

When and Why Noise Correlations Are Important in Neural Decoding

Hugo Gabriel Eyherabide^{1,2} and Inés Samengo¹

¹Centro Atómico Bariloche and Instituto Balseiro, R8402AGP San Carlos de Bariloche, Argentina; and ²Department of Computer Science and Helsinki Institute for Information Technology, University of Helsinki, 00560 Helsinki, Finland

Information may be encoded both in the individual activity of neurons and in the correlations between their activities. Understanding whether knowledge of noise correlations is required to decode all the encoded information is fundamental for constructing computational models, brain–machine interfaces, and neuroprosthetics. If correlations can be ignored with tolerable losses of information, the readout of neural signals is simplified dramatically. To that end, previous studies have constructed decoders assuming that neurons fire independently and then derived bounds for the information that is lost. However, here we show that previous bounds were not tight and overestimated the importance of noise correlations. In this study, we quantify the exact loss of information induced by ignoring noise correlations and show why previous estimations were not tight. Further, by studying the elementary parts of the decoding process, we determine when and why information is lost on a single-response basis. We introduce the minimum decoding error to assess the distinctive role of noise correlations under natural conditions. We conclude that all of the encoded information can be decoded without knowledge of noise correlations in many more situations than previously thought.

Introduction

A fundamental problem in neuroscience is to determine the simplest way to decode all the information encoded by neural populations. To decode all the information, it suffices to know the probabilistic mapping between the stimulus and the population activity (Oram et al., 1998; Knill and Pouget, 2004). When neurons are noise correlated (i.e., for each stimulus, their activities are correlated), the mapping must be built by measuring the joint activity of all neurons in the population; the construction demands large amounts of data and becomes experimentally and computationally intractable as the number of neurons increases (Nirenberg and Latham, 2003; Quiroga and Panzeri, 2009). However, when neurons are noise independent, the mapping can be built by measuring the activity of each neuron in the population one at a time, drastically reducing the amount of data required for the construction. If we assume that neurons are noise independent even when they are not, can we still decode all of the encoded information?

To answer this question, previous studies have estimated the inefficiency of decoders that were constructed assuming that neurons are noise independent (Nirenberg and Latham, 2003; Latham and Nirenberg, 2005; Ince et al., 2010; Oizumi et al.,

2010). Whenever all of these noise-independent (NI) decoders are inefficient, noise correlations are considered crucial for decoding. Otherwise, noise correlations are judged dispensable (Nirenberg et al., 2001; Averbeck et al., 2006). However, the conclusions drawn from these studies are still controversial. For pairs of neurons, the information lost by NI decoders was found to be <10% (Nirenberg et al., 2001; Graf et al., 2011; Pita-Almenar et al., 2011) and noise correlations were considered unimportant. However, pairs of neurons do not capture the complexity of large neural populations. Recent theoretical and experimental studies have revealed cases in which the information loss grows with the number of neurons (Averbeck et al., 2006; Klam et al., 2008; Ince et al., 2010; Oizumi et al., 2010), suggesting that noise correlations can indeed be important in neural decoding. Unfortunately, as we show here, the estimators used in these studies miscalculate the inefficiency of NI decoders in a context-dependent manner.

The exact estimation of the inefficiency of NI decoders is fundamental for assessing whether noise correlations are important in neural decoding and whether the losses are tolerable in practical applications. To that end, we here represent all NI decoders as sequences of transformations, separating the effect of the bare assumption that neurons are noise independent (the NI assumption) from the specific criteria used to select the decoded stimulus. We then quantify the information loss and the increment in the decoding error induced solely by the NI assumption and prove that the best NI decoder can achieve these bounds.

Equally important is determining how the neural code is transformed as a consequence of the NI assumption. We identify which response features are informative, which ones constitute noise, and which response features are preserved by the NI assumption and the subsequent transformations in the decoding

Received Jan. 21, 2013; revised Sept. 19, 2013; accepted Sept. 26, 2013.

Author contributions: H.G.E. and I.S. designed research; H.G.E. and I.S. performed research; H.G.E. and I.S. analyzed data; H.G.E. and I.S. wrote the paper.

This work was supported by the Consejo Nacional de Investigaciones Científicas y Técnicas, the Agencia de Promoción Científica y Tecnológica de Argentina, the Comisión Nacional de Energía Atómica, Universidad Nacional de Cuyo, and the Academy of Finland.

The authors declare no competing financial interests.

Correspondence should be addressed to Hugo Gabriel Eyherabide, Department of Computer Science, University of Helsinki, Gustaf Hällströmin katu 2b, 00560 Helsinki, Finland. E-mail: Hugo.Eyherabide@helsinki.fi.

DOI:10.1523/JNEUROSCI.0357-13.2013

Copyright © 2013 the authors 0270-6474/13/3317921-16\$15.00/0

process. Altogether, we provide a complete framework for assessing when and why information is lost by NI decoders.

Materials and Methods

Encoding, decoding, and neural information. The encoding process is the transduction of sensory stimuli S into responses $\mathbf{R} = [R_1, \dots, R_N]$ of a population of N neurons (R_n is the response of the n^{th} neuron). A priori, different stimuli S occur with probabilities $P(S)$. Responses \mathbf{R} are elicited with probabilities $P(\mathbf{R}|S)$. Posterior to the observation of \mathbf{R} , the stimulus probability becomes $P(S|\mathbf{R})$. When the probability distributions $P(S)$ and $P(S|\mathbf{R})$ are different, the population response \mathbf{R} contains information about S (i.e., \mathbf{R} may be used to infer S with higher precision than chance level). In units of bits, the mutual information $I(S;\mathbf{R})$ is quantified as

$$I(S; \mathbf{R}) = \underbrace{\mathbf{E}_{P(S)} [\log_2 P(S)]}_{H(S)} - \underbrace{\mathbf{E}_{P(S,\mathbf{R})} [\log_2 P(S|\mathbf{R})]}_{H(S|\mathbf{R})} \quad (1)$$

where $\mathbf{E}[X]$ represents the weighted mean of X with weights Y . The total entropy $H(S)$ and the noise entropy $H(S|\mathbf{R})$ quantify the average uncertainty of S prior and posterior to the observation of \mathbf{R} , respectively. The mutual information $I(S;\mathbf{R})$ represents the average reduction in the uncertainty of S due to the observation of \mathbf{R} .

