^{-1}, and when D is not injective, D^{-1} is a multi-valued mapping. The joint probability $P(s, \hat{s})$ of the presented and estimated stimuli, also called "confusion matrix" [12], is

$$P(s, \hat{s}) = \sum_{\mathbf{r} \in D^{-1}(\hat{s})} P(s, \mathbf{r}) , \tag{2}$$

where the sum runs over all responses $\mathbf{r}$ that are mapped onto $\hat{s}$ by D. The information that $\hat{S}$ preserves about S is $I_{\hat{S}}$, and can be calculated from the confusion matrix of Equation (2). The decoding accuracy above chance level is here defined as

$$A = \sum_s P(S{=}s, \hat{S}{=}s) - \max_s P(s)\,. \tag{3}$$

2.1.5. Optimal Decoding

Although all mappings D are formally admissible as decoders, not all are useful. The aim of a decoder is to make a good guess of the external stimulus S from the neural response $\mathbf{R}$. It is therefore important to be able to construct decoders that make good guesses, or at least, as good as the mapping from stimuli to responses allows. Optimal decoders (also called *Bayesian* or *maximum-a-posteriori* decoders, as well as *ideal homunculus*, or *observer*, among other names) are defined as [18,19]

$$\hat{s} = D_{\text{opt}}(\mathbf{r}) = \arg\max_s P(s|\mathbf{r}) = \arg\max_s P(s, \mathbf{r})\,. \tag{4}$$

This mapping selects, for each response $\mathbf{r}$, the stimulus $\hat{s}$ that most likely generated $\mathbf{r}$. It is optimal in the sense that any other decoding algorithm yields a confusion matrix with lower decoding accuracy. Equation (4) depends on $P(s, \mathbf{r})$, so the decoder cannot be defined before knowing the functional shape of the joint probability distribution between stimuli and responses. The process of estimating $P(s, \mathbf{r})$ from real data, and the subsequent insertion of the obtained distribution in Equation (4) is called the *training* of the decoder. The word "training" makes reference to a gradual process, originally stemming from a computational strategy employed to estimate the distribution progressively, while the data was being gathered. However, in this paper we do not discuss estimation strategies from limited samples, so for us, "training a decoder" is equivalent to constructing a decoder from Equation (4).

2.1.6. Extensions of Optimal Decoding

The study of Ince et al. [20] introduced the concept of ranked decoding, in which each response $\mathbf{r}$ is mapped onto a list of K stimuli $\hat{\mathbf{s}} = (\hat{s}_1, \ldots, \hat{s}_K)$ ordered according to their posterior probabilities so that $P(\hat{s}_k|\mathbf{r}) \geq P(\hat{s}_{k+1}|\mathbf{r})$ (with $1 \leq k < K$, and $K \leq$ the total number of stimuli in the experiment). Ranked decoding can provide useful models for intermediate stages in the decision pathway, and the information loss induced by ranked decoding was computed recently [17]. The joint probability associated with ranked decoding is

$$P(s, \hat{\mathbf{s}}) = \sum_{\mathbf{r} \in D^{-1}(\hat{\mathbf{s}})} P(s, \mathbf{r})\,, \tag{5}$$

where the sum runs over all response vectors $\mathbf{r}$ that produce the same ranking $\hat{\mathbf{s}}$. Although $P(s, \hat{\mathbf{s}})$ can be used to compute the information $I_{\hat{\mathbf{S}}}$ between S and $\hat{\mathbf{S}}$, it cannot be used to compute the decoding accuracy above chance level because the support of $\hat{\mathbf{S}}$ (i.e., the set of stimulus lists) is not contained in the support of S (i.e., the set of stimuli).

2.1.7. Approximations to Optimal Decoding

For given probabilities $P(\mathbf{r}|s)$ and $P(s)$, Equation (4) defines a mapping between each response $\mathbf{r}$ and a candidate stimulus $\hat{s}$. In the study of the neural code, scientists often wonder what would happen if responses were not governed by the experimentally recorded distribution $P_{\text{ex}}(\mathbf{r}|s)$, but by some other surrogate distribution $P_{\text{su}}(\mathbf{r}|s)$. If we replace $P_{\text{ex}}(\mathbf{r}|s)$ by $P_{\text{su}}(\mathbf{r}|s)$ in Equation (4), we define a new decoding algorithm

$$\hat{s} = D_{\text{su}}(\mathbf{r}) = \arg\max_s P_{\text{su}}(s|\mathbf{r}) = \arg\max_s P_{\text{su}}(s, \mathbf{r})\,. \tag{6}$$

which, as discussed below, may or may not be optimal, depending on how the decoder is used.

2.1.8. Two Different Decoding Strategies

One alternative, here referred to as "decoding method α" is that, for each response $\mathbf{r}$ obtained experimentally, one decodifies a stimulus $\hat{s}$ using the new mapping of Equation (6). In this case, the chain $s \rightarrow \mathbf{r} \rightarrow \hat{s}$ gives rise to the confusion matrix

$$P^{\alpha}(s,\hat{s}) = \sum_{\mathbf{r} \in D_{\mathrm{su}}^{-1}(\hat{s})} P_{\mathrm{ex}}(s,\mathbf{r}), \tag{7}$$

where the sum runs over all response vectors $\mathbf{r}$ that are mapped onto $\hat{s}$ by the new decoding algorithm D_{su}, and the probability $P_{\mathrm{ex}}(\mathbf{r},s)$ appearing in the right-hand side is the real one, since responses $\mathbf{r}$ are generated experimentally. It is easy to see that in this case, the decoding accuracy of the new algorithm is suboptimal, since responses $\mathbf{r}$ are generated with the original distribution $P_{\mathrm{ex}}(\mathbf{r}|s)$, and for that distribution, the optimal decoder is given by Equation (4) with $P = P_{\mathrm{ex}}$. In the literature, training a decoder with a probability $P_{\mathrm{su}}(\mathbf{r}|s)$ and then operating it on variables that are generated with $P_{\mathrm{ex}}(\mathbf{r}|s)$ is called *mismatched decoding*. In what follows, information values calculated from the distribution of Equation (7) are noted as $I_{\hat{S}}^{\alpha}$.

A second alternative, "decoding method β," is that, for each stimulus s, a surrogate response $\mathbf{R}_{\mathrm{su}}$ is drawn using the new distribution $P_{\mathrm{su}}(\mathbf{r}|s)$. If the sampled value is $\mathbf{R}_{\mathrm{su}} = \mathbf{r}$, the stimulus $\hat{s} = D_{\mathrm{su}}(\mathbf{r})$ is decoded. In this case, the confusion matrix is

$$P^{\beta}(s,\hat{s}) = \sum_{\mathbf{r} \in D_{\mathrm{su}}^{-1}(\hat{s})} P_{\mathrm{su}}(s,\mathbf{r}), \tag{8}$$

where as before, the sum runs over all response vectors $\mathbf{r}$ that are mapped onto $\hat{s}$ by the decoding algorithm $D_{\mathrm{su}}(\mathbf{r})$, but now the probability $P_{\mathrm{su}}(\mathbf{r},s)$ appearing in the right-hand side is the surrogate one, since responses $\mathbf{R}_{\mathrm{su}}$ are not generated experimentally. In this case, there is no mismatch between the construction and operation of the decoder, and D_{su} is optimal, in the sense that no other algorithm decodes $\mathbf{R}_{\mathrm{su}}$ with higher decoding accuracy. One should bear in mind, however, that the surrogate responses are not the responses observed experimentally, that they may well take values in a response set that does not coincide with the set of real responses, and that $\mathbf{R}_{\mathrm{su}}$ is not necessarily obtained by transforming the real response $\mathbf{R}_{\mathrm{ex}}$ with a stimulus-independent mapping (see below). In what follows, information values calculated from the distribution of Equation (8) are noted as $I_{\hat{S}}^{\beta}$. Methods α and β can be easily extended to encompass also ranked decoding, *mutatis mutandis*.

The two alternative decoding methods yield two different decoding accuracies. To distinguish them, we use the notation $A_{\mathbf{R}_1}^{\mathbf{R}_2}$. The superscript indicates the variable whose probability distribution is used to construct the decoder in Equation (4), and consequently, determines the set of $\mathbf{r} \in D_{\mathrm{su}}^{-1}(\hat{s})$ that contribute to the sums of Equations (7) and (8). The subscript indicates the variable upon which the decoder is applied, and its probability distribution is summed in the right-hand side of Equations (7) and (8). That is, $A_{\mathbf{R}_1}^{\mathbf{R}_2}$ is computed through Equation (3) with

$$\mathrm{P}_{\mathbf{R}_1}^{\mathbf{R}_2}(s,\hat{s}) = \sum_{\mathbf{r} \in D_{\mathbf{R}_2}^{-1}(\hat{s})} P(S=s, \mathbf{R}_1 = \mathbf{r}), \tag{9}$$

so that $\mathrm{P}^{\alpha}(s,\hat{s}) = \mathrm{P}_{\mathbf{R}_{\mathrm{ex}}}^{\mathbf{R}_{\mathrm{su}}}(s,\hat{s})$ and $\mathrm{P}^{\beta}(s,\hat{s}) = \mathrm{P}_{\mathbf{R}_{\mathrm{su}}}^{\mathbf{R}_{\mathrm{su}}}(s,\hat{s})$.

2.2. The Applicability of the Data-Processing Inequality

Assessing the relevance of a response feature typically involves a subtraction $\Delta I = I - I'$, where I and I' represent the mutual information between stimuli and a set of response features containing or not containing the tested feature, respectively. The magnitude of ΔI is often interpreted as the information provided by the tested feature. This interpretation requires ΔI to be positive, since intuitively, one would imagine that removing a response feature cannot increase the encoded information.

As shown below, a formal proof of this intuition may or may not be possible invoking the data processing inequality (see Section 2.1.3 and reference [16]), depending on the method used to eliminate the tested feature. As a consequence, there are cases in which ΔI is indeed negative (see below). In these cases, the tested feature is detrimental to information encoding [9].

2.2.1. Reduced Representations

There are several procedures by which the tested feature can be removed from the response. The validity of the data-processing inequalities (see definition in Section 2.1.3) depends on the chosen procedure. In order to specify the conditions in which the inequalities hold, we here introduce the concept of *reduced representations*. When the response feature under evaluation is removed from $\mathbf{R}_{\text{ex}}$ by a deterministic mapping $\mathbf{R}_{\text{su}} = f(\mathbf{R}_{\text{ex}})$, we call the obtained variable $\mathbf{R}_{\text{su}}$ a *reduced representation* of $\mathbf{R}_{\text{ex}}$. A required condition for a mapping to be a reduced representation is that the function f be stimulus-independent, that is, that the value of $\mathbf{R}_{\text{su}}$ be conditionally independent from s. Mathematically, this means that $P(\mathbf{r}_{\text{su}}, s|\mathbf{r}_{\text{ex}}) = P(\mathbf{r}_{\text{su}}|\mathbf{r}_{\text{ex}})\ P(s|\mathbf{r}_{\text{ex}})$. If the mapping f and the conditional response distribution $P_{\text{ex}}(\mathbf{r}|s)$ are known, the distribution $P_{\text{su}}(\mathbf{r}|s)$ can be derived using standard methods. The data processing inequality ensures that for all reduced representations, $I_{\mathbf{R}_{\text{ex}}} \geq I_{\mathbf{R}_{\text{su}}}$.

Reduced representations are usually employed when the response feature whose relevance is to be assessed takes a definite value in each trial, as happens for example, with the number of spikes in a fixed time window, the latency of the firing response, or the activity of a specific neuron in a larger population of neurons. In these cases it is easy to construct $\mathbf{R}_{\text{su}}$ simply by dropping from $\mathbf{R}_{\text{ex}}$ the tested feature, or by fixing its value with some deterministic rule.

Reduced representations can also be used in other cases, for example, when the relevance of the feature *response accuracy* is assessed. This feature does not take a specific value in each trial; only by comparing multiple trials can the response accuracy be determined. A widely-used strategy is to represent spike trains with temporal bins of increasing duration, and to evaluate how the amount of information decreases as the representation becomes coarser. A sequence of surrogate responses is thereby defined, by progressively disregarding the fine temporal precision with which spike trains were recorded (Figure 1).

Several studies have reported an information $I_{\mathbf{R}_{\text{su}}}$ that decreases monotonically with the duration δt of the time bin (for example [21–23]). If there is a specific temporal scale in which spike-time precision is relevant—the alleged argument goes—a sudden drop in $I_{\mathbf{R}_{\text{su}}}(\delta t)$ appears at the relevant scale. It should be noted, however, that the data processing inequality does not ensure that $I_{\mathbf{R}_{\text{su}}}(\delta t)$ be a monotonically decreasing function of δt. In the example of Figure 1, representations $\mathbf{R}^1_{\text{su}}$ and $\mathbf{R}^2_{\text{su}}$ are defined with long temporal bins, the durations of which are integer multiples of the bin used for $\mathbf{R}_{\text{ex}}$. Hence, $\mathbf{R}^1_{\text{su}}$ and $\mathbf{R}^2_{\text{su}}$ are reduced representations of $\mathbf{R}_{\text{ex}}$, and the data processing inequality does indeed guarantee that $I_{\mathbf{R}_{\text{ex}}} \geq I_{\mathbf{R}^1_{\text{su}}}$ and $I_{\mathbf{R}_{\text{ex}}} \geq I_{\mathbf{R}^2_{\text{su}}}$. However, $\mathbf{R}^2_{\text{su}}$ is not a reduced representation of $\mathbf{R}^1_{\text{su}}$, so there is no reason why $I_{\mathbf{R}^2_{\text{su}}}$ should be smaller than $I_{\mathbf{R}^1_{\text{su}}}$, and indeed, Figure 1b shows an example where it is not. The representation constructed with bins of intermediate duration, namely 10 ms, does not distinguish between the two stimuli, whereas those of shorter and longer duration, 5 and 15 ms, do. A similar effect can be observed in the experimental data (freely available online) of Lefebvre et al. [24], when analysed with bins of sizes 5, 10 and 15 ms in windows of total duration 60 ms. Although these examples are rare, they demonstrate that there is no theoretical substantiation to the expectation of $I_{\mathbf{R}_{\text{su}}}$ to drop monotonically with increasing δt.

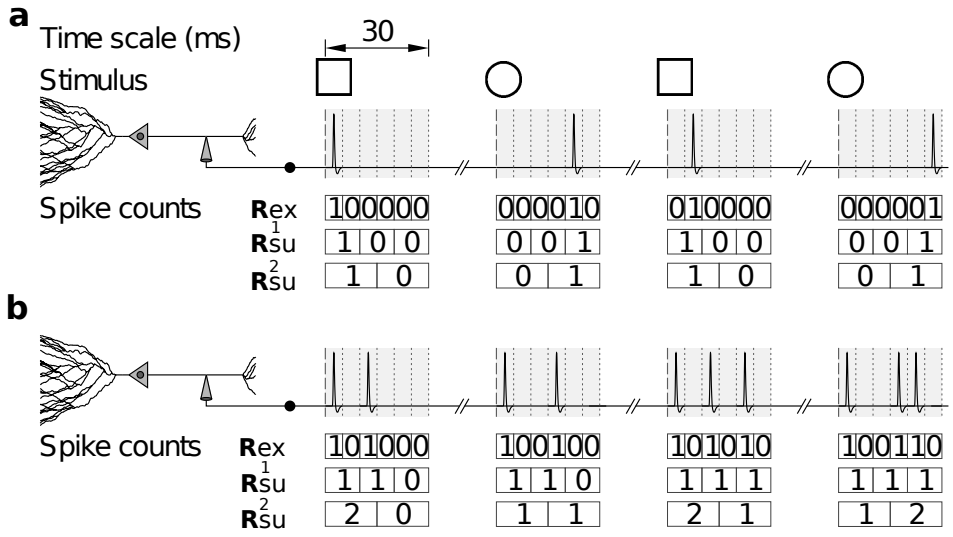

Figure 1. Assessing the relevance of response accuracy by varying the duration of the temporal bin. (**a**) Hypothetical intracellular recording of the spike patterns elicited by a single neuron after presenting in alternation two visual stimuli, □ and ○, each of which triggers two possible responses displayed in columns 1 and 3 for □, and 2 and 4 for ○. Stimulus probabilities and conditional response probabilities are arbitrary. Time is discretized in bins of 5 ms. The responses are recorded within 30 ms time-windows after stimulus onset. Spikes are fired with latencies that are uniformly distributed between 0 and 10 ms after the onset of □, and between 20 and 30 ms after the onset of ○. Responses are represented by counting the number of spikes within consecutive time-bins of size 5, 10 and 15 ms starting from stimulus onset, thereby yielding discrete-time sequences $\mathbf{R}_{\text{ex}}$, $\mathbf{R}^1_{\text{su}}$ and $\mathbf{R}^2_{\text{su}}$, respectively; (**b**) Same as a, but with stimuli producing two different types of response patterns composed of 2 or 3 spikes.