The decoding process is the transformation of the population response \mathbf{R} into an estimation S^{Dec} of the stimulus S . For all decoders, the decoded information $I(S;S^{Dec})$ is upper bounded by the encoded information $I(S;\mathbf{R})$. This bound is a consequence of the data-processing inequality, which states that no transformation of the population response \mathbf{R} can increase the amount of information about the stimulus S (Cover and Thomas, 1991; Quian Quiroga and Panzeri, 2009). The coding theorem (Shannon, 1948; Cover and Thomas, 1991) ensures that this bound is tight and can be achieved by decoding extended sequences of stimuli and responses. However, biological constraints (e.g., fast behavioral responses) may severely restrict the length of the sequences, thus reducing the decoded information below the bound. The information lost by a given decoding algorithm is defined as

$$\Delta I_{Dec} = I(S;\mathbf{R}) - I(S;S^{Dec}) \geq 0. \quad (2)$$

When ΔI_{Dec} is greater than zero, some information (ΔI_{Dec}) about the stimulus S , encoded in the population response \mathbf{R} , is lost during the decoding process. In other words, the decoder has ignored some information (ΔI_{Dec}) that may have improved the stimulus estimation. If ΔI_{Dec} is zero, the decoding process is optimal; that is, it decodes all of the encoded information. Further discussion on the meaning of ΔI_{Dec} can be found in Eyherabide and Samengo (2010) and references therein.

The family of NI decoders. NI decoders are here defined as probabilistic decoders (i.e., decoders that infer the stimulus from the conditional probability distribution of the response) constructed under the NI assumption (the assumption that neurons are noise independent). Mathematically, the NI assumption states that the probability $P(\mathbf{R}|S_k)$ of the population response $\mathbf{R} = [R_1, \dots, R_N]$ (N is the number of neurons in the population) elicited by the stimulus S_k can be inferred from the probability $P(R_n|S_k)$ of each neuron in the population as follows:

$$P(\mathbf{R}|S_k) = P_{NI}(\mathbf{R}|S_k) = \prod_{n=1}^N P(R_n|S_k). \quad (3)$$

Here, $P_{NI}(\mathbf{R}|S_k)$ is called the NI likelihood. By multiplying the probabilities of individual neurons $P(R_n|S_k)$, NI decoders neglect all noise correlations among neurons. Once the NI assumption is made, several NI decoders can be constructed as follows:

$$S^{NI}(\mathbf{R}) = f^{NIL}[P_{NI}(\mathbf{R}|S_1), \dots, P_{NI}(\mathbf{R}|S_K)]. \quad (4)$$

using different algorithms f^{NIL} for extracting the decoded stimulus S^{NI} from the NI likelihoods (K is the number of stimuli). This construction is here called the canonical NI decoder.

The best-known construction of NI decoders is here called classical NI decoder, which is based on the NI posterior probabilities:

$$P_{NI}(S_k|\mathbf{R}) = \frac{P_{NI}(\mathbf{R}|S_k)P(S_k)}{\sum_k P_{NI}(\mathbf{R}|S_k)P(S_k)} \quad (5)$$

obtained from Equation 3 by applying Bayes' rule. Different classical NI decoders can be constructed by using different algorithms f^{NIP} for inferring the decoded stimulus S^{NI} from the NI posteriors as follows:

$$S^{NI}(\mathbf{R}) = f^{NIP}[P_{NI}(S_1|\mathbf{R}), \dots, P_{NI}(S_K|\mathbf{R})], \quad (6)$$

the most common choice being the maximum-posterior criterion (Nirenberg and Latham, 2003; Latham and Nirenberg, 2005; Oizumi et al., 2010):

$$S^{NI}(\mathbf{R}) = \arg \max_{S_k} \{P_{NI}(S_1|\mathbf{R}), \dots, P_{NI}(S_K|\mathbf{R})\}. \quad (7)$$

Although classical NI decoders are by far the most popular, they are just a subset of all canonical NI decoders; that is, they are a restricted choice of all possible probabilistic decoders based on the NI assumption. Classical NI decoders based on the maximum-posterior criterion have often been claimed to be optimal within the family of NI decoders. This is indeed true if neurons are truly noise independent. Otherwise, optimality is not guaranteed, as we show in this study.