2.2.2. Stochastically Reduced Representations

When the response feature under evaluation is removed from the response variable $\mathbf{R}_{\text{ex}}$ by a stochastic mapping $\mathbf{R}_{\text{ex}} \rightarrow \mathbf{R}_{\text{su}}$, the obtained variable $\mathbf{R}_{\text{su}}$ is called a *stochastically reduced representation* of $\mathbf{R}_{\text{ex}}$. A required condition for a mapping to be a stochastically reduced representation is that the probability distribution of each $\mathbf{R}_{\text{su}}$ be dependent on $\mathbf{R}_{\text{ex}}$, but conditionally independent from s. In these circumstances, the data processing inequality ensures that $I_{\mathbf{R}_{\text{ex}}} \geq I_{\mathbf{R}_{\text{su}}}$. If the statistical properties of the noisy components of the mapping are known, as well as the conditional response probability distribution $P_{\text{ex}}(\mathbf{r}|s)$, the distribution $P_{\text{su}}(\mathbf{r}|s)$ can be derived using standard methods. Formally, stochastic representations $\mathbf{R}_{\text{su}}$ are obtained through stimulus-independent stochastic functions of the original representation $\mathbf{R}_{\text{ex}}$. After observing that $\mathbf{R}_{\text{ex}}$ adopted the value $\mathbf{r}_{\text{ex}}$, these functions produce a single value $\mathbf{r}_{\text{su}}$ for $\mathbf{R}_{\text{su}}$ chosen with transition probabilities $Q(\mathbf{r}_{\text{su}}|\mathbf{r}_{\text{ex}})$ such that

$$P_{\text{su}}(\mathbf{r}_{\text{su}}|s) = \sum_{\mathbf{r}_{\text{ex}}} P_{\text{ex}}(\mathbf{r}_{\text{ex}}|s)\, Q(\mathbf{r}_{\text{su}}|\mathbf{r}_{\text{ex}})\,. \tag{10}$$

To illustrate the utility of stochastically reduced representations, we discuss their role in providing alternative strategies when assessing the relevance of spike-timing precision, not by changing the bin size as in Figure 1, but by randomly manipulating the responses, as illustrated in Figure 2.

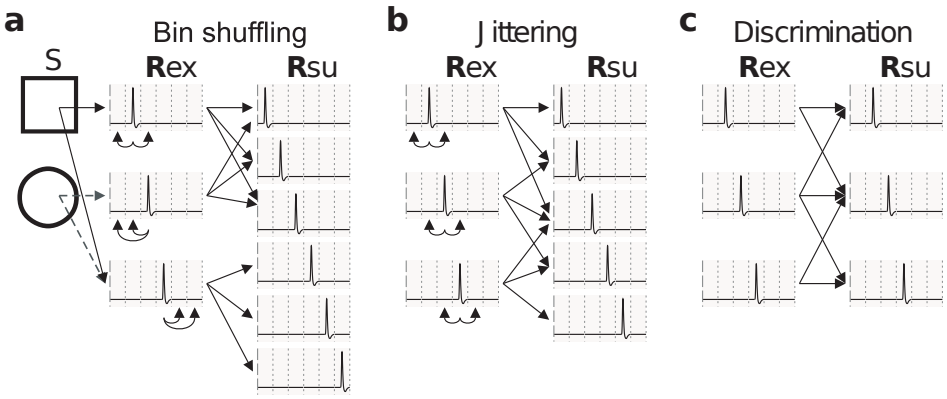

Figure 2. Examples of stochastic codes. Alternative ways of assessing the relevance of spike-timing precision. (**a**) Stochastic function (arrows on the left) modeling the encoding process. The elicited response $\mathbf{r}_{\text{ex}}$ is turned into a surrogate response $\mathbf{r}_{\text{su}}$ with a transition probability $Q(\mathbf{r}_{\text{su}}|\mathbf{r}_{\text{ex}})$ given by Equation (11). This function turns $\mathbf{R}_{\text{ex}}$ into a stochastic representation $\mathbf{R}_{\text{su}}$ by shuffling spikes and silences within bins of 15 *ms* starting from stimulus onset; (**b**) Responses $\mathbf{r}_{\text{ex}}$ in panel (a) are transformed by a stochastic function with $Q(\mathbf{r}_{\text{su}}|\mathbf{r}_{\text{ex}})$ given by Equation (12), which introduces jitter uniformly distributed within 15 *ms* windows centered at each spike; (**c**) Responses $\mathbf{r}_{\text{ex}}$ in panel (a) are transformed by a stochastic function with $Q(\mathbf{r}_{\text{su}}|\mathbf{r}_{\text{ex}})$ given by Equation (13), which models the inability to distinguish responses with spikes occurring in adjacent bins, or equivalently, with distances $d^{\text{spike}}[q=1] \leq 1$ or $d^{\text{interval}}[q=1] \leq 1$ (see [25,26] for further remarks on these distances). Notice that $\mathbf{R}_{\text{su}}$ samples the same response set as $\mathbf{R}_{\text{ex}}$.

The method of Figure 2a yields the same information $I_{\mathbf{R}_{\text{su}}}$ and response accuracy as the method producing $\mathbf{R}^2_{\text{su}}$ in Figure 1. Each method yields responses that can be related to the responses of the other method through a stimulus-independent deterministic or stochastic function. Both methods suffer from the same drawback: They treat spikes differently depending on their location within the 15 ms time window. Indeed, both methods preserve the distinction between two spikes located in different windows, but not within the same window, even if the separation between the spikes is the same. The mapping illustrated in Figure 2a has transition probabilities

$$Q(\mathbf{r}_{\text{su}}|\mathbf{r}_{\text{ex}}) = \frac{1}{3}\begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix}, \tag{11}$$

where rows enumerate the elements of the ordered set $\mathcal{R}_{ex}=\{[2],[3],[4]\}$ from where $\mathbf{R}_{\text{ex}}$ is sampled, and columns enumerate the elements of the ordered set $\mathcal{R}_{su}=\{[1],[2],[3],[4],[5],[6]\}$ from where $\mathbf{R}_{\text{su}}$ is sampled.

A third method, jittering, consists in shuffling the recorded spikes within time windows centered at each spike (Figure 2b). The responses generated by this method need not be obtainable from the responses generated by the mappings of Figure 2a or Figure 1 through stimulus-independent stochastic functions. Still, the method of Figure 2b inherently yields a stochastic code, and, unlike the methods discussed previously, treats all spikes in the same manner. The mapping illustrated in Figure 2b has transition probabilities

$$Q(\mathbf{r}_{\text{su}}|\mathbf{r}_{\text{ex}}) = \frac{1}{3}\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}, \tag{12}$$

where rows enumerate the elements of the ordered set $\mathcal{R}_{ex}=\{[2],[3],[4]\}$ from where $\mathbf{R}_{\text{ex}}$ is sampled, and columns enumerate the elements of the ordered set $\mathcal{R}_{su}=\{[1],[2],[3],[4],[5]\}$ from where $\mathbf{R}_{\text{su}}$ is sampled.

As a fourth example, consider the effect of response discrimination, as studied in the seminal work of Victor and Purpura [25]. There, two responses were considered indistinguishable when

some measure of distance between the responses was less than a predefined threshold. However, neural responses were transformed through a method based on cross-validation that is not guaranteed to be stimulus-independent. Depending on the case, hence, this fourth method may or may not be a stochastically reduced representation. The case chosen in Figure 2c is a successful example, and the associated matrix of transition probabilities is

$$Q(\mathbf{r}_{\mathrm{su}}|\mathbf{r}_{\mathrm{ex}}) = \frac{1}{6}\begin{bmatrix} 3 & 3 & 0 \\ 2 & 2 & 2 \\ 0 & 3 & 3 \end{bmatrix}, \tag{13}$$

where rows and columns enumerate the elements of the ordered set $\mathcal{R}_{ex}=\mathcal{R}_{su}=\{[2],[3],[4]\}$ from where both $\mathbf{R}_{\mathrm{ex}}$ and $\mathbf{R}_{\mathrm{su}}$ are sampled.

Other methods exist which merge indistinguishable responses, thereby yielding reduced representations. These methods, however, are limited to notions of similarity that are transitive, a condition not fulfilled, for example, by those based on Euclidean distance, edit distance, or by the case of Figure 2c.

Stochastically reduced representations include reduced representations as limiting cases. Indeed, when for each $\mathbf{r}_{\mathrm{ex}}$ there is a $\mathbf{r}_{\mathrm{su}}$ such that $Q(\mathbf{r}_{\mathrm{su}}|\mathbf{r}_{\mathrm{ex}}) = 1$, stochastic representations become reduced representations (Figure 3). The possibility to include stochasticity, however, broadens the range of alternatives. Consider for example the hypothetical experiment in Figure 3a, in which the neural responses $\mathbf{R}_{\mathrm{ex}}=[L, C]$ can be completely characterized by the first-spike latencies (L) and the spike counts (C). The importance of C can be studied for example by using a reduced code that replaces all C-values with a constant (Figure 3b). In this case,

$$Q(\mathbf{r}_{\mathrm{su}}|\mathbf{r}_{\mathrm{ex}}) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{14}$$

where rows enumerate the elements of the ordered set $\mathcal{R}_{ex}=\{[2,1],[3,1],[3,2],[4,2]\}$ from where $\mathbf{R}_{\mathrm{ex}}$ is sampled, and columns enumerate the elements of the ordered set $\mathcal{R}_{su}=\{[2,1],[3,1],[4,1]\}$ from where $\mathbf{R}_{\mathrm{su}}$ is sampled.

Another alternative is to assess the relevance of C by means of a stochastic code that shuffles the values of C across all responses with the same L (Figure 3c). In this case,

$$Q(\mathbf{r}_{\mathrm{su}}|\mathbf{r}_{\mathrm{ex}}) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & a & \bar{a} & 0 \\ 0 & a & \bar{a} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{15}$$

where rows enumerate the elements of the ordered set $\mathcal{R}_{ex}=\{[2,1],[3,1],[3,2],[4,2]\}$ from where $\mathbf{R}_{\mathrm{ex}}$ is sampled, and columns enumerate the elements of the ordered set $\mathcal{R}_{su}=\{[2,1],[3,1],[3,2],[4,2]\}$ from where $\mathbf{R}_{\mathrm{su}}$ is sampled. The parameter a is arbitrary, as long as $0 < a < 1$. We use the notation $\bar{a} = 1 - a$.

A third option is to use a stochastic code that preserves the original value of L but chooses the value of C from some possibly $L-$dependent probability distribution (Figure 3d), for which

$$Q(\mathbf{r}_{\mathrm{su}}|\mathbf{r}_{\mathrm{ex}}) = \begin{bmatrix} b & 0 & 0 & \bar{b} & 0 & 0 \\ 0 & c & 0 & 0 & \bar{c} & 0 \\ 0 & c & 0 & 0 & \bar{c} & 0 \\ 0 & 0 & d & 0 & 0 & \bar{d} \end{bmatrix} \tag{16}$$

where rows enumerate the elements of the ordered set $\mathcal{R}_{ex}=\{[2,1],[3,1],[3,2],[4,2]\}$ from where $\mathbf{R}_{\mathrm{ex}}$ is sampled, and columns enumerate the elements of the ordered set $\mathcal{R}_{su}=\{[2,1],[3,1],[4,1],[2,2],[3,2],[4,2]\}$ from where $\mathbf{R}_{\mathrm{su}}$ is sampled. The parameters a, b, c and d are arbitrary, as long as $0<a, b, c, d<1$; and we have used the notation $\bar{x}=1-x$ for any number x.

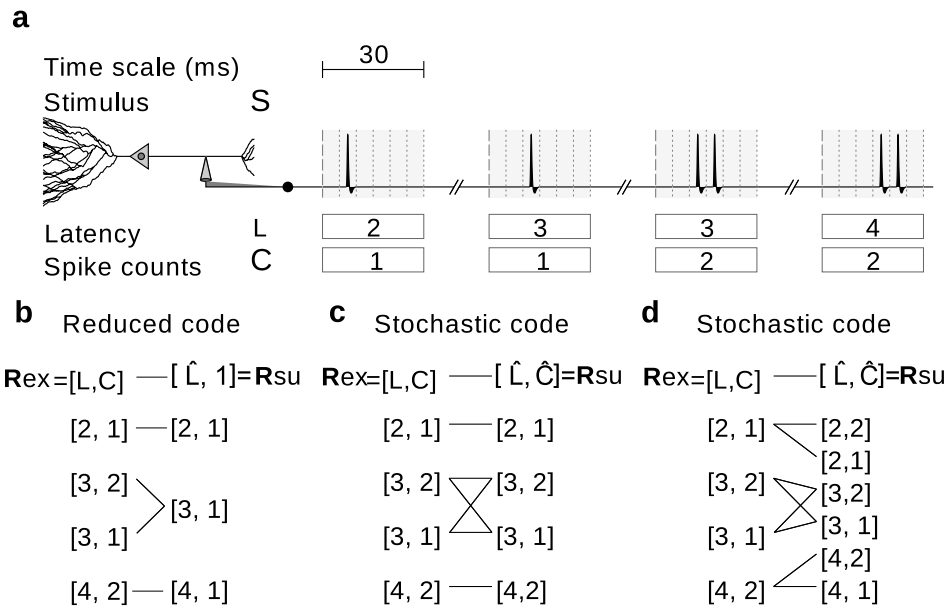

Figure 3. Stochastically reduced representations include and generalize deterministically reduced representations. (**a**) Analogous description to Figure 1a, but with responses characterized using a representation $\mathbf{R}_{ex} = [L, C]$ based on the first-spike latency (L) and the spike-count (C); (**b**) Deterministic transformation (arrows) of $\mathbf{R}_{ex}$ in panel a into a reduced code $\mathbf{R}_{su} = [\hat{L}, 1]$, which ignores the additional information carried in C by considering it constant and equal to unity. This reduced code can also be reinterpreted as a stochastic code with transition probabilities $Q(\mathbf{r}_{su}|\mathbf{r}_{ex})$ defined by Equation (14); (**c**) The additional information carried in C is here ignored by shuffling the values of C across all trails with the same L, thereby turning $\mathbf{R}_{ex}$ in panel a into a stochastic code $\mathbf{R}_{su}=[\hat{L}, \hat{C}]$ with transition probabilities $Q(\mathbf{r}_{su}|\mathbf{r}_{ex})$ defined by Equation (15); (**d**) The additional information carried in C is here ignored by replacing the actual value of C for one chosen with some possibly L-dependent probability distribution (Equation (16)).

2.2.3. Modification of the Conditional Response Probability Distribution

When the response feature under evaluation is removed by altering the real conditional response probability distribution $P_{ex}(\mathbf{r}|s)$, and transforming it into a surrogate distribution $P_{su}(\mathbf{r}|s)$, the obtained response model is here said to implement a *probabilistic removal* of the tested feature. Probabilistic removals are usually employed when assessing the relevance of correlations between neurons in a population, since correlations are not a variable that can be deleted from each individual response. For example, if $\mathbf{R}=(R_1, \ldots, R_n)$ represents the spike count of n different neurons, the real distribution $P_{ex}(r_1, \ldots, r_n|s)$ is replaced by a new distribution $P_{su}(r_1, \ldots, r_n|s)$ in which all neurons are conditionally independent, that is,

$$P_{su}(\mathbf{r}|s) = P_{NI}(\mathbf{r}|s) = \prod_{i=1}^{n} P_{ex}(r_i|s), \tag{17}$$

where, following the notation introduced previously [17], the generic subscript "su" was replaced by "NI" to indicate "noise-independent".

The probabilistic removal of a response feature may or may not be describable in terms of a deterministically or a stochastically reduced representation. In other words, there may or may not exist a mapping $\mathbf{R}_{ex} \rightarrow \mathbf{R}_{su}$, or equivalently, a matrix of transition probabilities $Q(\mathbf{r}_{su}|\mathbf{r}_{ex})$, that captures the replacement of $P_{ex}(\mathbf{r}|s)$ by $P_{su}(\mathbf{r}|s)$. It is important to assess whether such a matrix exists, since the data processing inequality is only guaranteed to hold with reduced representations, stochastic or not. If no reduced representation can capture the effect of a probabilistic removal, the data processing inequality may not hold, and $I_{\mathbf{R}_{su}}$ may well be larger than $I_{\mathbf{R}_{ex}}$.

In order to determine whether a stochastically reduced representation exists, the first step is to discern whether Equation (10) constitutes a compatible or an incompatible linear system for the matrix elements $Q(\mathbf{r}_{su}|\mathbf{r}_{ex})$. If the system is incompatible, there is no solution. In the compatible case, which is often indeterminate, a solution entirely composed of non-negative numbers that sum up to unity in each row is required. Given enough time and computational power, the problem can always be solved in the framework of linear programming [27]. In practical cases, however, the search is often hampered by the curse of dimensionality. To facilitate the labour, here we list a few necessary (though not sufficient) conditions that must be fulfilled for the mapping to exist. If any of the following properties does not hold, Equation (10) has no solution, so there is no need to begin a search.

Property 1. *Let $\mu(s)$ be a probability distribution defined in the set of stimuli that may or may not be equal to the actual distribution with which stimuli appear in the experiment under study. For any stimulus s, the inequality $I_\mu(\mathbf{R}_{su}; S=s) \leq I_\mu(\mathbf{R}_{ex}; S=s)$ between stimulus-specific informations [28,29] must hold, where*

$$I_\mu(\mathbf{R}; S=s) = \sum_{\mathbf{r}} P(\mathbf{r}|s)\ \log_2 \frac{P(\mathbf{r}|s)}{\sum_{s'} P(\mathbf{r}|s')\ \mu(s')}. \tag{18}$$

Proof. If $Q(\mathbf{r}_{su}|\mathbf{r}_{ex})$ exists, then Equation (10) can be inserted in Equation (18). Using the log-sum inequality [16], Property 1 follows. □

If we multiply both sides of the inequality by $\mu(s')$ and sum over s', we obtain an inequality between the mutual informations $I_\mu(\mathbf{R}_{su}; S) \leq I_\mu(\mathbf{R}_{ex}; S)$. If $\mu(s) = P(s)$, this result reduces to the data-processing inequality $I_{\mathbf{R}_{su}} \leq I_{\mathbf{R}_{ex}}$.