Estimators of the minimum information loss induced by NI decoders. Previous studies have proposed to estimate the minimum information loss ΔI_{NI}^{Min} induced by NI decoders using different criteria, namely:

$$\Delta I_{NI}^{Min} = \Delta I_{NI} = I(S;\mathbf{R}) - I(S;S^{NI}) \quad (8a)$$

$$\Delta I_{NI}^{Min} = \Delta I_{NI}^{LS} = I(S;\mathbf{R}) - I(S;S_1^{NI}, \dots, S_K^{NI}) \quad (8b)$$

$$\Delta I_{NI}^{Min} = \Delta I_{NI}^D = \mathbf{D}[P(S|\mathbf{R}) || P_{NI}(S|\mathbf{R})] \quad (8c)$$

$$\Delta I_{NI}^{Min} = \Delta I_{NI}^{DL} = \min_{\theta} \Delta I_{NI}^{DL}(\theta). \quad (8d)$$

In criterion 8a (Quian Quiroga and Panzeri, 2009; Ince et al., 2010), the information loss ΔI_{NI} measures the difference between the encoded information and the information decoded by a specific implementation of the NI decoder chosen by the researcher. In criterion 8b (Ince et al., 2010), S_k^{NI} ($1 \leq k \leq K$) depends on the population response \mathbf{R} and represents the k^{th} most likely stimulus if neurons were noise independent. Therefore, ΔI_{NI}^{LS} measures the difference between the encoded information and the information extracted by a decoder that ranks the set of stimuli with respect to their NI posterior probability. In criterion 8c (Nirenberg and Latham, 2003; Latham and Nirenberg, 2005), \mathbf{D} represents the conditional Kullback-Leibler divergence and therefore, ΔI_{NI}^D measures the departure of $P_{NI}(S|\mathbf{R})$ from $P(S|\mathbf{R})$. In criterion 8d (Latham and Nirenberg, 2005; Oizumi et al., 2010):

$$\Delta I_{NI}^{DL}(\theta) = \mathbf{D}[P(S|\mathbf{R}) || \hat{P}_{NI}(S|\mathbf{R},\theta)] \quad (9a)$$

$$\hat{P}_{NI}(S|\mathbf{R},\theta) \propto P(S) \prod_{n=1}^N [P(R_n|S)]^{\theta}, \quad (9b)$$

where θ is a real number the value of which is chosen to minimize $\Delta I_{NI}^{DL}(\theta)$. The quantity ΔI_{NI}^{DL} measures the information loss induced by a classical NI decoder when operating on long sequences of responses. Both ΔI_{NI}^D (criterion 8c) and ΔI_{NI}^{DL} (criterion 8d) are intended to provide information-theoretical estimators that do not require building specific NI decoders.

Decoding error. The decoding error is here defined as the average cost of mistakenly estimating the stimulus S that elicited a population response \mathbf{R} , namely:

$$\xi_{Dec} = \mathbf{E}_{P(S,S^{Dec})} [\mathcal{L}(S, S^{Dec})], \quad (10)$$

where $\mathcal{L}(S, S^{Dec})$ is the non-negative cost of inferring S^{Dec} when the encoded stimulus was S (Duda et al., 2000; Bishop, 2006).

The minimum decoding error for all possible decoders based on a specific representation \mathbf{R} of the population response is given by the following:

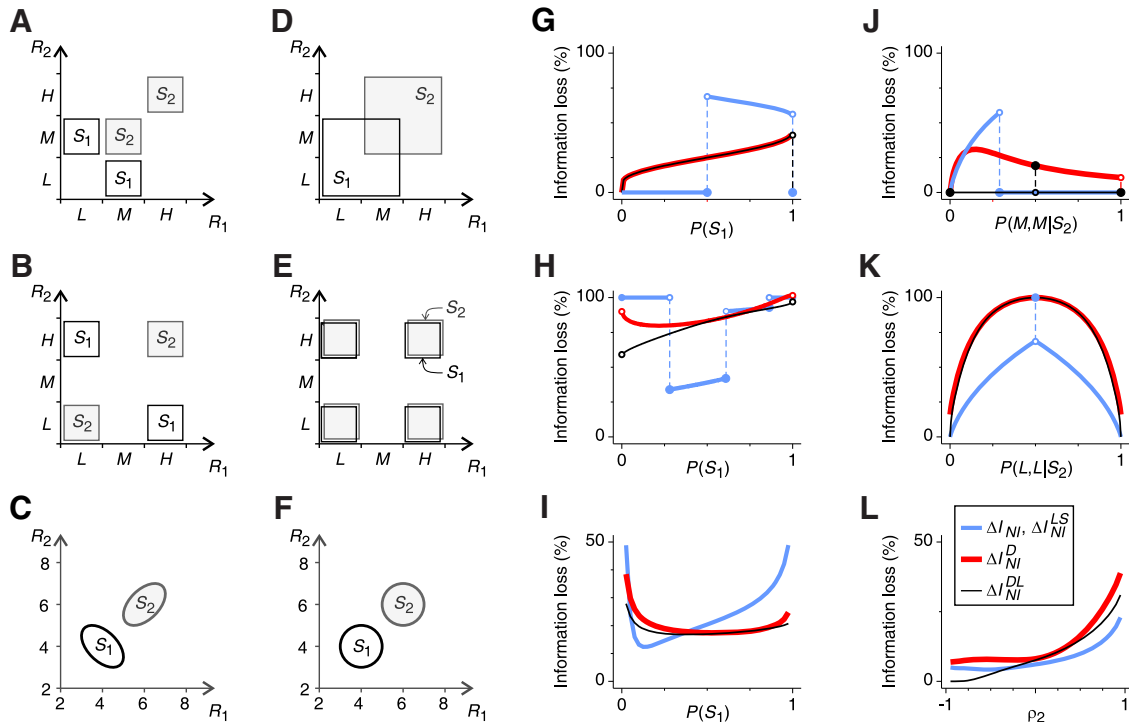


Figure 1. Previous estimations of ΔI_{NI}^{Min} are tighter or looser depending on the context. **A–C**, Examples of the simultaneous activity of two neurons elicited by two stimuli: S_1 (black) and S_2 (gray). **A, B**, Two population responses per stimulus. Responses to S_1 are negatively correlated and responses to S_2 are positively correlated. **C**, Populations responses have Gaussian distributions with mean values $\mu_1 = [4,4]$ and $\mu_2 = [6,6]$, variance equal to 1, and correlation coefficients ρ_1 and ρ_2 (for stimulus S_1 and S_2 , respectively). **D–F**, Surrogate NI population responses (see Materials and Methods). Because of the NI assumption, response distributions associated with different stimuli overlap. However, here we show that overlaps do not necessarily imply that information is lost (see text). For each example, ΔI_{NI}^{Min} was estimated using four estimators: ΔI_{NI} , ΔI_{NI}^{LS} , ΔI_{NI}^D , and ΔI_{NI}^{DL} (criteria 8a–8d). **G–L**, Variations of the estimations with: the stimulus probabilities (**G–I**), the response probabilities given S_2 (**J, K**), and the correlation coefficient ρ_2 (**L**). Only parameters specified in the x-axis are varied; the remaining parameters are constant. None of the estimators consistently lies below the others for all stimulus and response probabilities. Therefore, none of them constitutes a universal limit to the inefficiency of NI decoders and, depending on the context, they all overestimate, to a lesser or greater extent, the importance of noise correlations in neural decoding. Remaining parameters are as follows: in **G**, $P(M, L|S_1) = 0.5$ and $P(H, H|S_2) = 0.5$; in **H**, $P(H, L|S_1) = 0.8$ and $P(H, H|S_2) = 0.5$; in **I**, $\rho_1 = -0.9$ and $\rho_2 = 0.7$; in **J**, $P(M, L|S_1) = 0.5$ and $P(S_1) = 0.25$; in **K**, $P(H, L|S_1) = 0.5$ and $P(S_1) = 0.5$; in **L**, $P(S_1) = 0.2$ and $\rho_1 = -0.9$.

$$\xi^{Min}(\mathbf{R}, S) = \mathbf{E} \left[\min_{P(\mathbf{R})} \left\{ \mathbf{E} \left[\mathcal{L}(S, S^{Dec}) \right] \right\} \right] \quad (11)$$

where the minimization runs over all decoded stimuli (Bishop, 2006; Hastie et al., 2009). This minimum is achievable by a decoder defined as follows:

$$S^{Dec} = \arg \min_{\hat{s}} \left\{ \mathbf{E} \left[\mathcal{L}(S, \hat{s}) \right] \right\}. \quad (12)$$

The specific decoders that achieve the minimum decoding error depend on the shape of \mathcal{L} (Simoncelli, 2009). For example, the decoding-error probability (also known as fraction incorrect or error rate), can be obtained by setting $\mathcal{L}(S, S^{Dec})$ equal to zero if S and S^{Dec} coincide or to unity otherwise. The decoder that achieves the minimum decoding-error probability is given by the following:

$$S^{Dec} = \arg \min_s \{P(S|\mathbf{R})\}. \quad (13)$$

Decoders based on surrogate responses. Previous studies have also proposed to study the importance of noise correlations using decoders that were optimized for decoding the surrogate population activity that would be elicited if neurons were truly noise independent (Nirenberg et al., 2001; Berens et al., 2012). For each stimulus, this set of surrogate NI population responses $\{\mathbf{R}_{NI}\}$ is computed as the Cartesian product of the sets $\{R_n\}$, the elements of which are the individual responses of the n^{th} neuron:

$$\{\mathbf{R}_{NI}\} = \{R_1\} \times \dots \times \{R_N\}. \quad (14)$$

To train the decoder, surrogate NI population responses \mathbf{R}_{NI} are drawn from the set $\{\mathbf{R}_{NI}\}$ with probabilities given by Equation 3.

Results

Shortcomings of previous measures of information loss

The importance of noise correlations in neural decoding has been assessed by comparing the encoded information with the information extracted by NI decoders, which can be constructed in many different ways. The minimum difference between these two quantities is the minimum information loss ΔI_{NI}^{Min} induced by NI decoders. If ΔI_{NI}^{Min} is greater than zero, then noise correlations are important: By neglecting them, information is lost. The minimum information loss ΔI_{NI}^{Min} has been estimated in several ways. In this section, we compare the four most widely used estimators and show that they all tend to overestimate the importance of noise correlations in neural decoding. The results are illustrated with three examples in Figure 1. Previous studies have concluded that, in these examples, noise correlations are important for decoding. We demonstrate, however, that these conclusions are not valid in general: they may or may not hold depending on the stimulus and response probabilities and on the amount of correlation in the responses. In the first two examples (Fig. 1A, B), responses to stimulus S_1 are negatively correlated and responses to stimulus S_2 are positively correlated. In the third example (Fig. 1C), continuous responses are analyzed and the amount of correlation or anticorrelation of responses to stimulus S_2 is not fixed, but rather depends on the value of a given parameter.

Previous studies have estimated ΔI_{NI}^{Min} as the actual information loss ΔI_{NI} (criterion 8a; Fig. 1G–L, blue line) induced by a specific implementation of the NI decoder chosen by the researcher. The most common choice is here called the classical NI decoder (Eq. 7), which, for each population response, decodes the stimulus with the highest NI posterior probability $P_{NI}(S|\mathbf{R})$ (Wu et al., 2001; Nirenberg and Latham, 2003; Latham and Nirenberg, 2005; Ince et al., 2010). For example, in Figure 1A, the classical NI decoder always estimates the correct stimulus except for response $[R_1, R_2] = [M, M]$. Whenever

$$P_{NI}(S_2|M, M) < P_{NI}(S_1|M, M), \quad (15)$$

the classical NI decoder infers that the response $[M, M]$ was elicited by stimulus S_1 , whereas an optimal decoder constructed with knowledge of noise correlations always decodes the stimulus S_2 . Hence, ΔI_{NI} is greater than zero. However, if Equation 15 does not hold, the stimuli decoded by these two decoders always coincide and thus ΔI_{NI} is zero. When varying the stimulus and response probabilities in a continuous manner as in Figure 1, G and J, the transition between these two different situations is reflected as a discontinuity in the representation of ΔI_{NI} , resulting in a broken line. Whenever the classical NI decoder is optimal (i.e., ΔI_{NI} is zero), noise correlations are irrelevant for decoding. The converse, however, is not necessarily true. The classical NI decoder is only one among many ways of constructing a NI decoder. Other NI decoders may be more efficient or even optimal. In the latter case, noise correlations are irrelevant regardless of the value of ΔI_{NI} .

For example, a NI decoder can be constructed just like the classical NI decoder but with the stimulus prior probabilities $\hat{P}(S)$ differing from those $P(S)$ set in the experiment (Oram et al., 1998). It may be puzzling to see that such a NI decoder constructed with unrealistic prior probabilities may operate more efficiently than the classical NI decoder constructed with the real priors. Indeed, altering an optimal decoder constructed with the real probabilities $P(\mathbf{R}|S)$ cannot increase the information and might actually reduce it. However, there is no reason to believe that, when altering a suboptimal decoder constructed with unrealistic (NI) probabilities, the information may not increase. Of course, only carefully chosen alterations can do the job. In Figure 1G, a classical NI decoder with $\hat{P}(S_1)$ fixed at a value between 0 and 0.5 is capable of decoding the stimulus without error (i.e., Eq. 15 is never fulfilled) regardless of the true stimulus probabilities.

In an attempt to avoid the arbitrariness of the choice of a NI decoder (Averbeck et al., 2006; Quiñ Quiroga and Panzeri, 2009), Nirenberg et al. (2001) proposed measuring ΔI_{NI}^{Min} as the divergence ΔI_{NI}^D between the probability distributions of the stimulus given the response computed with and without the NI assumption (criterion 8c; Fig. 1G–L, red line). This method aims at estimating the information loss without decoding explicitly the population response, but unfortunately, it may severely overestimate ΔI_{NI}^{Min} . For the example shown in Figure 1A, ΔI_{NI}^D becomes:

$$\Delta I_{NI}^D = -P(M, M; S_2) \log_2 P_{NI}(S_2|M, M), \quad (16)$$

and thus ΔI_{NI}^D is always greater than zero, even though we showed that the classical NI decoder is optimal for a wide range of stimulus and response probabilities. The overestimation problem was first shown by Schneidman et al. (2003), who also showed that, strangely, ΔI_{NI}^D can even exceed the encoded information (Fig. 1H, top right). Indeed, if ΔI_{NI} is zero, then ΔI_{NI} is zero and noise correlations are unimportant, but the converse is not necessarily true. Last, Ince et al. (2010) showed that, in rat somatosensory

cortex, the size of the overestimation increases with the number of neurons in the population.

Oizumi et al. (2010) proposed to solve the overestimation problem by computing another quantity, here called ΔI_{NI}^{DL} (criterion 8d; Fig. 1G–L, black line), which measures the performance of a classical NI decoder when decoding long sequences of responses (Latham and Nirenberg, 2005). Unfortunately, ΔI_{NI}^{DL} also overestimates ΔI_{NI}^{Min} in a context-dependent manner. Consider the example shown in Figure 1A. The estimation of ΔI_{NI}^{DL} involves a minimization problem (Eq. 8d), leading to:

$$\Delta I_{NI}^{DL} = \begin{cases} 0 & \text{if } P_{NI}(M, M|S_1) \neq P_{NI}(M, M|S_2) \\ \Delta I_{NI}^D & \text{if } P_{NI}(M, M|S_1) = P_{NI}(M, M|S_2) \end{cases}. \quad (17)$$

By comparing Equations 17 and 15, we find that whenever $P_{NI}(M, M|S_1) = P_{NI}(M, M|S_2)$ and $P(S_2) > P(S_1)$, the classical NI decoder is optimal despite $\Delta I_{NI}^{DL} > 0$. The same occurs for the other examples shown in Figure 1, where ΔI_{NI}^{DL} may be larger or smaller than ΔI_{NI} depending on the stimulus and response probabilities. Therefore, ΔI_{NI}^{DL} does not constitute a limit to the inefficiency of NI decoders.

Recently, Ince et al. (2010) proposed another alternative, estimating ΔI_{NI}^{Min} as the information loss ΔI_{NI}^{LS} induced by a NI decoder that associates each response with a list of stimuli ordered according to how likely they would be if neurons were noise independent (criterion 8b; Fig. 1G–L, blue line). In Figure 1, ΔI_{NI}^{LS} coincides with ΔI_{NI} (the same occurs in any experiment involving two stimuli) and thus exhibits the same drawbacks. In general, ΔI_{NI}^{LS} represents a tighter bound than ΔI_{NI} (Ince et al., 2010). However, as discussed in the next sections, it still overestimates ΔI_{NI}^{Min} for reasons analogous to those of ΔI_{NI} .

The estimators mentioned above are based on probabilistic NI decoders, that is, decoders that infer the stimulus from the probability of the population responses computed with the NI assumption (Eq. 3). Previous studies have also proposed another alternative: to base the estimation of ΔI_{NI}^{Min} on decoders (generally linear) in which parameters are optimized for decoding surrogate NI population responses; that is, artificial responses that are generated under the NI assumption (Nirenberg et al., 2001; Averbeck and Lee, 2006; Quiñ Quiroga and Panzeri, 2009; Berens et al., 2012; see Materials and Methods and Fig. 1D–F). By comparing these two alternatives, however, we find that they may lead to opposite conclusions: noise correlations may turn out to be irrelevant in one of them and essential in the other (Fig. 2). Most importantly, using the second approach, one may conclude that noise correlations are essential even though neurons are actually noise independent (Fig. 2B). At first glance, one might think of this issue as an overestimation problem that, as in the case of ΔI_{NI} , could be avoided if one considered all possible decoding algorithms, not just the linear. Unfortunately, without constraining the type of decoding algorithms and optimization criteria, this estimator trivially underestimates ΔI_{NI}^{Min} and yields the conclusion that noise correlations are always irrelevant for neural decoding.