Property 2. *If $Q(\mathbf{r}_{su}|\mathbf{r}_{ex})$ exists, then $Q(\mathbf{r}_{su}|\mathbf{r}_{ex}) = 0$ whenever $P_{ex}(s, \mathbf{r}_{ex}) > 0$ and $P_{su}(s, \mathbf{r}_{su}) = 0$ for at least some s.*

Proof. Suppose that $Q(\mathbf{r}_{su}|\mathbf{r}_{ex}) > 0$ when $P_{ex}(s, \mathbf{r}_{ex}) > 0$ for some s. Then, Equation (10) yields $P_{su}(\mathbf{r}_{su}|s) > 0$, contradicting the hypothesis that $P_{su}(\mathbf{r}_{su}|s) = 0$. Hence, $Q(\mathbf{r}_{su}|\mathbf{r}_{ex})$ must vanish. □

For example, in Figure 4a, we decorrelate first-spike latencies (L) and spike counts (C) by replacing the true conditional distribution $P_{ex}(\mathbf{r}|s)$ (left panel) by its noise-independent version $P_{su} = P_{NI}(\mathbf{r}|s)$ defined in Equation (17) (middle panel). Before searching for a mapping $\mathbf{R}_{ex} \to \mathbf{R}_{su}$, we verify that the condition $I_{\mathbf{R}_{ex}} > I_{\mathbf{R}_{su}}$ holds. Moreover, for several choices of $\mu(\bigcirc)$ and $\mu(\square)$, one may confirm that $I_\mu(\mathbf{R}_{ex}; S=\bigcirc) > I_\mu(\mathbf{R}_{su}; S=\bigcirc)$, as well as $I_\mu(\mathbf{R}_{ex}; S=\square) > I_\mu(\mathbf{R}_{su}; S=\square)$. These results motivate the search for a solution of Equation (10) for $Q(\mathbf{r}_{su}|\mathbf{r}_{ex})$. The transition probability must be zero at least whenever $\mathbf{R}_{su} \in \{[1,3]; [2,3]; [3,3]; [3,2]; [3,1]\}$ and $\mathbf{R}_{ex} \in \{[1,2]; [2,1]\}$ (Property 2). One possible solution is

$$Q(\mathbf{r}_{su}|\mathbf{r}_{ex}) = \frac{1}{2} \begin{bmatrix} 2b & \bar{b}c & \bar{c}\bar{b} & \bar{b}c & 0 & 0 & \bar{c}\bar{b} & 0 & 0 \\ \bar{a} & 2a & 0 & 0 & \bar{a} & 0 & 0 & 0 & 0 \\ a & 0 & 0 & 2\bar{a} & a & 0 & 0 & 0 & 0 \\ 0 & b & 0 & b & 2\bar{b}c & \bar{c}\bar{b} & 0 & \bar{c}\bar{b} & 0 \\ 0 & 0 & b & 0 & 0 & \bar{b}c & b & \bar{b}c & 2\bar{c}\bar{b} \end{bmatrix}. \tag{19}$$

where each response is defined by a vector $[L, C]$, and rows and columns enumerate the elements of the ordered sets $\mathcal{R}_{ex} = \{[1,1], [1,2], [2,1], [2,2], [3,3]\}$ and $\mathcal{R}_{su} = \{[1,1], [1,2], [1,3], [2,1], [2,2], [2,3], [3,1], [3,2], [3,3]\}$ from where $\mathbf{R}_{ex}$ and $\mathbf{R}_{su}$ are sampled, respectively. In Equation (19), $a = P_{ex}([1,2]|\square)$; $b = P_{ex}([1,1]|\bigcirc)$; and $c = P_{ex}([2,2]|\bigcirc)/\bar{b}$.

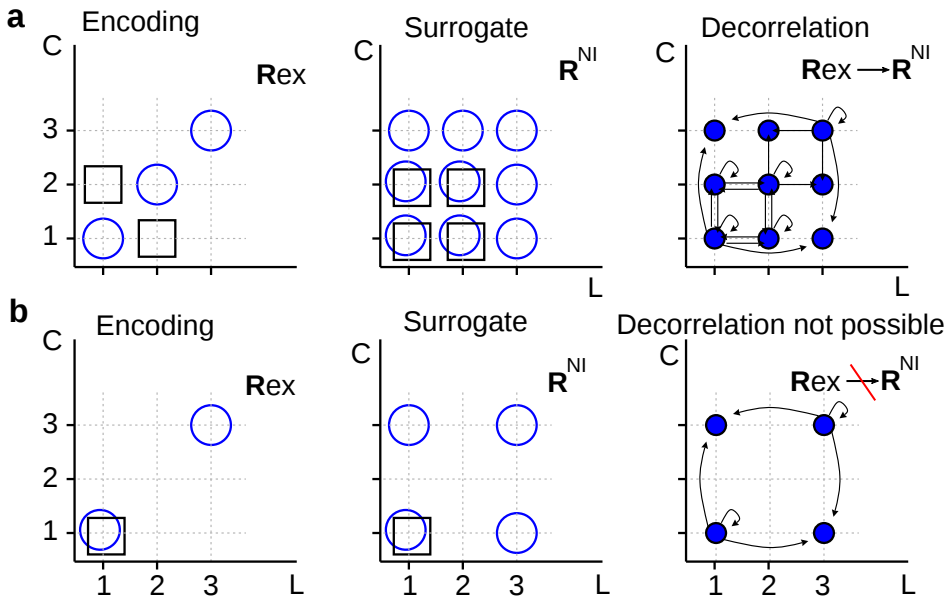

Figure 4. Relation between probabilistic removal and stochastic codes. (**a**) Cartesian coordinates depicting: on the left, responses $\mathbf{R}_{ex}$ of a neuron for which L and C are positively correlated when elicited by ○, and negatively correlated when elicited by □; in the middle, the surrogate responses $\mathbf{R}_{su} = \mathbf{R}_{NI}$ that would occur should L and C be noise independent (middle); and on the right, a stimulus-independent stochastic function that turns $\mathbf{R}_{ex}$ into $\mathbf{R}_{su}$ with $Q(\mathbf{r}_{su}|\mathbf{r}_{ex})$ given by Equation (19); (**b**) Same description as in (**a**), but with L and C noise independent given □, and with the stochastic function depicted on the right turning $\mathbf{R}_{ex}$ into $\mathbf{R}_{NI}$ given ○ but not □.

However, stochastically reduced representations are not always guaranteed to exist. For example, in Figure 4b, it is easy to verify that the condition $I_\mu(\mathbf{R}_{ex}; S = \square) < I_\mu(\mathbf{R}_{su}; S = \square)$ holds for any $\mu(\bigcirc) \neq 0$. Therefore, no stochastic mapping can transform $\mathbf{R}_{ex}$ into $\mathbf{R}_{su}$ in such a way that $P_{ex}(\mathbf{r}|s)$ is converted into $P_{su}(\mathbf{r}|s)$. Schneidman et al. [9] employed an analogous example, but involving different neurons instead of response aspects. The two examples of Figure 4 motivate the following theorem:

Theorem 1. *No deterministic mapping* $\mathbf{R}_{ex} \rightarrow \mathbf{R}_{su}$ *exists transforming the conditional probability* $P_{ex}(\mathbf{r}|s)$ *into its noise-independent version* $P_{su} = P_{NI}(\mathbf{r}|s)$ *defined in Equation (17). Stochastic mappings* $\mathbf{R}_{ex} \rightarrow \mathbf{R}_{su}$ *may or may not exist, depending on the conditional probability* $P_{ex}(\mathbf{r}|s)$.

Proof. See Appendix B.2. □

In addition, when a stochastic mapping $\mathbf{R}_{ex} \rightarrow \mathbf{R}_{su}$ exists, the values of the probabilities $Q(\mathbf{r}_{su}|\mathbf{r}_{ex})$ may well depend on the discarded response aspect, as well as on the preserved response aspects. We mention this fact, because when assessing the relevance of noise correlations, the marginals $P_{ex}(r_i|s)$ suffice for us to write down the surrogate distribution $P_{su}(\mathbf{r}|s) = P_{NI}(\mathbf{r}|s)$, with no need to know the full distribution $P_{ex}(\mathbf{r}|s)$ containing the noise correlations. One could have hoped that perhaps also the mapping $\mathbf{R}_{ex} \rightarrow \mathbf{R}_{su}$ (assuming that such a mapping exists) could be calculated with no knowledge of the noise correlations. This is, however, not always true, as stated in the theorem below. Two experiments with the same marginals and different amounts of noise correlations may require different mappings to eliminate noise correlations, as illustrated in the the example of Figure 5. More formally:

Theorem 2. *The transition probabilities $Q(\mathbf{r}_{su}|\mathbf{r}_{ex})$ of stochastic codes that ignore noise correlations may depend both on the marginal likelihoods (preserved at the output of the mapping), and on the noise correlations (eliminated at the output of the mapping).*

Proof. See Appendix B.3. □

The solution of Equation (10) for the example of Figure 5 is

$$Q(\mathbf{r}_{su}|\mathbf{r}_{ex}) = \frac{1}{2}\begin{bmatrix} \bar{a} & 2a & 0 & \bar{a} & 0 & 0 & 0 \\ a & 0 & 2\bar{a} & a & 0 & 0 & 0 \\ 0 & 0 & 0 & b & 2\bar{b} & 0 & b \\ 0 & 0 & 0 & \bar{b} & 0 & 2b & \bar{b} \end{bmatrix}, \tag{20}$$

where each response is defined by a vector $[L, C]$, and rows and columns enumerate the elements of the ordered sets $\mathcal{R}_{ex} = \{[1,2],[2,1],[2,3],[3,2]\}$ and $\mathcal{R}_{su} = \{[1,1],[1,2],[2,1],[2,2],[2,3],[3,2],[3,3]\}$ from where $\mathbf{R}_{ex}$ and $\mathbf{R}_{su}$ are sampled, respectively. In Equation (20), $a = P(\mathbf{R}_{ex} = [1,2]|S = \square)$; and $b = P(\mathbf{R}_{ex} = [3,2]|S = \bigcirc)$. The fact that the matrix in Equation (20) bears an explicit dependence on these parameters–and not only on $P_{ex}(L|S)$ and $P_{ex}(C|S)$–implies that the transformation between $\mathbf{R}_{ex}$ and $\mathbf{R}_{su}$ depends on the amount of noise correlations in $\mathbf{R}_{ex}$.

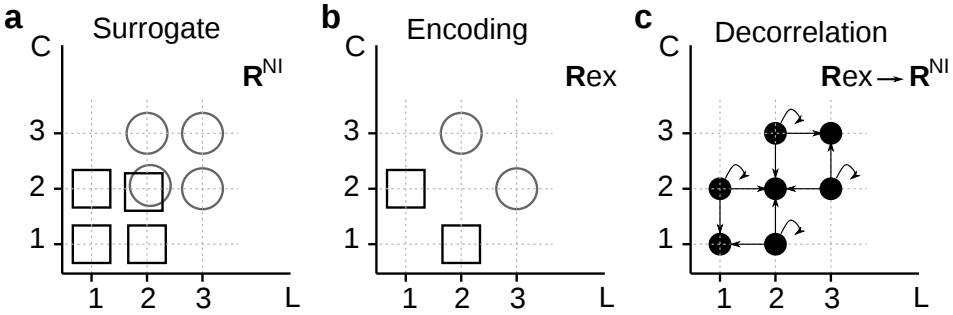

Figure 5. Stochastically reduced representations that ignore noise correlations may depend on them.(**a**) Cartesian coordinates representing a hypothetical experiment in which two different stimuli, □ and ○, elicit single neuron responses ($\mathbf{R}_{su} = \mathbf{R}_{NI}$) that are completely characterized by their first-spike latency (L) and spike counts (C). Both L and C are noise independent; (**b**) Cartesian coordinates representing a hypothetical experiment with the same marginal probabilities $P_{ex}(l|s)$ and $P_{ex}(c|s)$ as in panel (a), with one among many possible types of noise correlations between L and C; (**c**) Stimulus-independent stochastic function transforming the noise-correlated responses $\mathbf{R}_{ex}$ of panel (b) into the noise-independent responses $\mathbf{R}_{su} = \mathbf{R}_{NI}$ of panel (a). The transition probabilities $Q(\mathbf{r}_{su}|\mathbf{r}_{ex})$ are given in Equation 20, and they bear an explicit dependence on the amount of noise correlations.

2.3. Multiple Measures to Assess the Relevance of a Specific Response Feature

The importance of a specific response feature has been previously quantified in many ways (see [17,30] and references therein), which have oftentimes led to heated debates about their merits and drawbacks [9,11,12,17,31–33]. Here we consider several measures, to underscore the diversity of the meanings with which the relevance of a given feature has been assessed so far. They are mathematically defined as

$$\Delta I_{\mathbf{R}_{su}} = I_{\mathbf{R}_{ex}} - I_{\mathbf{R}_{su}} \tag{21}$$

$$\Delta I_{\hat{\mathbf{S}}} = I_{\mathbf{R}_{ex}} - I^{\beta}_{\hat{\mathbf{S}}} \tag{22}$$

$$\Delta I_{\hat{S}} = I_{\mathbf{R}_{ex}} - I^{\beta}_{\hat{S}} \tag{23}$$

$$\Delta A_{\mathbf{R}_{su}} = A^{\mathbf{R}_{ex}}_{\mathbf{R}_{ex}} - A^{\mathbf{R}_{su}}_{\mathbf{R}_{su}} \tag{24}$$

$$\Delta I^{D} = \sum_{s,\mathbf{r}} P_{ex}(s,\mathbf{r}) \ln \frac{P_{ex}(s|\mathbf{r})}{P_{su}(s|\mathbf{r})} \tag{25}$$

$$\Delta I^{DL} = \min_{\theta} \sum_{s,\mathbf{r}} P_{ex}(s,\mathbf{r}) \ln \frac{P_{ex}(s|\mathbf{r})}{P_{su}(s|\mathbf{r},\theta)} \tag{26}$$

$$\Delta I^{LS} = I_{\mathbf{R}_{ex}} - I^{\alpha}_{\hat{\mathbf{S}}} \tag{27}$$

$$\Delta I^{B} = I_{\mathbf{R}_{ex}} - I^{\alpha}_{\hat{S}} \tag{28}$$

$$\Delta A^{B} = A^{\mathbf{R}_{ex}}_{\mathbf{R}_{ex}} - A^{\mathbf{R}_{su}}_{\mathbf{R}_{ex}} \tag{29}$$

Equations (22)–(24) are based on matched decoders, that is, decoders operating on responses governed by the same probability distribution involved in their construction (method β). Instead, Equations (25)–(28) are based on the operation of mismatched decoders (method α). Each measure of Equations (21)–(24) has one or two homologous measures in Equations (25)–(29), as illustrated in Figure 6.

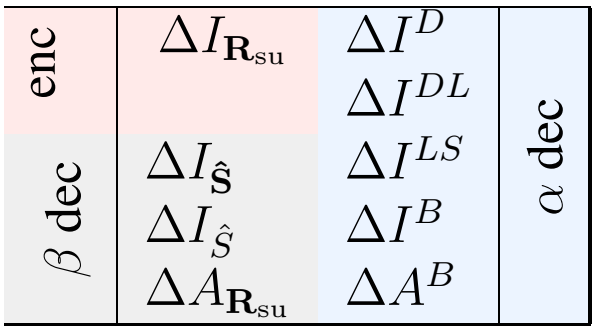

Figure 6. Relations between the measures defined in Equations (21)–(29). The four measures on the left are either encoding-oriented ($\Delta I_{\mathbf{R}_{su}}$, on a pink background), or half-way between encoding- and decoding-oriented (the last three, gray background). The five measures on the right are all decoding-oriented (light-blue background). Each measure on the left has a conceptually related measure on the right on the same line, except for $\Delta I_{\mathbf{R}_{su}}$, which has two associated decoding-oriented measures: ΔI^{D} and ΔI^{LD}. The distinction between the measures on pink and on gray background relies on the fact that $\Delta I_{\mathbf{R}_{su}}$ does not involve a decoding process. Instead, $\Delta I_{\hat{\mathbf{S}}}$, $\Delta I_{\hat{S}}$ and $\Delta A_{\mathbf{R}_{su}}$ decode a stimulus (or rank the stimuli) with decoding method β. This decoding is not meant to be applicable to real experiments, since (as opposed to the truly decoding-oriented measures on the right, that operate with method α) the decoding is applied to the surrogate responses $\mathbf{R}_{su}$, not the real ones $\mathbf{R}_{ex}$.

We here describe the measures briefly, and refer the interested reader to the original papers.

In Equation (21), $I_{\mathbf{R}_{ex}}$ and $I_{\mathbf{R}_{su}}$ are the mutual informations between the set of stimuli and a set of responses governed by the distributions $P_{ex}(\mathbf{r}|s)$ and $P_{su}(\mathbf{r}|s)$, respectively. Thus, $\Delta I_{\mathbf{R}_{su}}$ is the simplest way in which the information encoded by the true responses can be compared with that of the surrogate responses. This comparison has been employed for more than six decades in neuroscience [34,35] to study, for example, the encoding of different stimulus features in spike counts, in synchronous spikes, and in other forms of spike patterns, both in single neurons and populations (see [30] and references therein).