To prove this, notice first that the set of surrogate NI population responses $\{\mathbf{R}_{NI}\}$ can be constructed by adding to the set of population responses $\{\mathbf{R}\}$ all those responses that would occur only if neurons were noise independent. Therefore, any decoding algorithm that maps $\{\mathbf{R}_{NI}\}$ into the set of stimuli $\{S\}$ can be written as follows:

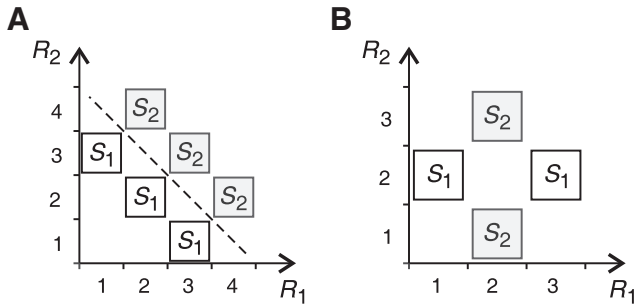


Figure 2. Comparison of different strategies to construct decoders that ignore noise correlations. Each panel shows the simultaneous responses of two neurons R_1 and R_2 elicited by two stimuli S_1 and S_2 . In both examples, stimuli and responses are equally likely. **A**, Linear decoders trained with surrogate NI population responses (dashed line) extract all the encoded information, whereas no probabilistic NI decoder can do so. Specifically, a probabilistic NI decoder is inefficient for a range of probabilities $P(R_1, R_2 | S_k)$ complying with the two following conditions: $P(2, 2 | S_1)^2 = P(1, 3 | S_1)P(3, 1 | S_1)$ and $P(3, 3 | S_2)^2 = P(2, 4 | S_2)P(4, 2 | S_2)$. **B**, Although neurons are noise independent, no linear decoder is capable of extracting all of the encoded information.

$$S^{Dec} = \begin{cases} f_1^{Dec}(\mathbf{R}_{NI}) & \text{if } \mathbf{R}_{NI} \in \{S\} \\ f_2^{Dec}(\mathbf{R}_{NI}) & \text{otherwise} \end{cases} \quad (18)$$

Among all possible mappings from $\{\mathbf{R}_{NI}\}$ to $\{S\}$, there always exists at least one for which f_1^{Dec} coincides with an optimal criterion to decode the population responses \mathbf{R} .

Previous studies have hypothesized that the NI assumption increases the number of real responses lying in the overlap between the surrogate NI responses associated with different stimuli (compare Fig. 1A–C with Fig. 1D–F), thereby introducing ambiguity in the NI decoding process and losing information. This conclusion, however, is not always true. The decoding rule may well evaluate the magnitude of each NI posterior probability (Eq. 6) and, with this information, always decode the same stimulus as a decoder constructed with knowledge of noise correlations (Bishop, 2006). For example, in Figure 1A, the population response $[R_1, R_2] = [M, M]$ is only elicited by stimulus S_2 . However, if neurons were noise independent, $[M, M]$ would be elicited by both S_1 and S_2 (Fig. 1D). Nevertheless, the classical NI decoder operates optimally for a wide range of stimulus and response probabilities (Fig. 1G, J). The presence of real responses in the overlap between the surrogate NI responses associated with each stimulus constitutes a necessary, but not a sufficient, condition for a NI decoder to be lossy.

In summary, we have shown that, without appropriate constraints, previous approaches using surrogate NI population responses may not be suitable for the analysis of the importance of noise correlations in neural decoding. Other approaches based on probabilistic NI decoders do not exhibit this problem because the construction of the NI decoder is purely based on the NI assumption. However, the estimators used in these approaches tend to overestimate the minimum inefficiency of probabilistic NI decoders in a context-dependent manner and none of them constitutes a universal bound. Unfortunately, the overestimation problem cannot be solved by simply taking the minimum estimation, though this strategy is better than relying on one estimator alone. In the next section, we show how to evaluate the exact value of the minimum information loss.

Exact measure of the minimum information loss

Previous estimators fail to tightly bound the minimum information loss of NI decoders. The reasons for the failure depend on the estimator, as shown in the previous section. In the case of ΔI_{NI} ,

the failure is due to the fact that the optimal decoder is not searched among all possible NI decoders but, at best, among subsets of limited size. However, the failure can be avoided by extending the search to all possible NI decoders. To that end, in this section, we first introduce the notion of canonical NI decoders: the set of all decoders that comply with the NI assumption (Eq. 3). Using a fundamental theorem in information theory, the coding theorem, we then determine exactly the amount of information lost by the best canonical NI decoder. This information loss is smaller than the bounds analyzed in the previous section (Fig. 1).

All probabilistic NI decoders, here called canonical NI decoders (Eq. 4), can be described as a 2-stage process (Fig. 3). Without loss of generality, consider the population response $\mathbf{R} = [R_1, \dots, R_N]$ (where N is the number of neurons) elicited by a stimulus S_k ($1 \leq k \leq K$, where K is the number of stimuli). In the first stage, the population response \mathbf{R} is internally represented as a vector \mathbf{R}^{NIL} of NI likelihoods (defined in Eq. 3), given by the following:

$$\mathbf{R}^{NIL} = [P_{NI}(\mathbf{R}|S_1), \dots, P_{NI}(\mathbf{R}|S_K)]. \quad (19)$$

This step, and only this step, embodies the NI assumption. The second stage represents the transformation of \mathbf{R}^{NIL} into the decoded stimulus S^{NI} and embodies the estimation criterion used to decode the stimulus.