The measure ΔI^{D} defined in Equation (25) was introduced by Nirenberg et al. [8] to study the role of noise correlations, and was later extended to arbitrary deterministic mappings [10,12,13]. Here we

use the supra-script D to indicate that the measure is the "divergence" (in the Kullback-Leibler sense) between the posterior stimulus distributions calculated with the real and the surrogate responses, respectively. In [10], Nirenberg and Latham argued that the important feature of ΔI^D is that it represents the information loss of a mismatched decoder trained with $P_{\text{su}}(\mathbf{r}|s)$ but operated on the real responses, sampled from $P_{\text{ex}}(\mathbf{r}|s)$. Not before long, Schneidman et al. [9] noticed that ΔI^D can exceed $I_{\mathbf{R}_{\text{ex}}}$. The interpretation of ΔI^D as a measure of information loss would imply that decoders trained with surrogate responses can lose more information than the one encoded by the real response. In fact, ΔI^D tends to infinity if $P_{\text{su}}(s|\mathbf{r}) \to 0$ when $P(s|\mathbf{r}) > 0$ for some s. In the limit, ΔI^D becomes undefined when $P_{\text{su}}(\mathbf{r}) = 0$ and $P_{\text{ex}}(\mathbf{r}) > 0$. To avoid this peculiar behavior, Latham and Nirenberg generalized the theoretical framework used to derive ΔI^D [11], giving rise to the measure ΔI^{DL} of Equation (26). Here, the supra-script DL makes reference to "Divergence Lowest", since the measure was presented as the lowest possible information loss of a decoder trained with $P_{\text{su}}(\mathbf{r}|s)$. In the definition of ΔI^{LD}, the parameter θ is a real scalar. The distribution $P_{\text{su}}(s|\mathbf{r},\theta)$ was defined by Latham and Nirenberg [11] as proportional to $P(s)\,P_{\text{su}}(\mathbf{r}|s)^\theta$. This definition has several problems, as discussed in [11,17,36–39]. In Appendix B.1 we demonstrate a theorem that resolves the issues appearing in previous definitions, and justifies the use of

$$P_{\text{su}}(s|\mathbf{r},\theta) \propto \begin{cases} P(s) & \text{if } \exists \hat{s}, \hat{\mathbf{r}} \text{ such that } P_{\text{ex}}(\hat{\mathbf{r}}|\hat{s}) > P_{\text{su}}(\hat{\mathbf{r}}|\hat{s}) = 0 \\ 0 & \text{if } P_{\text{su}}(\mathbf{r}|s) = P_{\text{ex}}(\mathbf{r}|s) = 0 \text{ for some but not all } s \\ P(s)\,P_{\text{su}}(\mathbf{r}|s)^\theta & \text{otherwise} \end{cases} \tag{30}$$

From the conceptual point of view, ΔI^{DL} represents the information loss of a mismatched decoder trained with $P_{\text{su}}(\mathbf{r}|s)$ and operated on $\mathbf{R}_{\text{ex}}$. Latham and Nirenberg [11] showed that, unlike ΔI^D, it is possible to demonstrate that $\Delta I^{DL} \leq I_{\mathbf{R}_{\text{ex}}}$. Hence, ΔI^{DL} never yields a tested feature encoding more information than the full response. The proof in [11] ignored a few specific cases that we discuss in the Theorem A1 of Appendix B.1. Still, even in those additional cases, the inequality $\Delta I^{DL} \leq I_{\mathbf{R}_{\text{ex}}}$ holds.

In Equations (22) and (23), $\hat{\mathbf{S}}$ and $\hat{S}$ denote a sorted stimulus list and the most-likely stimulus, respectively, both decoded by evaluating Equation (6) (or its ranked version) on a response $\mathbf{r}$ sampled from the surrogate distribution $P_{\text{su}}(\mathbf{r}|s)$ (method β). Estimating mutual informations using decoders can be traced back at least to Gochin et al. [40], and comparing the estimations of two decoders that take different response features into account, at least to Warland et al. [41].

The measures $\Delta I_{\hat{\mathbf{S}}}$ and $\Delta I_{\hat{S}}$ are paired with ΔI^{LS} and ΔI^B, respectively, since the latter are obtained from the former when replacing the decoding method from β to α. The measure ΔI^{LS} was introduced by Ince et al. [20], and quantifies the difference between the information in $\mathbf{R}_{\text{ex}}$, and the one in the output of decoders that, after observing a variable $\mathbf{r}$ sampled with distribution $P_{\text{ex}}(\mathbf{r}|s)$ (method α), produce a stimulus list sorted according to $P_{\text{su}}(s|\mathbf{r})$. The supra-script LS indicates "List of Stimuli". Similarly, ΔI^B, quantifies the difference between the information encoded in $\mathbf{R}_{\text{ex}}$ and that encoded in the output of a decoder trained by inserting $P_{\text{su}}(s|\mathbf{r})$ into Equation (6), and operated on $\mathbf{r}$ sampled with distribution $P_{\text{ex}}(\mathbf{r}|s)$ (method α). The supra-script B stands for the "Bayesian" nature of the involved decoder. The use of these measures can be traced back at least to Nirenberg et al. [8], although in that case, decoders were restricted to be linear. The measure $\Delta I_{\hat{\mathbf{S}}}$ of Equation (22) is new, and we have introduced it here as the homologous of ΔI^{LS}. When the number of stimuli is two, $\Delta I_{\hat{\mathbf{S}}} = \Delta I_{\hat{S}}$, since selecting the optimal stimulus is (as a computation) in one-to-one correspondence with ranking the two candidate stimuli.

The accuracy loss $\Delta A_{\mathbf{R}_{\text{su}}}$ defined in Equation (24) entails the comparison between the performance of two decoders, one trained with and applied on $\mathbf{R}_{\text{ex}}$, and one trained with and applied on $\mathbf{R}_{\text{su}}$. Such comparisons have also a long history in neuroscience [42,43] (see [9,12] for further discussion). The accuracy loss ΔA^B also compares two decoders. The first, is the same as for $\Delta A_{\mathbf{R}_{\text{su}}}$, but the second is trained with $\mathbf{R}_{\text{su}}$ and applied on $\mathbf{R}_{\text{ex}}$.

The measures ΔI^{LS}, ΔI^B, and ΔA^B are undefined if the actual responses $\mathbf{R}_{\text{ex}}$ are not contained in the set of surrogate responses $\mathbf{R}_{\text{su}}$. In other words, a decoder constructed with $P_{\text{su}}(\mathbf{r}|s)$ does not know what output to produce when evaluated in a response $\mathbf{r}$ for which $P_{\text{su}}(\mathbf{r}) = 0$. This situation

never happens when evaluating the relevance of noise correlations with $P_{\text{su}} = P_{NI}$, but it may well be encountered in more general situations, as for example, in Figure 3B.

2.4. Relating the Values Obtained with Different Measures

If a mapping $\mathbf{R}_{\text{ex}} \rightarrow \mathbf{R}_{\text{su}}$ exists transforming $P_{\text{ex}}(\mathbf{r}|s)$ into $P_{\text{su}}(\mathbf{r}|s)$, we may use the decoding procedure of Equation (6) to construct the transformation chain $\mathbf{R}_{\text{ex}} \rightarrow \mathbf{R}_{\text{su}} \rightarrow \hat{\mathbf{S}} \rightarrow \hat{S}$ [17,44]. Consequently, $\Delta I_{\mathbf{R}_{\text{su}}}$, $\Delta I_{\hat{\mathbf{S}}}$ and $\Delta I_{\hat{S}}$ can be interpreted as accumulated information losses after the first, second and third transformations, respectively, and $\Delta A_{\mathbf{R}_{\text{su}}}$, as the accuracy loss after the first transformation. The data processing theorems (Section 2.1.3) ensure that these measures are never negative. This property, however, cannot be guaranteed in the absence of a reduced transformation $\mathbf{R}_{\text{ex}} \rightarrow \mathbf{R}_{\text{su}}$, stochastic or deterministic. Indeed, in the example of Figure 4b, if both stimuli are equiprobable, and both responses $\mathbf{R}_{\text{ex}}$ associated with ◯ are equiprobable, then $\Delta I_{\mathbf{R}_{\text{su}}} = \Delta I_{\hat{\mathbf{S}}} = \Delta I\ \hat{S} \approx -79\,\%$ of $I_{\mathbf{R}_{\text{ex}}} \approx 0.31$ bits, implying that the surrogate responses encode more information about the stimulus than the original, experimental responses. Removing the correlations between spike count and latency, hence, increases the information, so correlations can be concluded to be detrimental to information encoding.

Irrespective of whether a (deterministic or stochastic) mapping $\mathbf{R}_{\text{ex}} \rightarrow \mathbf{R}_{\text{su}}$ exists, the data processing inequality guarantees that $\Delta I_{\mathbf{R}_{\text{su}}} \leq \Delta I_{\hat{\mathbf{S}}} \leq \Delta I_{\hat{S}}$, since $\hat{\mathbf{S}}$ is a deterministic function of $\mathbf{R}_{\text{su}}$, and $\hat{S}$ is a deterministic function of $\hat{\mathbf{S}}$. The inequality holds irrespective of the sign of each measure.

All decoder-oriented measured are guaranteed to be non-negative. The very definitions of ΔI^{D} and of ΔI^{DL} imply they cannot be negative, since they are both Kullback-Leibler divergences between two probability distributions. The sequence of reduced transformations $\mathbf{R}_{\text{ex}} \rightarrow \hat{\mathbf{S}} \rightarrow S$, in turn, guarantees the non-negativity of ΔI^{LS}, ΔI^{B} and ΔA^{B}, through the Data Processing Inequalities.

In order to assess whether decoding-oriented measures are always larger or smaller than their encoding (or gray) counterparts, we performed a numerical exploration comparing each encoding/gray-oriented measure with its decoding-oriented homologue. The exploration was conducted by calculating the values of these measures for a large collection of possible stimulus prior probabilities $P(s)$, and response conditional probabilities $P_{\text{ex}}(\mathbf{r}|s)$ in the examples of Figures 2–4 and 7. The details of the numerical exploration are in Appendix A. The measures in the first group were sometimes greater and sometimes smaller than those of the second group, depending on the case and the probabilities (Table 1). Consequently, our results demonstrate that there is no general rule by which measures of one type bound the measures of the other type.

The exploration also included the example of Figure 7a. In panel (a), the transition probabilities are

$$Q(\mathbf{r}_{\text{su}}|\mathbf{r}_{\text{ex}}) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}, \tag{31}$$

where rows and columns enumerate the elements of the ordered sets $\mathcal{R}_{ex} = \mathcal{R}_{su} = \{[1], [2], [3], [4]\}$ from where both $\mathbf{R}_{\text{ex}}$ and $\mathbf{R}_{\text{su}}$ are sampled. For panel b,

$$Q(\mathbf{r}_{\text{su}}|\mathbf{r}_{\text{ex}}) = \frac{1}{2} \begin{bmatrix} 2\,\bar{a} & a & a & 0 & 0 & 0 \\ 0 & b & b & 2\,\bar{b} & 0 & 0 \\ 0 & 0 & 0 & 2\,\bar{b} & b & b \\ 2\,\bar{a} & 0 & 0 & 0 & a & a \end{bmatrix}, \tag{32}$$

with $0 < a, b < 1$, rows enumerating the elements of $\mathcal{R}_{ex} = \{[2], [3], [5], [6]\}$, and columns those of $\mathcal{R}_{su} = \{[1], [2], [3], [4], [5], [6]\}$.

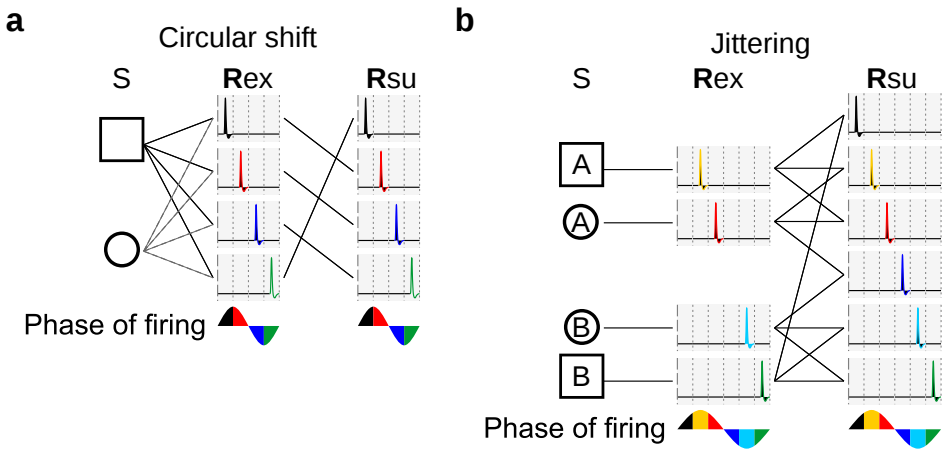

Figure 7. Stochastic codes may play different roles in encoding and decoding. (**a**) Hypothetical experiment with two stimuli □ and ○, which are transformed (solid and dashed lines) into neural responses containing a single spike ($C = 1$) fired at different phases (Φ) with respect to a cycle of 20 ms period starting at stimulus onset. The phases have been discretized in intervals of size $\pi/2$ and wrapped to the interval $[0, 2\pi)$. The encoding process is followed by a circular phase-shift that transforms $\mathbf{R}_{\text{ex}}{=}\Phi$ into another code $\mathbf{R}_{\text{su}}{=}\hat{\Phi}$ with transition probabilities $Q(\mathbf{r}_{\text{su}}|\mathbf{r}_{\text{ex}})$ defined by Equation (31). The set of all $\mathbf{R}_{\text{su}}$ coincides with the set of all $\mathbf{R}_{\text{ex}}$; (**b**) Same as (a), except that stimuli are four (🄰, Ⓐ, 🄱, and Ⓑ), and phases are measured with respect to a cycle of 30 *ms* period and discretized in intervals of size $\pi/3$. The encoding process is followed by a stochastic transformation (lines on the right) that introduces jitter, thereby transforming $\mathbf{R}_{\text{ex}}{=}\Phi$ into another code $\mathbf{R}_{\text{su}}{=}\hat{\Phi}$ with transition probabilities $Q(\mathbf{r}_{\text{su}}|\mathbf{r}_{\text{ex}})$ defined by Equation (32).

Table 1. Numerical exploration of the maximum and minimum differences between several measures of information and accuracy losses. The values are expressed as percentages of $I_{\mathbf{R}_{\text{ex}}}$ (the information encoded in $\mathbf{R}_{\text{ex}}$) or $A^{\mathbf{R}_{\text{ex}}}_{\mathbf{R}_{\text{ex}}}$ (the maximum accuracy above chance level when decoders operate on $\mathbf{R}_{\text{ex}}$). All examples involve two stimuli, so $\Delta I_{\hat{\mathbf{S}}}{=}\Delta I_{\hat{S}}$ and $\Delta I^{LS}{=}\Delta I^{B}$. The absolute value of $\Delta A_{\mathbf{R}_{\text{su}}}{-}\Delta A^{B}$ can become extremely large when $A^{\mathbf{R}_{\text{su}}}_{\mathbf{R}_{\text{su}}}{\approx}0$. Dashes represent cases in which decoding-oriented measures are undefined, as explained in Section 2.4.

Cases		Figure 4a	Figure 2b	Figure 2c	Figure 3d	Figure 2a	Figure 3b	Figure 7a
$\Delta I_{\hat{\mathbf{R}}}{-}\Delta I^{D}$	min	−79	−51	−34	0	0	—	≤999
	max	26	32	51	0	0	—	−20
$\Delta I_{\hat{\mathbf{R}}}{-}\Delta I^{DL}$	min	−34	−32	−16	0	0	−100	−100
	max	59	41	98	0	0	0	0
$\Delta I_{\hat{\mathbf{R}}}{-}\Delta I^{B}$	min	−67	−62	−46	−63	−87	—	−100
	max	57	81	96	0	0	—	0
$\Delta I_{\hat{S}}{-}\Delta I^{D}$	min	−79	−48	−34	0	0	—	≤999
	max	67	92	93	63	87	—	70
$\Delta I_{\hat{S}}{-}\Delta I^{DL}$	min	−34	−27	−16	0	0	-100	−100
	max	91	92	99	63	87	0	97
$\Delta I_{\hat{S}}{-}\Delta I^{B}$	min	−51	−31	−17	0	0	—	−100
	max	59	91	98	0	0	—	100
$\Delta A_{\hat{\mathbf{R}}}{-}\Delta A^{B}$	min	−386	−200	−150	0	0	—	≤999
	max	95	67	100	0	0	—	0

An important issue is to identify the situations in which $\Delta I_{\mathbf{R}_{\text{su}}}$ gives exactly the same result as either ΔI^{D} or ΔI^{DL}. It is not easy to determine the conditions for the equality between $\Delta I_{\mathbf{R}_{\text{su}}}$ and ΔI^{DL}. Yet, for the equality between $\Delta I_{\mathbf{R}_{\text{su}}}$ and ΔI^{D}, and in the specific case in which $P_{\text{su}}(\mathbf{r}|s) = P_{NI}(\mathbf{r}|s)$ as given by Equation (17), the following theorem holds.

Theorem 3. *When assessing the relevance of noise correlations,* $\Delta I^D = \Delta I_{\mathbf{R}_{su}}$ *if and only if*

$$\lambda = \sum_{\mathbf{r}} [P_{ex}(\mathbf{r}) - P_{su}(\mathbf{r})] \log_2 [P_{su}(\mathbf{r})] = 0. \tag{33}$$

Moreover, $\lambda \lessgtr 0$ *implies that* $\Delta I^D \lessgtr \Delta I_{\mathbf{R}_{su}}$.

Proof. See Appendix B.4. □

Equation (33) implies that neither the prior stimulus probabilities $P(s)$ nor the conditional response probabilities $P_{ex}(\mathbf{r}|s)$ intervene in the condition for the equality, beyond the effect they have in fixing the value of $P_{ex}(\mathbf{r})$ and $P_{su}(\mathbf{r})$. Each response $\mathbf{r}$ makes a contribution to the value of λ, which favours ΔI^D whenever $P_{su}(\mathbf{r}) > P_{ex}(\mathbf{r})$, and $I_{\mathbf{R}_{ex}}$ in the opposite case. As pointed out by [10], all responses $\mathbf{r}$ for which $P_{ex}(\mathbf{r}) = 0$ and $P_{su}(\mathbf{r}) > 0$ give a null contribution to ΔI^D, and a negative contribution to $I_{\mathbf{R}_{ex}}$, implying that correlations in such responses are irrelevant for decoding, and detrimental to encoding.

The fact that encoding-oriented measures neither bound nor are bounded by decoding-oriented measures is a daunting result. If, when working in a specific example, one gets a positive value with one measure and a negative value with another, the interpretation must carefully distinguish between the two paradigms. One may wonder, however, if such distinction is also required when correlations are absolutely essential for one of the measures, in that they capture the whole of the encoded information. Could the other measure conclude that they are irrelevant? Or that they are only mildly relevant? Luckily, in this case, the answer is negative. In other words, when the tested feature is fundamental, then ΔI^D and $\Delta I_{\mathbf{R}_{su}}$ coincide, and no conflict arises between encoding and decoding, as proven by the following theorem:

Theorem 4. $\Delta I^{DL} = I_{\mathbf{R}_{ex}}$ *if and only if* $\Delta I_{\mathbf{R}_{su}} = I_{\mathbf{R}_{ex}}$*, regardless of whether stochastic codes exist that map the actual responses* $\mathbf{R}_{ex}$ *into the surrogate responses* $\mathbf{R}_{su} = \mathbf{R}_{NI}$ *generated assuming noise independence.*

Proof. See Appendix B.5. □

The conclusion is that if a given feature is 100% relevant for encoding, then it is also 100% relevant for decoding, and vice versa. Hence, although $\Delta I_{\mathbf{R}_{su}}$ and ΔI^{DL} often differ in the relevance they ascribe to a given feature, the discrepancy is only encountered when the tested feature is not the only informative feature in play. When the removal of the feature is catastrophic (in the sense that it brings about a complete information loss), then both $\Delta I_{\mathbf{R}_{su}}$ and ΔI^{DL} diagnose the situation equally.

2.5. Relation between Measures Based on Decoding Strategies α and β

The results of Table 1 may seem puzzling because decoding happens after encoding. Therefore—one may naively reason—the data processing theorems should have forbidden both $\Delta I_{\mathbf{R}_{su}}$ to surpass ΔI^D, ΔI^{DL}, or ΔI^B, as well as $\Delta A_{\mathbf{R}_{su}}$ to surpass ΔA^B. However, even though decoding indeed happens after encoding, the data processing theorem is not violated. The theorem certainly ensures that $\Delta I_{\mathbf{R}_{su}}$ and $\Delta A_{\mathbf{R}_{su}}$ constitute lower bounds for measures related to decoders that operate on responses generated by $P_{su}(\mathbf{r}|s)$, but not for measures related to decoders that operate on responses generated by $P_{ex}(\mathbf{r}|s)$, such as happens with ΔI^D, ΔI^{DL}, ΔI^B, and ΔA^B.

This observation about the validity of the data processing inequality is different from the one discussed in Section 2.2. There, we discussed the conditions under which $\Delta I_{\mathbf{R}_{su}}$ could be guaranteed to be non-negative, the crucial factor being the existence of a stochastic mapping $\mathbf{R}_{ex} \to \mathbf{R}_{su}$. Now we are discussing a different aspect, regarding whether decoding-related measures can or cannot be bounded by encoding-oriented measures. The conclusion is that in general terms, the answer is negative, because decoding-related measures operate with decoding strategy α, a strategy never addressed by the encoding measures. The surrogate variable $\mathbf{R}_{su}$ participating in the encoding measure $\Delta I_{\mathbf{R}_{su}}$ is *not* the response decoded by the measures of Equations (25)–(28), so the data processing inequalities need

not hold. That being said, there are specific instances in which both types of measures coincide, two of them discussed in Theorems 3 and 4 and a third case later in Theorem 5.

Other explanations have been given in the literature for the fact that sometimes, decoding oriented measures surpass their encoding counterparts. For example, it has been alleged [10] that when $\Delta I^D, \Delta I^{DL}$ or ΔI^B are smaller than $\Delta I_{\mathbf{R}_{su}}$, this is either due to (a) the impossibility to define a stimulus-independent reduction $\mathbf{R}_{ex} \rightarrow \mathbf{R}_{su}$ that yields $P_{ex}(\mathbf{r}|s) \rightarrow P_{su}(\mathbf{r}|s)$ (and therefore the data-processing inequality is not guaranteed to hold), or due to (b) the fact that surrogate responses often sample values of response space that are never reached by real responses (and therefore, the losses of matched decoders may be larger than the ones of mismatched ones). However, Figure 2c constitutes a counterexample of both arguments, since there, the stimulus-independent stochastic reduction exists, and the response set of $\mathbf{R}_{ex}$ and $\mathbf{R}_{su}$ coincide.

One could also wonder whether the discrepancy between the values obtained with encoding-oriented measures and decoding-oriented measures only occurs in examples where a stochastic reduction $\mathbf{R}_{ex} \rightarrow \mathbf{R}_{su}$ exists, and the involved transition matrix $Q(\mathbf{r}_{su}|\mathbf{r}_{ex})$ depends on the joint probabilities $P_{ex}(\mathbf{r}, s)$, and not only on the marginals, as discussed in Theorem 2. However, Figure 2b,c provide examples in which $Q(\mathbf{r}_{su}|\mathbf{r}_{ex})$ does not depend on $P(\mathbf{r}, s)$, and yet, the discrepancies are still observed.

The distinction between decoding strategies α and β is also crucial when using the measure ΔI^D. This measure was introduced by Nirenberg et al. [8] for the specific case in which the tested feature is the amount of noise correlations, that is, when $P_{su}(s|\mathbf{r}) = P_{NI}(s|\mathbf{r})$. The measure was later extended to arbitrary deterministic mappings $\mathbf{R}_{su} = f(\mathbf{R}_{ex})$ [10,12,13], with the instruction to use an expression like Equation (25), but with $P_{su}(s|\mathbf{r})$ replaced by $P(s|\mathbf{R}_{su} = f(\mathbf{r})) = P_{su}(s|f(\mathbf{r}))$. It should be noted, however, that as soon as this replacement is made, ΔI^D becomes exactly equal to $\Delta I_{\mathbf{R}_{su}}$. Specifically, the measure ΔI^D now describes the information loss of a decoder that operates on a response variable generated with the surrogate distribution $P_{su}(\mathbf{r}|s)$ (decoding method β). If we want to keep the original spirit, and associate ΔI^D with a decoder that operates on a response variable generated with the real distribution $P_{ex}(\mathbf{r}|s)$ (decoding method α), in Equation 25, $P_{su}(s|\mathbf{r})$ should not be modified. Only the evaluation of the surrogate variable $\mathbf{R}_{su}$ in the experimentally observed value $\mathbf{R}_{ex} = \mathbf{r}$ describes a mismatched decoder constructed with $P_{su}(\mathbf{r}|s)$ and operated on $\mathbf{R}_{ex}$ (mathematical details in Appendix C).

2.6. Assessing the Type of Information Encoded by Individual Response Features

When the stimulus contains several attributes (as shape, color, sound, etc.), by removing a specific response feature it is possible to assess not only *how much* information is encoded by the feature, but also, *what type* of information. Identifiying the type of encoded information implies determining the stimulus feature represented by the tested response feature. As shown in this section, the type of encoded information is as dependent on the method of removal as is the amount. In other words, the different measures defined in Equations (21)–(29) sometimes associate a feature with the encoding of different stimulus attributes.

In the example of Figure 8, we use four compound stimuli $S = [S_F, S_L]$, generated by choosing independently a frame ($S_F = \square$ or $\bigcirc$) and a letter ($S_L =$ A or B), thereby yielding 🄰, Ⓐ, 🄱, and Ⓑ. Stimuli are transformed into neural responses $\mathbf{R} = [L, C]$ with different number of spikes ($1 \leq C \leq 5$) fired at different first-spike latencies ($1 \leq L \leq 4$; time has been discretized in 5 ms bins). Latencies are only sensitive to frames whereas spikes counts are only sensitive to letters, thereby constituting independent-information streams: $P(s, \mathbf{r}) = P(s_F, l)\, P(s_L, c)$ [33]. The equality in the numerical value of two measures does not imply that both measures assign the same meaning to the information encoded by the tested response feature. Indeed, the two measures may sometimes report the tested response feature to encode two different aspects of the set of stimuli. Consider a decoder that is trained using the noisy data $\mathbf{R}_{su}$ shown in Figure 8a, but it is asked to operate on either the same noisy data with which it was trained (strategy β), or with the quality data $\mathbf{R}_{ex}$ of Figure 8b (strategy α). The information

losses $\Delta I_{\mathbf{R}_{su}}$, ΔI^{D}, and ΔI^{DL} are all equal to 50 % of $I(S, \mathbf{R}_{ex}) = 2$ bits. Therefore, the information loss is independent of whether, in the operation phase, the decoder is fed with responses generated with $P_{su}(\mathbf{r}|s)$ or with $P_{ex}(\mathbf{r}|s)$.

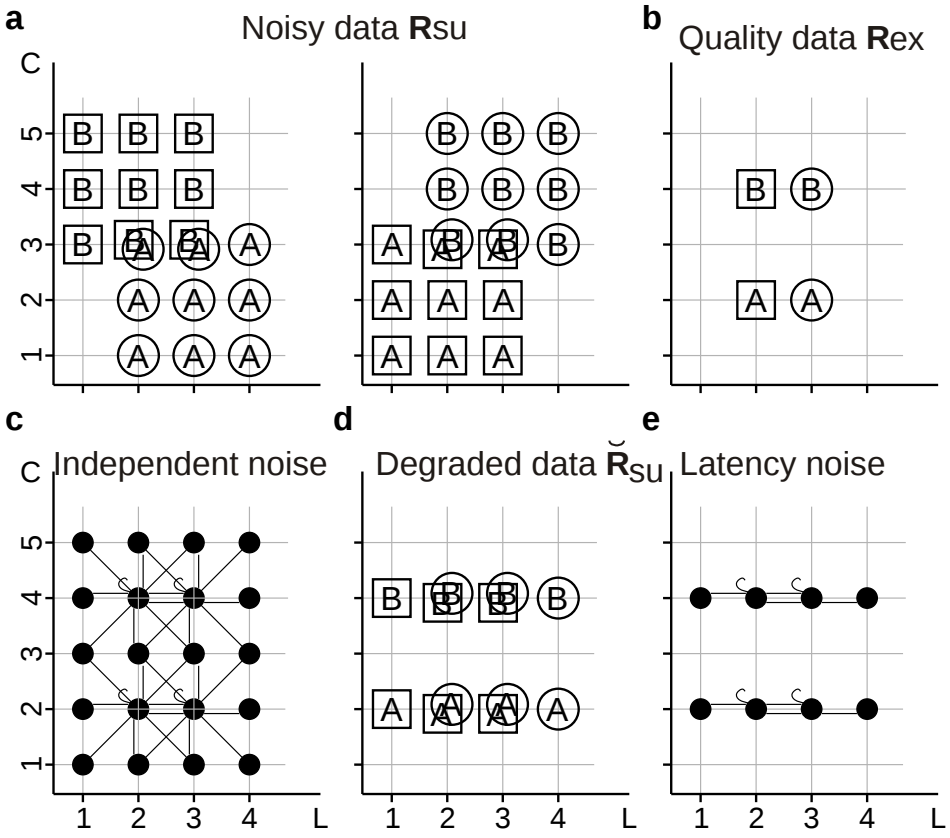

Figure 8. Assessing the amount and type of information encoded by . (**a**) Noisy data $\mathbf{R}_{su} = [L, C]$ recorded in response of the compound stimulus $S = [S_F, S_L]$; (**b**) Quality data $\mathbf{R}_{ex} = [L, C]$ recorded in the case of panel (a), but without noise; (**c**) Stimulus-independent stochastic transformation with transition probabilities $Q(\mathbf{r}_{su}|\mathbf{r}_{ex})$ given by Equation (34), that introduces independent noise both in the latencies and in the spike counts, thereby transforming $\mathbf{R}_{ex}$ into $\mathbf{R}_{su}$ and rendering $\mathbf{R}_{su}$ a stochastic code; (**d**) Degraded data $\check{\mathbf{R}}$ obtained by adding latency noise to the quality data; (**e**) Representation of the stimulus-independent stochastic transformation $\mathbf{R}_{ex} \to \check{\mathbf{R}}$ with transition probabilities $Q(\check{\mathbf{r}}|\mathbf{r}_{ex})$ given by Equation (35) that adds latency noise in panel (d).

The transformation $Q(\mathbf{r}_{su}|\mathbf{r}_{ex})$ causes some responses $\mathbf{R}_{su}$ to occur for all stimuli, so when decoding with method β, some information about frames is lost (that is, $I(S_F, \mathbf{R}_{su}) \approx 33\,\%$ of $I(S_F, \mathbf{R}_{ex}) = 1$ bit), as well as some information about letters (that is, $I(S_L, \mathbf{R}_{su}) \approx 67\,\%$ of $I(S_L, \mathbf{R}_{ex}) = 1$ bit). In other words, decoding $\mathbf{R}_{su}$ causes a partial information loss $\Delta I_{\mathbf{R}_{su}}$ that is composed of both frame and letter information. Instead, when decoding $\mathbf{R}_{ex}$ with method α, there is no information loss about letters: For the responses $\mathbf{R}_{ex}$ that actually occur, the decoder trained with $\mathbf{R}_{su}$ can perfectly identify the letters, because $P_{su}(C = 2|S_L = \mathrm{A}) = P_{su}(C = 4|S_L = \mathrm{B}) = 1$. The information about frames, on the other hand, is completely lost, since $P_{su}(l|\square) = P_{su}(l|\bigcirc)$ whenever l adopts a value that actually occurs in $\mathbf{R}_{ex}$, namely 2 or 3. This example shows that the fact that two decoding procedures give the same numerical loss does not mean that they draw the same conclusions regarding the role of the tested feature in the neural code. Ananalogous computations yield analogous results for the hypothetical experiment shown in Figure 7b.

If responses $\mathbf{r}_{\text{ex}}$ and $\mathbf{r}_{\text{su}}$ are written as vectors $[L, C]$, and the values of $Q(\mathbf{r}_{\text{su}}|\mathbf{r}_{\text{ex}})$ are arranged in a rectangular structure, in Figure 8c the transition probabilities are

$$Q(\mathbf{r}_{\text{su}}|\mathbf{r}_{\text{ex}}) = \frac{1}{9}\begin{bmatrix} 1&1&1&0&1&1&1&0&1&1&1&0&0&0&0&0&0&0&0&0\\ 0&1&1&1&0&1&1&1&0&1&1&1&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&1&1&1&0&1&1&1&0&1&1&1&0\\ 0&0&0&0&0&0&0&0&0&1&1&1&0&1&1&1&0&1&1&1 \end{bmatrix}, \tag{34}$$

where rows and columns indicate the ordered sets $\mathcal{R}_{ex}=\{[2,2],[3,2],[2,4],[3,4]\}$ and $\mathcal{R}_{su}=\{1,2,3,4\}\times\{1,2,3,4,5\}$, where $\times$ denotes the Cartesian product with colexicographical order, that is, ordered as $[1,1],[2,1],[3,1],[4,1],[1,2]$, etc. In Figure 8e

$$Q(\check{\mathbf{r}}|\mathbf{r}_{\text{ex}}) = \frac{1}{3}\begin{bmatrix} 0&0&0&0&1&1&1&0&0&0&0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&1&1&1&0&0&0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&0&0&0&1&1&1&0&0&0&0 \end{bmatrix}, \tag{35}$$

with rows and columns with the same convention as in Equation (34).

Finally, the noisy data (Figure 8a) can be obtained by transforming the degraded data (Figure 8d) with the transition matrix

$$Q(\mathbf{r}_{\text{su}}|\check{\mathbf{r}}) = \frac{1}{3}\begin{bmatrix} 1&0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&0&0&0\\ 0&1&0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&0&0\\ 0&0&1&0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0&0\\ 0&0&0&1&0&0&0&1&0&0&0&1&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&1&0&0&0\\ 0&0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&1&0&0\\ 0&0&0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&1&0\\ 0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&1 \end{bmatrix}. \tag{36}$$

with rows and columns indicating the ordered sets $\{[1,2,3,4]\times[2,4]\}$ and $\{1,2,3,4\}\times\{1,2,3,4,5\}$, respectively, where $\times$ denotes the Cartesian product with colexicographical order.