As stated by the data-processing inequality (see Materials and Methods), each transformation in the sequence cannot increase the information about the stimulus and may potentially induce an information loss. The information lost in each stage of the decoding process can be determined using standard methods previously developed for the analysis of neural codes (Borst and Theunissen, 1999; Panzeri et al., 2007; Eyherabide and Samengo, 2010). In particular, the actual information loss ΔI_{NI} induced by canonical NI decoders can be separated as follows:

$$\Delta I_{NI} = \underbrace{I(S; \mathbf{R}) - I(S; \mathbf{R}^{NIL})}_{\Delta I_{NI}^{NIL}} + \overbrace{I(S; \mathbf{R}^{NIL}) - I(S; S^{NI})}^{\Delta I_{NI}^{Est}}. \quad (20)$$

Here, ΔI_{NI}^{NIL} is the information loss induced by the NI assumption (first stage) and ΔI_{NI}^{Est} is the information loss induced by the estimation process (second stage).

The NI assumption (first stage) is common to all NI decoders, and therefore ΔI_{NI}^{NIL} constitutes a lower bound to the information loss induced by all NI decoders. Mathematically, Equation 20 decomposes the actual information loss ΔI_{NI} into two non-negative terms, thereby proving that ΔI_{NI}^{NIL} is a lower bound of ΔI_{NI} . Nevertheless, ΔI_{NI}^{NIL} could still underestimate the minimum information loss induced by all NI decoders. To prove that ΔI_{NI}^{NIL} is tight, we need to prove that a NI decoder exists for which ΔI_{NI} coincides with ΔI_{NI}^{NIL} , as we do next.

Different estimation criteria induce different information losses ΔI_{NI}^{Est} . However, the coding theorem (Shannon, 1948; Cover and Thomas, 1991) demonstrates the existence of a decoding procedure that, operating on long sequences of responses, can make ΔI_{NI}^{Est} negligible, thus extracting all the information $I(S; \mathbf{R}^{NIL})$ that is preserved after the NI assumption. Because the NI assumption is common to all NI decoders, $I(S; \mathbf{R}^{NIL})$ constitutes the maximum amount of information that can be extracted by any canonical NI decoder, and the difference:

$$\Delta I_{NI}^{Min} = \Delta I_{NI}^{NIL} = I(S; \mathbf{R}) - I(S; \mathbf{R}^{NIL}) \geq 0, \quad (21)$$

constitutes the minimum information loss induced by any decoder embodying the NI assumption; that is, the single fundamental property needed to evaluate the relevance of correlations (Nirenberg and Latham, 2003).

We have invoked the coding theorem to demonstrate Equation 21. This theorem was demonstrated by Shannon (1948) and promoted the development of information theory as a full discipline. In the context of our work, the theorem applies to the mapping $S \rightarrow R \rightarrow R^{NIL}$, which can be abbreviated as $S \rightarrow R^{NIL}$. This mapping, in the notation of Shannon, constitutes a channel that transforms each S into R^{NIL} . Repeated uses of the channel transform sequences of Q stimuli $[S_1, \dots, S_Q]$ into sequences $[R_1^{NIL}, \dots, R_Q^{NIL}]$. Shannon’s proof involved actual decoders that mapped sequences of Q responses R^{NIL} into sequences of Q -decoded stimuli S^{NI} . By making Q sufficiently large, he showed that there is at least one decoder for which the information transmission rate $I(S, S^{NI})$ can be made as close as desired to $I(S, R^{NIL})$ with negligible decoding error, thus yielding Equation 21. Two things, however, should be noticed. First, Shannon did not display his decoder explicitly—and neither do we. He simply demonstrated its existence. Second, in order for the theorem to hold, long sequences of stimuli and responses may be required. This requirement may be undesirable when studying the neural code under behavioral contexts, so later we introduce the minimum decoding error to circumvent this inconvenience.

To illustrate how Equations 20 and 21 improve the analysis of the role of noise correlations in neural decoding, consider the examples shown in Figure 4. Each panel shows how the population response is transformed throughout the decoding process by the canonical NI decoder. These examples were analyzed in Figure 1, in which we showed that, for a wide range of stimulus and response probabilities, all previous estimators indicated that noise correlations are important. However, as we show next, these estimators include not only the information loss induced by the NI assumption, but also the information loss induced by underlying assumptions constraining the estimation criteria, thus overestimating the importance of noise correlations.

In the example shown in Figure 4A (previously analyzed in Fig. 1A), ΔI_{NI}^{NIL} is zero regardless of the values of the stimulus and response probabilities. Indeed, after the first stage (where the NI assumption takes place) population responses elicited by different stimuli remain different (i.e., they are associated with different R^{NIL} ; Fig. 4A, middle), as we formally show after the next section. Therefore, the losses reported by previous estimators necessarily occur during the estimation stage. However, among all mappings between the representation R^{NIL} and the stimulus, there is at least one capable of correctly estimating the stimulus. Therefore, all the encoded information can be extracted without any loss and noise correlations are irrelevant for decoding. Notice, nevertheless, that finding an optimal estimation criterion explicitly is unnecessary. As we showed above, the minimum