2.7. Conditions for Equality of the Amount and Type of Information Loss Reported by Different Measures

We now derive the conditions under which encoding/gray-oriented measures coincide with their decoding-oriented counterparts, as observed in Figures 2a and 3d. That is, we derive the conditions under which the following equalities hold:

$$\Delta I_{\mathbf{R}_{\text{su}}}=\Delta I^{D}=\Delta I^{DL}, \tag{37}$$

$$\Delta I_{\hat{\mathbf{S}}}=\Delta I^{LS}, \tag{38}$$

$$\Delta I_{\hat{S}}=\Delta I^{B}, \tag{39}$$

$$\Delta A_{\mathbf{R}_{\text{su}}}=\Delta A^{B}. \tag{40}$$

The example in Figure 7a showed that the existence of deterministic mappings does not suffice for a qualitative and quantitative equivalence of different measures. Furthermore, the example of Figure 3b showed that the equalities require the space of $\mathbf{R}_{\text{su}}$ to include the space of $\mathbf{R}_{\text{ex}}$, or else the decoding method α may be undefined. We demonstrate that the Equations (37)–(40) arise, and moreover, that there is no discrepancy in the type of information assessed by these different measures, whenever the mapping from $\mathbf{R}_{\text{ex}}$ into $\mathbf{R}_{\text{su}}$ can be described using positive-diagonal idempotent stochastic matrices [45]. Specifically, we prove the following theorem:

Theorem 5. *Consider a stimulus-independent stochastic function f from a representation $\mathbf{R}_{\text{ex}}$ into another representation $\mathbf{R}_{\text{su}}$, such that the range $\mathcal{R}$ of $\mathbf{R}_{\text{su}}$ includes that of $\mathbf{R}_{\text{ex}}$, and with transition probabilities $Q(\mathbf{r}_{\text{su}}|\mathbf{r}_{\text{ex}})$ that can be written as positive-diagonal idempotent right stochastic matrices with row and column indices that enumerate the elements of $\mathcal{R}$ in the same order. Then, Equations (37)–(40) hold.*

Proof. See Appendix B.6. □

The theorem states that the equalities of Equations (37)–(40) can be guaranteed whenever the removal of the tested response feature involves a (deterministic or) stochastic mapping $\mathbf{R}_{ex} \rightarrow \mathbf{R}_{su}$ that induces a partition within the set of real responses $\mathbf{R}_{ex}$, and $\mathbf{R}_{su}$ is obtained by rendering all responses inside each partition indistinguishable (but not across partitions). To sample $\mathbf{R}_{su}$, the probabilities of individual responses inside each partition are re-assigned, rendering their distinction uninformative [30].

This theorem provides sufficient but not necessary conditions for the equalities to hold. The important aspect, however, is that it ensures that the equalities hold not only in numerical value, but also, in the type of information that different measures ascribe to the tested feature. Two different methods preserve or lose information of different type if, when decoding a stimulus, the trials with decoding errors tend to confound different attributes of the stimulus, as in the example of Figure 8. The conditions of Theorem 5, however, ensure that the strategies α and β always decode exactly the same stimulus (see Appendix B.6), so there can be no difference in the confounded attributes. Pushing the argument further, one could even argue that responses (real or surrogate) encode more information than the identity of the stimulus that originated them. For a fixed decoded stimulus, the response still contains additional information [46], that refers to (a) the degree of certainty with which the stimulus is decoded, and (b) the rank of the alternative stimuli, in case the decoded stimulus was mistaken [20]. Both meanings are embodied in the whole rank of a posteriori probabilities $P_{su}(s|\mathbf{r})$, not just the maximal one. Yet, under the conditions of the theorem, the entire rankings obtained with methods α and β coincide (see Appendix B.6). Therefore, even within this broader interpretation, there can be no difference in the qualitative aspects of the information preserved or lost by one and the other.

For example, in Figure 7b, we found that all information losses are equal (that is, $\Delta I_{\hat{\mathbf{R}}}$, $\Delta I_{\hat{\mathbf{S}}}$, $\Delta I_{\hat{S}}$, ΔI^{D}, ΔI^{DL}, ΔI^{LS}, and ΔI^{B} are all 50 %), and both accuracy losses are equal (that is, $\Delta A_{\hat{\mathbf{R}}}$ and ΔA^{B} are both ≈67 %). However, the conditions of Theorem 5 do not hold. The matrix of Equation (32) is not block-diagonal, nor it can be taken to that shape by incorporating new rows (to make it square), and permuting both rows and columns, in such a way that the response vectors are enumerated in the same order by both indices. For this reason, the losses are not guaranteed to be of the same type.

Instead, the transition probabilities of Equations (15) and (16) can be turned into positive-diagonal idempotent right stochastic matrices. Equation (15) is already in the required format. To take Equation (16) to the conditions of Theorem 5, two new rows need to be incorporated, associated to the responses $[4,1]$ and $[2,2]$, that do not occur experimentally. Those rows can contain arbitrary values, since the condition $P_{ex}([4,1]|S) = P_{ex}([2,2]|S) = 0$, $\forall\, S$ renders them irrelevant. Arranging the columns so that both rows and columns enumerate the same list of responses, Equation (16) can be written as

$$Q(\mathbf{r}_{su}|\mathbf{r}_{ex}) = \begin{bmatrix} b & \bar{b} & 0 & 0 & 0 & 0 \\ b & \bar{b} & 0 & 0 & 0 & 0 \\ 0 & 0 & c & \bar{c} & 0 & 0 \\ 0 & 0 & c & \bar{c} & 0 & 0 \\ 0 & 0 & 0 & 0 & d & \bar{d} \\ 0 & 0 & 0 & 0 & d & \bar{d} \end{bmatrix}, \tag{41}$$

with $\mathcal{R}_{ex}{=}\mathcal{R}_{su}{=}\{[2,1],[2,2],[3,1],[3,2],[4,1],[4,2]\}$. Hence, in these two examples, both the amount and type of information of encoding and decoding-based measures coincide.

2.8. Improving the Performance of Decoders Operating with Strategy α

In a previous paper [17], we demonstrated that neither ΔI^{D} nor ΔI^{DL} constitute lower bounds on the information loss induced by decoders constructed by disregarding the tested response feature. This means that some decoders may exist, that perform better than $D_{su}(\mathbf{r})$ defined in Equation (6). In this section we discuss one possible way in which some of these improved decoders may be constructed, inspired in the example of Figure 8. Quite remarkably, the construction involves the addition of noise to the real responses, before feeding them to the decoder of Equation (6). Panel (a) shows a decoder constructed with noisy data ($\mathbf{R}_{su}$), and then employed to decode quality data ($\mathbf{R}_{ex}$;

Figure 8b), thereby yielding information losses $\Delta I^D = \Delta I^{DL} = 50\,\%$. These losses can be decreased by feeding the decoder with a degraded version $\check{\mathbf{R}}$ of the quality data (Figure 8d) generated through a stimulus-independent transformation that adds latency noise (Figure 8e). Decoding $\mathbf{R}_{\text{ex}}$ as if it were $\mathbf{R}_{\text{su}}$ by first transforming $\mathbf{R}_{\text{ex}}$ into $\check{\mathbf{R}}$ results in $\Delta I^D = \Delta I^{DL} \approx 33\,\%$, thereby recovering 33 % of the information previously lost. On the contrary, adding spike-count noise will tend to increase the losses. Thus, adding suitable amounts and type of noise can increase the performance of approximate decoders, and the result is not limited to the case in which the response aspect is the amount of noise correlations. In addition, this result also indicates that, contrary to previously thought [47], decoding algorithms need not match the encoding mechanisms for performing optimally from an information-theoretical standpoint. All these results are a consequence of the fact that decoders operating with strategy α are not optimal, so it is possible to improve their performance by deterministic or stochastic manipulations of the response. In practice, our results open up the possibility of increasing the efficiency of decoders constructed with approximate descriptions of the neural responses, usually called approximate or mismatched decoders, by adding suitable amounts and types of noise to the decoder input.

3. Related Issues

3.1. Relation to Decomposition-Based Methods

Many measures of different types have been developed to assess how different response features of the neural code interact with each other. Some are based on direct comparisons between the information encoded by individual features, or collections of features (see for example [48–50], to cite just a few among many). Others distinguish between two or more potential dynamical models of brain activity [51], for example, by differentiating between conditional and unconditional correlations between neurons in the frequency domain [52]. Yet others, rely on decompositions or projections based on information geometry. In those, the mutual information between stimuli and responses $I_{\mathbf{R}}$ is broken down as $I_{\mathbf{R}} = \sum_i I'_{R_i} + \text{Synergy Terms} + \text{Redundancy Terms}$, where I'_{R_i} represents the information contributed by the individual response feature R_i, and the remaining terms incorporate the synergy or redundancy between them. In the original approaches [53–57], the terms I'_{R_i} represented the information $I(R_i;S)$ encoded in single response aspects irrespective of what be encoded in other aspects. In later studies, [58–62], these terms accounted for the information that is *only* encoded in individual aspects, taking care of excluding whatever be redundant with other aspects. The approach discussed in this paper is in the line of the studies Nirenberg et al. [8] and Schneidman et al. [9] and all their consequences. This line has some similarities and some discrepancies with the decomposition-based studies. We here comment on some of these relations.

- First, the measure $\Delta I_{\mathbf{R}_{\text{su}}}$ quantifies the relevance of a given feature with the difference $I_{\mathbf{R}_{\text{ex}}} - I_{\mathbf{R}_{\text{su}}}$. When the surrogate response $\mathbf{R}_{\text{su}}$ is equal to the original response $\mathbf{R}_{\text{ex}}$ with just a single component R_i eliminated, $\Delta I_{\mathbf{R}_{\text{su}}}$ is equal to $I(R_i;s|\bar{R}_i)$, where $\bar{R}_i$ is the collection of all response aspects except R_i. In this case, $\Delta I_{\mathbf{R}_{\text{su}}}$ coincides with the sum of the unique and the synergistic contributions of the dual decompositions in the newest set of methods [63].
- Second, when assessing the relevance of a given response feature, we are often inclined to draw conclusions about the cost of ignoring the tested feature when aiming to decode the original stimulus. As shown in this paper, those conclusions depend not only on how stimuli are encoded, but also, on how they are decoded. The decomposition-based methods are mainly focused in the encoding problem, so they are less suited to draw conclusions about decoding.
- Finally, as discussed in Figure 8, not only the amount of (encoded or decoded) information matters, but also, what type. Decomposition-based methods, although not yet reaching a full consensus in their formulation, provide a valuable attempt to characterize how both the type and the amount of information is structured within the set of analyzed variables, in a way that is complementary to the present approach, specifically in analyzing the structure of the lattices obtained by associating different response features [58,63].

3.2. The Problem of Limited Sampling

Throughout the paper we assumed that the distribution $P_{\text{ex}}(s, \mathbf{r})$ is known, or is accessible to the experimenter. In the examples, when we calculated information values, we plugged the true distributions into the formulas, without discussing the fact that such distribution may not be easily estimated with finite amounts of data. Whichever method is used to estimate $P_{\text{ex}}(s, \mathbf{r})$, to a larger or lesser degree, the outcome is no more than an approximation. Hence, even $I_{\mathbf{R}_{\text{ex}}}$ (which is supposed to be the full information) is estimated approximately. Since $P_{\text{su}}(s, \mathbf{r})$ is a modified version of $P_{\text{ex}}(s, \mathbf{r})$, also $P_{\text{su}}(s, \mathbf{r})$ can only be estimated approximately. Information measures, including Kullback-Leibler divergences, are highly sensitive to variations in the involved probabilities [20,32,64–69], and the latter are unavoidable in high-dimensional response spaces. The assessment of the relevance of a given feature, hence, requires experiments that contain sufficient samples so as to ensure that the correcting methods work. When the response space is large, the measures ΔI_S, ΔI^B and the loss of accuracies are less sensitive to limited sampling than $\Delta I_{\mathbf{R}_{\text{su}}}$, ΔI^D and ΔI^{LD}.

In addition, the problem of finite sampling can also be formulated as an attempt to determine the relevance of the feature "Accuracy in the estimation of $P_{\text{ex}}(\mathbf{r}|s)$". This feature is not a property of the nervous system, but rather, of our ability to characterise it. Still, the framework developed here can also handle this methodological problem. The estimated distribution can be interpreted as a stochastic modification $P_{\text{su}}(\mathbf{r}|s)$ of the true distribution $P_{\text{ex}}(\mathbf{r}|s)$. As long as the caveats discussed in this paper are taken into account, the measures of Equations (21)–(29) may serve to evaluate the cost of modeling $P_{\text{ex}}(\mathbf{r}|s)$ out of finite amounts of data.

4. Conclusions

Several measures have been proposed in the literature to assess the relevance of specific response features in the neural code. All proposals are based on the idea that by removing the tested feature from the response, the neural code deteriorates, and the lost information is a useful measure of the relevance of the feature. In this paper, we demonstrated that the neural code may or may not deteriorate when removing a response feature, depending on the nature of the tested feature, and on the method of removal, in ways previously unseen. First, we determined the conditions under which the data processing inequality can be invoked. Second, we showed that decoding-oriented measures may result in larger or smaller losses than their encoding (or gray) counterparts, even for response aspects that, unlike noise correlations, can be modeled as stimulus-independent transformations of the full response. Third, we demonstrated that both types of measures coincide under the conditions of Theorem 5. Fourth, we showed that evaluating the role of a response feature in the neural code involves not only an assessment of its contribution to the amount of encoded information, but also, to the meaning of that information. Such meaning is as dependent as the amount on the measure employed to assess it. Finally, our results open up the possibility that simple and cheap decoding strategies, based on the addition of an adequate type and amount of noise, be more efficient and resilient than previously thought. We conclude that the assessment of the relevance of a specific response feature cannot be performed without a careful justification for the selection of a specific method of removal.

Author Contributions: Conceptualization, methodology, software, validation, formal analysis, investigation, writing, visualization: H.G.E. Formal analysis, resources, writing, editing: I.S.

Funding: This work was supported by the Ella and Georg Ehrnrooth Foundation, Consejo Nacional de Investigaciones Científicas y Técnicas of Argentina (06/C444), Universidad Nacional de Cuyo (PIP 0256), and Agencia Nacional de Promoción Científica y Tecnológica (grant PICT Raíces 2016 1004).

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

Appendix A. On the Information and Accuracy Differences

Each value in Table 1 (except for those associated with Figures 3b; see below) was computed using the Nelder-Mead simplex algorithm for optimization, as implemented by the function `fminsearch` of Matlab 2016. For accuracy reasons, only examples in which $I_{\mathbf{R}_{ex}} \geq 10^{-6}$ bits and $A^{\mathbf{R}_{ex}}_{\mathbf{R}_{ex}} \geq 10^{-6}$ were considered. Furthermore, parameters defining the joint stimulus-response probabilities and the transition matrices were restricted to the interval $[0.05, 0.95]$. Each difference between two measures defined in Equations (21)–(29) was computed repeatedly, with random initial values for the stimulus-response probabilities and the transition matrices, until the value of the difference failed to increase or decrease in 20 consecutive runs.

The values in Table 1 for Figure 3b were computed analytically with $P_{ex}([3,2]) > 0$ or $P_{ex}([4,2]) > 0$, but not both. In those cases, the measures ΔI^D, ΔI^B, and ΔA^B are undefined, whereas $\Delta I^{DL} = 100\,\%$, for the reasons given in Section 2.4. However, $\Delta I_{\mathbf{R}_{su}}$ and $\Delta I_{\hat{S}}$ can vary between 0 % and 100 %, for example, attaining 0 % when $P_{ex}([3,1]) \to 0$, and 100 % when $P_{ex}([2,1]) \to 0$ and $P_{ex}([4,2]) \to 0$. The information $I_{\mathbf{R}_{ex}}$ equals the stimulus entropy, regardless of the response probabilities. The values in Table 1 for Figure 3d were computed by setting $b = c = d = 0.5$ in Equation (16). The values in Section 2.4 for Figure 7b were obtained by setting $P_{ex}(s, \mathbf{r}) = 1/4$ for the stimulus-response pairs shown in the figure, and are valid for any transition probability matrix set as in Equation (32) with $b{=}a$. The values in Section 2.4 for Figure 8 were obtained by setting $P_{ex}(s, \mathbf{r}) = 1/4$ for the stimulus-response pairs shown in the figure.

Appendix B. Proofs

Appendix B.1. Derivation of Equation (30)

The definition of ΔI^{DL} involves the probability $P_{su}(s|\hat{\mathbf{r}}, \theta)$ defined in [11,36,38] as proportional to $P(S)\prod_i P^\theta(R_i|S)$, where the exponent θ is chosen so as to maximize ΔI^{DL}. This definition has been recently shown to be invalid when $\exists\, \mathbf{r}$,s such that $P_{su}(\mathbf{r}|s) = 0$ for a stimulus s or a response $\mathbf{r}$ for which $P_{ex}(\mathbf{r}|s) \neq 0$ [17]. This problem never appears when evaluating the relevance of noise correlations with $P_{su}(\mathbf{r}|s) = P_{NI}(\mathbf{r}|s)$ as stated by Equation (17). Yet, it may well appear in more general cases, including those arising from stochastically reduced codes. To overcome it, we prove the theorem

Theorem A1. *The probability $P(s|\mathbf{r}, \theta)$ that appears in the definition of ΔI^{DL} is*

$$P_{su}(s|\mathbf{r}, \theta) \propto \begin{cases} P(s) & \text{if } \exists \hat{s}, \hat{\mathbf{r}} \text{ such that } P_{ex}(\hat{\mathbf{r}}|\hat{s}) > P_{su}(\hat{\mathbf{r}}|\hat{s}) = 0 \\ 0 & \text{if } P_{su}(\mathbf{r}|s) = P_{ex}(\mathbf{r}|s) = 0 \text{ for some but not all } s \\ P(s)\, P_{su}(\mathbf{r}|s)^\theta & \text{otherwise} \end{cases}$$

Proof. According to Latham and Nirenberg [11], the probability $P_{su}(s|\mathbf{r}, \theta)$ is the one that minimizes the Kullback-Leibler divergence $D_{KL}[P^*(\mathbf{r}, s)||p(\mathbf{r})p(s)]$ with respect to the distribution $P^*(\mathbf{r}, s)$, subject to the constraints

$$\langle \log_2 P_{su}(\mathbf{r}|s) \rangle_{P^*(\mathbf{r},s)} = \langle \log_2 q(\mathbf{r}|s) \rangle_{P(s,\mathbf{r})} \quad \text{(A1)}$$

$$\sum_s P^*(\mathbf{r}, s) = P(\mathbf{r}). \quad \text{(A2)}$$

The minimization problem can be formulated in terms of an objective function to be minimized, in which the constraints appear with Lagrange multipliers, and θ is the one accompanying Equation (A1). Using the standard conventions that $0 \log 0 = 0$ and $x \log 0 = \infty$ for $x > 0$, Equation (A1) is fulfilled if $\exists\, \hat{\mathbf{r}}$,$\hat{s}$ such that $P(\hat{s}|\hat{\mathbf{r}}, \theta) > 0$ if $P_{ex}(\hat{\mathbf{r}}, \hat{s}) > P_{su}(\hat{\mathbf{r}}|\hat{s}) = 0$. The first part of the theorem immediately follows by solving Equation (B15) in [11] as there indicated with $\beta = 0$. If $\nexists\, \hat{\mathbf{r}}$,$\hat{s}$ such that $P_{ex}(\hat{\mathbf{r}}, \hat{s}) > P_{su}(\hat{\mathbf{r}}|\hat{s}) = 0$, then Equation (A1) is fulfilled only if $P(s, \mathbf{r}|\theta) = 0$ when $P_{su}(\mathbf{r}|s) = P_{ex}(\mathbf{r}, s) = 0$. The second and third parts of the theorem immediately follows using Bayes' rule. □

Appendix B.2. Proof of Theorem 1

Proof. The second part is proved by the two examples in Figure 4. The first part was proved in [9], at least for cases in which the set of the surrogate responses $\mathbf{R}_{\text{su}} = \mathbf{R}_{NI}$ differ from the set of the real responses $\mathbf{R}_{\text{ex}}$. When they both coincide, we can prove the first part by contradiction, assuming that a deterministic mapping exists from $\mathbf{R}_{\text{ex}}$ into $\mathbf{R}_{NI}$. If both variables sample the same response space, the deterministic mapping must be one-to-one, otherwise the variable $\mathbf{R}_{NI}$ would sample a smaller set. Therefore, both $\mathbf{R}_{NI}$ and $\mathbf{R}_{\text{ex}}$ maximize the conditional entropy given S over the probability distributions with the same marginals, since one-to-one mappings do not modify the entropy, and $\mathbf{R}_{NI}$ is defined as the distribution with maximal conditional entropy with fixed marginals. Because the probability distribution achieving this maximum is unique [16], $P_{\text{su}}(\mathbf{r}|s)$ and $P_{\text{ex}}(\mathbf{r}|s)$ must be the same, thereby proving the theorem. □

Appendix B.3. Proof of Theorem 2

Proof. We prove the dependency on the marginal likelihoods by computing $Q(\mathbf{r}_{\text{su}}|\mathbf{r}_{\text{ex}})$ for the hypothetical experiment of Figure 4a, and observing that the result depends on the marginal likelihood $P_{\text{ex}}(L|s)$. To that end, we rewrite Equation (10) for $\mathbf{R}_{\text{su}} = [1,2]$ as

$$P_{\text{su}}([1,2]|\square) = P_{\text{ex}}([1,2]|\square)\, Q([1,2]||[1,2]) + P_{\text{ex}}([2,1]|\square)\, Q([1,2]||[2,1])\,.$$

Note that $P_{\text{ex}}([1,2]|\square){=}1{-}P_{\text{ex}}([2,1]|\square){=}P_{\text{ex}}(L{=}1|\square)$ and $P_{\text{su}}([1,2]|\square){=}P_{\text{ex}}(L{=}1|\square)^2$. Using this and rearranging the terms, we obtain the quadratic equation

$$P_{\text{ex}}(L{=}1|\square)^2 + P_{\text{ex}}(L{=}1|\square)\ \{Q([1,2]||[2,1]) - Q([1,2]||[1,2])\} - Q([1,2]||[2,1]) = 0\,,$$

that is solved by

$$P_{\text{ex}}(L{=}1|\square) = 0.5\ \left[\delta q + \left(\delta q^2 + 4\, Q([1,2]||[2,1])\right)^{0.5}\right]\,,$$

where $\delta q{=}Q([1,2]||[1,2]) - Q([1,2]||[2,1])$. Hence, any change in $P_{\text{ex}}(L=1|\square)$ must be followed by some change in $Q(\mathbf{r}_{\text{su}}|\mathbf{r}_{\text{ex}})$, thereby proving the first part.

We prove the dependency on the noise correlations by computing $Q(\mathbf{r}_{\text{su}}|\mathbf{r}_{\text{ex}})$ for the hypothetical experiment of Figure 5, and observing that the result not only depends on the marginal likelihoods $P_{\text{ex}}(L|s)$ and $P_{\text{ex}}(C|s)$, but in many cases, it also depends on the joint distributions $P_{\text{ex}}(L,C|s)$. Hence, varying the amount of noise correlations, even if keeping the marginals fixed, yields a variation in the mapping $Q(\mathbf{r}_{\text{su}}|\mathbf{r}_{\text{ex}})$.

We proceed by reductio ad absurdum. If $Q(\mathbf{r}_{\text{su}}|\mathbf{r}_{\text{ex}})$ does not depend on the amount of noise correlations in $P_{\text{ex}}(\mathbf{r}|s)$, we may assume that if we vary $P_{\text{ex}}(\mathbf{r}|s)$ but keep the marginals $P_{\text{ex}}(r_i|s)$ fixed, the transition probabilities $Q(\mathbf{r}_{\text{su}}|\mathbf{r}_{\text{ex}})$ remain unchanged. Under this hypothesis, Equation (10) is valid for many choices of $P_{\text{ex}}(\mathbf{r}|s)$. In this context, consider the set of all response distributions with the same marginals as $P_{\text{ex}}(\mathbf{r}|s)$ that can be turned into $P_{\text{su}}(\mathbf{r}|s)$ through $Q(\mathbf{r}_{\text{su}}|\mathbf{r}_{\text{ex}})$. This set includes $P_{\text{su}}(\mathbf{r}|s)$, and therefore, $Q(\mathbf{r}_{\text{su}}|\mathbf{r}_{\text{ex}})$ should be able to transform $P_{\text{su}}(\mathbf{r}|s)$ into itself. In addition, Property 2 requires that $Q(\mathbf{r}_{\text{su}}|[2,2]) = 0$ when $\mathbf{r}_{\text{su}} \neq [2,2]$ because either $P(\mathbf{r}_{\text{su}}|\square) = 0$ or $P(\mathbf{r}_{\text{su}}|\bigcirc) = 0$ for those responses. Normalization yields $Q([2,2]||[2,2]) = 1$. Furthermore, computing Equation (10) for $\mathbf{R}_{\text{su}} = [2,2]$ yields

$$0 = P_{\text{ex}}([1,1]|\square)\, Q([2,2]||[1,1]) + P_{\text{ex}}([1,2]|\square)\, Q([2,2]||[1,2]) + P_{\text{ex}}([2,1]|\square)\, Q([2,2]||[2,1])\,,$$

which shows that $Q([2,2]|\mathbf{r}_{\text{ex}}){=}0$ when $\mathbf{r}_{\text{ex}} \in \{[1,1],[1,2],[2,1]\}$. Consequently, the resulting $Q(\mathbf{r}_{\text{su}}|\mathbf{r}_{\text{ex}})$ yields through Equation 10 that $P_{\text{su}}([2,2]|\square) = P_{\text{ex}}([2,2]|\square)$. After noticing that

$$P_{\text{su}}([2,2]|\square){=}P_{\text{ex}}(L=2|\square)\, P_{\text{ex}}(C=2|\square)\,,$$

and that

$$P_{\text{ex}}([2,2]|\square) = P_{\text{ex}}(L=2|\square){-}P_{\text{ex}}([1,2]|\square) = P_{\text{ex}}(C=2|\square){-}P_{\text{ex}}([2,1]|\square)\,,$$

we can show that, after some straightforward algebra, Equation (10) only holds if $P_{\text{su}}(\mathbf{r}|\square) = P_{\text{ex}}(\mathbf{r}|\square)$ for all $\mathbf{r}$. Thus, the initial hypothesis yields a transition matrix $Q(\mathbf{r}_{\text{su}}|\mathbf{r}_{\text{ex}})$ that is unable to transform $\mathbf{R}_{\text{ex}}$ into $\mathbf{R}_{\text{su}}$ when $\mathbf{R}_{\text{ex}}$ is noise correlated, and thus $Q(\mathbf{r}_{\text{su}}|\mathbf{r}_{\text{ex}})$ necessarily depends on the amount of noise correlations in $\mathbf{R}_{\text{ex}}$. □

Appendix B.4. Proof of Theorem 3

Proof. The condition $\Delta I^D = \Delta I_{\mathbf{R}_{\text{su}}}$ implies that

$$\sum_{s\mathbf{r}} P_{\text{ex}}(s,\mathbf{r}) \ \log\left[\frac{P_{\text{ex}}(s|\mathbf{r})}{P_{\text{su}}(s|\mathbf{r})}\right] = I_{\mathbf{R}_{\text{ex}}} - I_{\mathbf{R}_{\text{su}}}. \tag{A3}$$

However,

$$\Delta I^D = I_{\mathbf{R}_{\text{ex}}} - \sum_{s\mathbf{r}} P_{\text{ex}}(s,\mathbf{r}) \ \log\left[\frac{P_{\text{su}}(\mathbf{r}|s)}{P_{\text{su}}(\mathbf{r})}\right].$$

Hence, Equation (A3) becomes

$$-\sum_{s\mathbf{r}} P_{\text{ex}}(s,\mathbf{r}) \ \log\left[\frac{P_{\text{su}}(\mathbf{r}|s)}{P_{\text{su}}(\mathbf{r})}\right] = -I_{\mathbf{R}_{\text{su}}} \tag{A4}$$

In addition, when evaluating the relevance of noise correlations, $P_{\text{su}}(\mathbf{r},s) = P(s)\ P_{NI}(\mathbf{r}|s)$ as established by Equation (17). Hence,

$$-\sum_{s\mathbf{r}} P_{\text{ex}}(s,\mathbf{r}) \ \log\left[\frac{P_{\text{su}}(\mathbf{r}|s)}{P_{\text{su}}(\mathbf{r})}\right] = \sum_j H(R_j|s) + \sum_{s\mathbf{r}} P_{\text{ex}}(\mathbf{r},s) \ \log[P_{\text{su}}(\mathbf{r})] \tag{A5}$$

$$-I_{\mathbf{R}_{\text{su}}} = \sum_j H(R_j|s) + \sum_{s\mathbf{r}} P_{\text{su}}(s,\mathbf{r}) \ \log[P_{\text{su}}(\mathbf{r})]. \tag{A6}$$

Replacing Equations (A5) and (A6) in Equation (A4),

$$\sum_{s\mathbf{r}} P_{\text{ex}}(s,\mathbf{r}) \ \log[P_{\text{su}}(\mathbf{r})] = \sum_{s\mathbf{r}} P_{\text{su}}(s,\mathbf{r}) \ \log[P_{\text{su}}(\mathbf{r})].$$

Summing in s, and rearranging,

$$\sum_{\mathbf{r}} [P_{\text{ex}}(\mathbf{r}) - P_{\text{su}}(\mathbf{r})] \ \log[P_{\text{su}}(\mathbf{r})] = 0.$$

If instead of an equality, we start with an inequality, that same inequality can be kept all through the proof. □

Appendix B.5. Proof of Theorem 4

Proof. Consider a neural code $\mathbf{R}_{\text{ex}} = [R_1, \ldots, R_N]$ and recall that the range of $\mathbf{R}_{NI}$ includes that of $\mathbf{R}_{\text{ex}}$. Therefore, $\Delta I^{DL} = I_{\mathbf{R}_{\text{ex}}}$ implies that the minimum in Eqution (26) is attained when $\theta = 0$. In that case, Equation (B13a) in [11] yields

$$\sum_{s,r_n} P(s,r_n) \ \log_2 P(r_n|s) = \sum_{s,r_n} P(s)\ P_{\text{ex}}(r_n) \ \log_2 P_{\text{ex}}(r_n|s), \quad \forall\, 1 \leq n \leq N, \quad \forall\, n.$$

After some more algebra and recalling that the Kullback-Leibler divergence is never negative, this equation becomes $I_{R_n} = 0$, implying that when read isolatedly, single responses contain no information about the stimulus. Consequently $\Delta I_{\mathbf{R}_{NI}} = I_{\mathbf{R}_{\text{ex}}}$, thereby proving the "only if" part. For the "if" part, it is sufficient to notice that the last equality implies that $P_{NI}(\mathbf{r}|s) = P_{NI}(\mathbf{r})$. □

Appendix B.6. Proof of Theorem 5

Proof. The conditions on f and $Q(\mathbf{r}_{su}|\mathbf{r}_{ex})$ ensure that $Q(\mathbf{r}_{su}|\mathbf{r}_{ex})$ can be written as a block-diagonal matrix, each block composed of the same rows with no zeros, and that each block can be associated with a non-overlapping partition $\mathcal{R}_1, \ldots, \mathcal{R}_M$ of the range of f. Under these conditions, $P(\mathbf{r}_{su}|\mathbf{r}_{ex}) = P(\mathbf{r}_{su}|\mathcal{R}_m)$ when $\mathbf{r}_{ex} \in \mathcal{R}_m$. Hence, for $\mathbf{r}_{su} \in \mathcal{R}_m$, $P(\mathbf{r}_{su}|s) = P(\mathbf{r}_{su}|\mathcal{R}_m)\, P(\mathcal{R}_m|s)$, yielding $P(s|\mathbf{r}_{su}) = P(s|\mathcal{R}_m)$ and $P(s|\mathbf{r}_{su}, \theta) = P(s|\mathcal{R}_m, \theta)$. Recomputing Equations (21)–(29) with these equalities in mind immediately yields the equalities in the theorem.

Even when the amount of information is equal, differences in the type of information may arise because the measures are based on different decoding strategies, here denoted α and β. However, under the conditions of the theorem, decoding strategy α and decoding strategy β are one and the same. Because $P(s|\mathbf{r}_{su}) = P(s|\mathcal{R}_m)$, both decoding strategies choose s only based on the partition $\mathcal{R}$ of $\mathbf{r}_{ex}$ or $\mathbf{r}_{su}$, respectively. Mathematically, both choose s according to

$$\hat{s} = \arg\max_s P(s|\mathcal{R}(\mathbf{r}))\,,$$

where $\mathcal{R}(\mathbf{r})$ denotes the mapping from $\mathbf{r}$ into $\mathcal{R}$, which is the same regardless of whether $\mathbf{r}$ is $\mathbf{r}_{ex}$ or $\mathbf{r}_{su}$. Because $Q(\mathbf{r}_{su}|\mathbf{r}_{ex})$ maps each partition onto itself, the responses within each partition of $\mathbf{r}_{su}$ is completely generated by the responses in each partition of $\mathbf{r}_{ex}$, and thus the decoding strategies are applied to the same set of $\mathbf{r}_{ex}$. Hence, both decoding strategies are defined and operate in the same manner, yielding the same information. □

Appendix C. On the Computation of ΔI^D

The information loss caused by mismatched decoders (decoding strategy α) when $\mathbf{R}_{su}{=}f(\mathbf{R}_{ex})$ has previously been computed as ΔI^D but with $P_{su}(s|\mathbf{r})$ replaced by $P(s|\mathbf{R}_{su} = f(\mathbf{r})) = P_{su}(s|f(\mathbf{r}))$ [10,12,13]. The latter represents the probability of s given that $\mathbf{R}_{su}$ takes the value $f(\mathbf{r})$, thereby limiting f to deterministic mappings. However, the probabilities $P_{su}(s|\mathbf{r})$ and $P_{su}(s|f(\mathbf{r}))$ are not equivalent, since

$$\begin{aligned} P_{su}(s|\mathbf{r}) &\propto \sum_{\mathbf{r} = f(\hat{\mathbf{r}})} P_{ex}(\hat{\mathbf{r}}, s) \\ P_{su}(s|f(\mathbf{r})) &\propto \sum_{f(\mathbf{r}) = f(\hat{\mathbf{r}})} P_{ex}(\hat{\mathbf{r}}, s) \end{aligned}$$

These two definitions raise the question of which alternative is the appropriate one when computing the information loss caused by mismatched decoders.

To resolve this question, notice that replacing $P_{su}(s|\mathbf{r})$ with $P_{su}(s|f(\mathbf{r}))$ in Equation (6) yields the decoding algorithm

$$\hat{s} = \arg\max_s P_{su}(s|f(\mathbf{r}))\,.$$

This algorithm entails first transforming the observed $\mathbf{r}$ into $\mathbf{r}_{su}{=}f(\mathbf{r})$, and then choosing the stimulus $\hat{s} = D_{su}(\hat{\mathbf{r}})$ with a matched probability. Hence, its operation is analogous to the decoding algorithm β, and not, as originally intended, to the decoding algorithm α.

To illustrate the difference, recall the experiment in Figure 7a and suppose that the observed response is $\mathbf{r} = 0.25\pi$. The decoding algorithm α reads this value, computes $P_{su}(s|0.25\pi)$, and decodes $\hat{s} = \square$. Instead, the decoding algorithm proposed in [10,12,13], first transforms the value of $\mathbf{r} = 0.25\pi$ into $f(\mathbf{r}) = 0.75\pi$, then computes $P_{su}(s|0.75\pi)$, and finally decodes $\hat{s} = \bigcirc$. This mode of operation corresponds to the decoding algorithm β.

The above discrepancy can also be seen from the change in the operational meaning of ΔI^D caused by the replacement. To that end, recall that ΔI^D was first introduced as a comparison between the average number of binary questions required to identify s after observing $\mathbf{r}$ when using two optimal

question-asking strategies, one tailored for $P_{\text{ex}}(s|\mathbf{r})$ and the other for $P_{\text{su}}(s|\mathbf{r})$ [8]. Mathematically, this difference can be written as

$$\Delta I^D = \sum_{s,\mathbf{r}} P_{\text{ex}}(s,\mathbf{r}) \, \log_2 P_{\text{ex}}(s|\mathbf{r}) - \sum_{s,\mathbf{r}} P_{\text{ex}}(s,\mathbf{r}) \, \log_2 P_{\text{su}}(s|\mathbf{r}) \, . \tag{A7}$$

In each term, the argument of the logarithms is determined by the question-asking strategy, whereas the weight of the averages is determined by the probability distribution of the variables on which the strategy is applied [8,10,16]. Equation (A7) describes the decoding strategy α.

Replacing $P_{\text{su}}(s|\mathbf{r})$ with $P_{\text{su}}(s|f(\mathbf{r}))$ turns Equation (A7) into

$$\begin{aligned}
\Delta \tilde{I} &= \sum_{s,\mathbf{r}} P_{\text{ex}}(s,\mathbf{r}) \, \log_2 P_{\text{ex}}(s|\mathbf{r}) - \sum_{s,\mathbf{r}} P_{\text{ex}}(s,\mathbf{r}) \, \log_2 P(s|f(\mathbf{r})) \\
&= \sum_{s,\mathbf{r}} P_{\text{ex}}(s,\mathbf{r}) \, \log_2 P_{\text{ex}}(s|\mathbf{r}) - \sum_{s,\mathbf{r}_{\text{su}}} P_{\text{su}}(s,\mathbf{r}_{\text{su}}) \, \log_2 P_{\text{su}}(s|\mathbf{r}_{\text{su}}) \\
&= \Delta I_{\mathbf{R}_{\text{su}}} \, .
\end{aligned}$$

Unlike ΔI^D, this difference compares the average number of binary questions required to identify s after observing $\mathbf{r}$ using a question-asking strategy that is optimal for $P_{\text{ex}}(s|\mathbf{r})$, with the average number of binary questions required to identify s after observing $\mathbf{r}_{\text{su}}$ using the a question-asking strategy that is optimal for $P_{\text{su}}(s|\mathbf{r}_{\text{su}})$. This is the way the decoding strategy β operates, not α.

Naively, one may think that a change in $P_{\text{su}}(s|\mathbf{r})$, regardless of its size, may turn the measure $\Delta I_{\mathbf{R}_{\text{su}}}$, typically regarded as an encoding-oriented measure and here linked to the decoding algorithm β, into the decoding-oriented measure ΔI^D. However, notice that this change cannot occur through the equations above due to the change induced in $P_{\text{su}}(s,\mathbf{r}_{\text{su}})$. For that to actually occur, one must write $\Delta I_{\mathbf{R}_{\text{su}}}$ differently, as for example:

$$\Delta I_{\mathbf{R}_{\text{su}}} = \sum_{s,\mathbf{r}} P_{\text{ex}}(s,\mathbf{r}) \, \log_2 P_{\text{ex}}(s|\mathbf{r}) - \sum_{s,\mathbf{r}_{\text{su}}} P_{\text{ex}}(s,\mathbf{r}) \, \log_2 P_{\text{su}}(s|\mathbf{r}_{\text{su}}) \, .$$

In this reformulation, the second term can be interpreted as the average number of binary questions required to identify s after observing $\mathbf{r}$ using a question-asking strategy that is optimal for $P_{\text{su}}(s|\mathbf{r}_{\text{su}})$, but only after converting $\mathbf{r}$ into $\mathbf{r}_{\text{su}}$. Any change in $P_{\text{su}}(s|\mathbf{r}_{\text{su}})$ immediately renders $P_{\text{su}}(s|\mathbf{r}_{\text{su}})$ a mismatched probability for $\mathbf{r}_{\text{su}}$, and makes the second term represent the average number of binary questions required to identify s after observing $\mathbf{r}$ using the question-asking strategy that is optimal for an altered version of $P_{\text{su}}(s|\mathbf{r}_{\text{su}})$ but only after converting $\mathbf{r}$ into $\mathbf{r}_{\text{su}}$, which need not resemble the meaning of the second term in ΔI^D.

References

1. Adrian, E.D. The impulses produced by sensory nerve endings. *J. Physiol.* **1926**, *61*, 49–72. [CrossRef] [PubMed]
2. Hubel, D.H.; Wiesel, T.N. Receptive fields of single neurones in the cat's striate cortex. *J. Physiol.* **1959**, *148*, 173–180. [CrossRef]
3. Thorpe, S.; Fize, D.; Marlot, C. Speed of processing in the human visual system. *Nature* **1996**, *6582*, 520–522. [CrossRef] [PubMed]
4. Abeles, M. *Corticonix: Neural Circuits of the Cerebral Cortex*; Cambridge University Press: Cambridge, UK, 1991.
5. Gray, C.M.; König, P.; Engel, A.K.; Singer, W. Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature* **1989**, *6213*, 334–337. [CrossRef] [PubMed]
6. Franke, F.; Fiscella, M.; Sevelev, M.; Roska, B.; Hierlemann, A.; da Silveira, R.A. Structures of Neural Correlation and How They Favor Coding. *Neuron* **2016**, *89*, 409–422. [CrossRef] [PubMed]
7. O'Keefe, J. Hippocampues, theta, and spatial memory. *Curr. Opin. Neurobiol.* **1993**, *6*, 917–924. [CrossRef]
8. Nirenberg, S.; Carcieri, S.M.; Jacobs, A.L.; Latham, P.E. Retinal ganglion cells act largely as independent encoders. *Nature* **2001**, *411*, 698–701. [CrossRef] [PubMed]
9. Schneidman, E.; Bialek, W.; Berry, M.J. Synergy, redundancy, and independence in population codes. *J. Neurosci.* **2003**, *23*, 11539–11553. [CrossRef] [PubMed]

10. Nirenberg, S.; Latham, P.E. Decoding neuronal spike trains: How important are correlations? *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 7348–7353. [CrossRef] [PubMed]
11. Latham, P.E.; Nirenberg, S. Synergy, redundancy, and independence in population codes, revisited. *J. Neurosci.* **2005**, *25*, 5195–5206. [CrossRef] [PubMed]
12. Quiroga, R.Q.; Panzeri, S. Extracting information from neuronal populations: Information theory and decoding approaches. *Nat. Rev. Neurosci.* **2009**, *10*, 173–185. [CrossRef] [PubMed]
13. Latham, P.E.; Roudi, Y. Role of correlations in population coding. In *Principles of Neural Coding*; Panzeri, S., Quian Quiroga, R., Eds.; CRC Press: Boca Raton, FL, USA, 2013; Chapter 7, pp. 121–138.
14. Casella, G.; Berger, R.L. *Statistical Inference*, 2nd ed.; Duxbury Press: Duxbury, MA, USA, 2002.
15. Panzeri, S.; Brunel, N.; Logothetis, N.K.; Kayser, C. Sensory neural codes using multiplexed temporal scales. *Trends Neurosci.* **2010**, *33*, 111–120. [CrossRef] [PubMed]
16. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley-Interscience: New York, NY, USA, 2006.
17. Eyherabide, H.G.; Samengo, I. When and why noise correlations are important in neural decoding. *J. Neurosci.* **2013**, *33*, 17921–17936. [CrossRef] [PubMed]
18. Knill, D.C.; Pouget, A. The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends Neurosci.* **2004**, *27*, 712–719. [CrossRef] [PubMed]
19. van Bergen, R.S.; Ma, W.J.; Pratte, M.S.; Jehee, J.F.M. Sensory uncertainty decoded from visual cortex predicts behavior. *Nat. Neurosci.* **2015**, *18*, 1728–1730. [CrossRef] [PubMed]
20. Ince, R.A.A.; Senatore, R.; Arabzadeh, E.; Montani, F.; Diamond, M.E.; Panzeri, S. Information-theoretic methods for studying population codes. *Neural Netw.* **2010**, *23*, 713–727. [CrossRef] [PubMed]
21. Reinagel, P.; Reid, R.C. Temporal coding of visual information in the thalamus. *J. Neurosci.* **2000**, *20*, 5392–5400. [CrossRef] [PubMed]
22. Panzeri, S.; Petersen, R.S.; Schultz, S.R.; Lebedev, M.; Diamond, M.E. The Role of Spike Timing in the Coding of Stimulus Location in Rat Somatosensory Cortex. *Neuron* **2001**, *29*, 769–777. [CrossRef]
23. Rokem, A.; Watzl, S.; Gollisch, T.; Stemmler, M.; Herz, A.V.M.; Samengo, I. Spike-timing precision underlies the coding efficiency of auditory receptor neurons. *J. Neurophysiol.* **2006**, *95*, 2541–2552. [CrossRef] [PubMed]
24. Lefebvre, J.L.; Zhang, Y.; Meister, M.; Wang, X.; Sanes, J.R. γ-Protocadherins regulate neuronal survival but are dispensable for circuit formation in retina. *Development* **2008**, *135*, 4141–4151. [CrossRef] [PubMed]
25. Victor, J.D.; Purpura, K.P. Nature and precision of temporal coding in visual cortex: A metric-space analysis. *J. Neurophysiol.* **1996**, *76*, 1310–1326. [CrossRef] [PubMed]
26. Victor, J.D. Spike train metrics. *Curr. Opin. Neurobiol.* **2005**, *15*, 585–592. [CrossRef] [PubMed]
27. Boyd, S.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: Cambridge, UK, 2004.
28. Fano, R.M. *Transmission of Information*; The MIT Press: Cambridge, MA, USA, 1961.
29. DeWeese, M.R.; Meister, M. How to measure the information gained from one symbol. *Netw. Comput. Neural Syst.* **1999**, *10*, 325–340. [CrossRef]
30. Eyherabide, H.G.; Samengo, I. Time and category information in pattern-based codes. *Front. Comput. Neurosci.* **2010**, *4*, 145. [CrossRef] [PubMed]
31. Eckhorn, R.; Pöpel, B. Rigorous and extended application of information theory to the afferent visual system of the cat. I. Basic concepts. *Kybernetik* **1974**, *16*, 191–200. [CrossRef] [PubMed]
32. Panzeri, S.; Treves, A. Analytical estimates of limited sampling biases in different information measures. *Network* **1996**, *7*, 87–107. [CrossRef] [PubMed]
33. Eyherabide, H.G. Disambiguating the role of noise correlations when decoding neural populations together. *arXiv* **2016**, arXiv:1608.05501.
34. MacKay, D.M.; McCulloch, W.S. The limiting information capacity of a neuronal link. *Bull. Math. Biophys.* **1952**, *14*, 127–135. [CrossRef]
35. Fitzhugh, R. The statistical detection of threshold signals in the retina. *J. Gen. Physiol.* **1957**, *40*, 925–948. [CrossRef] [PubMed]
36. Merhav, N.; Kaplan, G.; Lapidoth, A.; Shamai Shitz, S. On information rates for mismatched decoders. *IEEE Trans. Inf. Theory* **1994**, *40*, 1953–1967. [CrossRef]
37. Oizumi, M.; Ishii, T.; Ishibashi, K.; Hosoya, T.; Okada, M. A general framework for investigating how far the decoding process in the brain can be simplified. In *Advances in Neural Information Processing Systems*; The MIT Press: Cambridge, MA, USA, 2009; pp. 1225–1232.

38. Oizumi, M.; Ishii, T.; Ishibashi, K.; Hosoya, T.; Okada, M. Mismatched decoding in the brain. *J. Neurosci.* **2010**, *30*, 4815–4826. [CrossRef] [PubMed]
39. Oizumi, M.; Amari, S.I.; Yanagawa, T.; Fujii, N.; Tsuchiya, N. Measuring Integrated Information from the Decoding Perspective. *PLoS Comput. Biol.* **2016**, *12*, e1004654. [CrossRef] [PubMed]
40. Gochin, P.M.; Colombo, M.; Dorfman, G.A.; Gerstein, G.L.; Gross, C.G. Neural ensemble coding in inferior temporal cortex. *J. Neurophysiol.* **1994**, *71*, 2325–2337. [CrossRef] [PubMed]
41. Warland, D.K.; Reinagel, P.; Meister, M. Decoding visual information from a population of retinal ganglion cells. *J. Neurophysiol.* **1997**, *78*, 2336–2350. [CrossRef] [PubMed]
42. Optican, L.M.; Richmond, B.J. Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex. III. Information theoretic analysis. *J. Neurophysiol.* **1987**, *57*, 162–178. [CrossRef] [PubMed]
43. Salinas, E.; Abbott, L.F. Transfer of coded information from sensory neurons to motor networks. *J. Neurosci.* **1995**, *10*, 6461–6476. [CrossRef]
44. Geisler, W.S. Sequential ideal-observer analysis of visual discriminations. *Psychol. Rev.* **1989**, *96*, 267–314. [CrossRef] [PubMed]
45. Högnäs, G.; Mukherjea, A. *Probability Measures on Semigroups: Convolution Products, Random Walks and Random Matrices*, 2nd ed.; Springer: New York, NY, USA, 2011.
46. Samengo, I.; Treves, A. The information loss in an optimal maximum likelihood decoding. *Neural Comput.* **2002**, *14*, 771–779. [CrossRef] [PubMed]
47. Shamir, M. Emerging principles of population coding: In search for the neural code. *Curr. Opin. Neurobiol.* **2014**, *25*, 140–148. [CrossRef] [PubMed]
48. Gawne, T.J.; Richmond, B.J. How independent are the messages carried by adjacent inferior temporal cortical neurons? *J. Neurosci.* **1993**, *13*, 2758–2771. [CrossRef] [PubMed]
49. Gollisch, T.; Meister, M. Rapid Neural Coding in the Retina with Relative Spike Latencies. *Science* **2008**, *5866*, 1108–1111. [CrossRef] [PubMed]
50. Reifenstein, E.T.; Kemptner, R.; Schreiber, S.; Stemmler, M.B.; Herz, A.V.M. Grid cells in rat entorhinal cortex encode physical space with independent firing fields and phase precession at the single-trial level. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 6301–6306. [CrossRef] [PubMed]
51. Park, H.J.; Friston, K. Nonlinear multivariate analysis of neurophysiological signals. *Science* **2013**, *6158*, 1238411. [CrossRef] [PubMed]
52. Dahlhaus, R.; Eichler, M.; Sandkühler, J. Identification of synaptic connections in neural ensembles by graphical models. *J. Neurosci. Methods* **1997**, *77*, 93–107. [CrossRef]
53. Panzeri, S.; Schultz, S.R.; Treves, A.; Rolls, E.T. Correlations and the encoding of information in the nervous system. *Proc. R. Soc. B Biol. Sci.* **1999**, *266*, 1001–1012. [CrossRef] [PubMed]
54. Schultz, S.R.; Panzeri, S. Temporal Correlations and Neural Spike Train Entropy. *Phys. Rev. Lett.* **2001**, *25*, 5823–5826. [CrossRef] [PubMed]
55. Panzeri, S.; Schultz, S.R. A Unified Approach to the Study of Temporal, Correlational, and Rate Coding. *Neural Comput.* **2001**, *13*, 1311–1349. [CrossRef] [PubMed]
56. Pola, G.; Thiele, A.; Hoffmann, K.P.; Panzeri, S. An exact method to quantify the information transmitted by different mechanisms of correlational coding. *Network* **2003**, *14*, 35–60. [CrossRef] [PubMed]
57. Hernández, D.G.; Zanette, D.H.; Samengo, I. Information-theoretical analysis of the statistical dependencies between three variables: Applications to written language. *Phys. Rev. E.* **2015**, *92*, 022813. [CrossRef] [PubMed]
58. Williams, P.L.; Beer, R.D. Nonnegative decomposition of multivariate information. *arXiv* **2010**, arXiv:1004.2515.
59. Harder, M.; Salge, C.; Polani, D. Bivariate Measure of Redundant Information. *Phys. Rev. E.* **2013**, *87*, 012130. [CrossRef] [PubMed]
60. Griffith, V.; Koch, C. Quantifying Synergistic Mutual Information. In *Guided Self-Organization: Inception*; Prokopenko, M., Ed.; Springer: New York, NY, USA, 2014; Chapter 6, pp. 159–190.
61. Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J.; Ay, N. Quantifying unique information. *Entropy* **2014**, *16*, 2161–2183. [CrossRef]
62. Ince, R.A.A. Measuring Multivariate Redundant Information with Pointwise Common Change in Surprisal. *Entropy* **2017**, *19*, 318. [CrossRef]

63. Chicharro, D.; Panzeri, S. Synergy and Redundancy in Dual Decompositions of Mutual Information Gain and Information Loss. *Entropy* **2017**, *19*, 71. [CrossRef]
64. Wolpert, D.H.; Wolf, D.R. Estimating functions of probability distributions from a finite set of samples. *Phys. Rev. E.* **1996**, *52*, 6841–6973. [CrossRef]
65. Samengo, I. Estimating probabilities from experimental frequencies. *Phys. Rev. E* **2002**, *65*, 046124. [CrossRef] [PubMed]
66. Nemenman, I.; Bialek, W.; de Ruyter van Steveninck, R. Entropy and information in neural spike trains: Progress on the sampling problem. *Phys. Rev. E.* **2004**, *69*, 056111. [CrossRef] [PubMed]
67. Paninski, L. Estimation of entropy and mutual information. *Neural Comput.* **2003**, *6*, 1191–1253. [CrossRef]
68. 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. [CrossRef] [PubMed]
69. Montemurro, M.A.; Senatore, R.; Panzeri, S. Tight data-robust bounds to mutual information combining shuffling and model selection techniques. *Neural Comput.* **2007**, *11*, 2913–2957. [CrossRef] [PubMed]