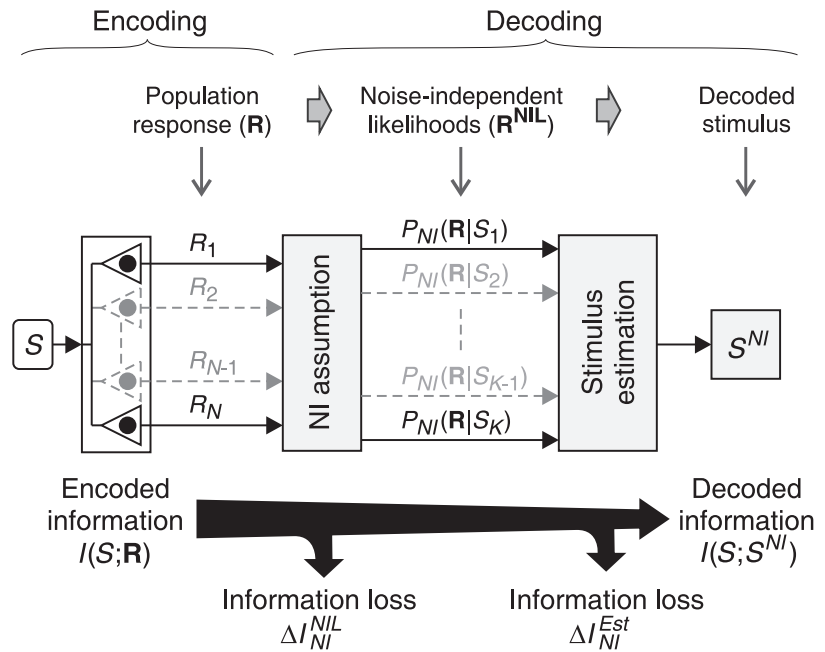


Figure 3. The canonical NI decoder is modeled as a sequence of transformations of the population response. The first stage involves the NI assumption, transforming the population response R into a vector R^{NIL} of NI likelihoods. The second stage involves the stimulus estimation, transforming R^{NIL} into the decoded stimulus S^{NI} . At each stage, information may be lost.

inefficiency of NI decoders, and the importance of noise correlations in neural decoding, can both be assessed by using ΔI_{NI}^{NIL} even before considering any estimation criterion.

Of course, information may be lost before the stimulus estimation takes place due to the NI assumption. Consider the examples shown in Figure 4, B and C (previously analyzed in Fig. 1B). When both $P(H, L|S_1)$ and $P(H, H|S_2)$ are set to 0.5 (Fig. 4B), ΔI_{NI}^{NIL} is equal to the encoded information. Noise correlations are thus crucial for decoding. Indeed, after the NI assumption, all population responses become indistinguishable (Fig. 4B, middle) and no estimation criterion is capable of extracting any information about the stimulus (Gawne and Richmond, 1993). These carefully chosen response probabilities, however, constitute an isolated case. For other values of the response probabilities (Fig. 4C), all population responses are represented differently after the NI assumption (as we formally show after the next section), and therefore ΔI_{NI}^{NIL} is zero. Therefore, except for the isolated case of Figure 4B, noise correlations are irrelevant for decoding.

When and why information is lost throughout the decoding process

In the previous section, we first modeled the NI decoder as a sequence of transformations of the population response (Fig. 3) and then quantified the average information loss induced by each stage of the NI-decoding process (Eq. 20) over all population responses (see Materials and Methods). However, information losses need not be evenly distributed among responses. In this section, we determine which are the specific responses that induce losses and the amount and type of information that is lost. We demonstrate that losses may only appear in those responses in which the decoding mapping is not injective. To localize the loss, we analyze how each successive transformation in the decoding process merges distinct representations of two or more responses into a single one so that their distinction is lost. The approach is similar to previous studies of the neural code (Eyherabide and

Samengo, 2010, and references therein), revealing what sort of information about the stimulus is preserved or lost and which response features encode such information.

After each transformation in the decoding process, two or more population responses whose distinction is informative may be represented in identical manner so their distinction is no longer available for subsequent transformations. Whenever that happens, information is lost. To assess whether the distinction between two (or more) population responses \mathbf{R}_A and \mathbf{R}_B is informative or constitutes noise, we first construct a representation $\tilde{\mathbf{R}}$ that treats those responses as if they were the same, but keeps the distinction between all other responses and then we compare the encoded information with and without the distinction (Eyherabide and Samengo, 2010)

$$\Delta I_{\mathbf{R} \rightarrow \tilde{\mathbf{R}}} = I(\mathbf{R}; S) - I(\tilde{\mathbf{R}}; S) \geq 0. \quad (22)$$

Whenever $\Delta I_{\mathbf{R} \rightarrow \tilde{\mathbf{R}}}$ is zero, the distinction between \mathbf{R}_A and \mathbf{R}_B provides no additional information and only constitutes noise. For example, in Figure 4B, this is the case for population responses $[R_1, R_2] = [L, H]$ and $[H, L]$ or responses $[L, L]$ and $[H, H]$. Notice that, when responses vary in a continuum, single responses are typically associated with a probability density. In that case, a representation $\tilde{\mathbf{R}}$ that treats two single responses \mathbf{R}_A and \mathbf{R}_B as equivalent induces no information loss. In the continuous case, an information loss occurs only when $\tilde{\mathbf{R}}$ treats a set of population responses as equivalent and, in addition, the probability of the set is nonzero.

The condition for the distinction between responses to be informative (Eq. 22) can also be written as a direct comparison between the real posterior probabilities $P(S|\mathbf{R})$. The distinction between two (or more) population responses \mathbf{R}_A and \mathbf{R}_B constitutes noise if and only if:

$$P(S_k|\mathbf{R}_A) = P(S_k|\mathbf{R}_B), \quad (23)$$

for all stimuli S_k (this comes directly from Eq. 22). Otherwise, their distinction is informative, and ignoring it induces an information loss $\Delta I_{\mathbf{R} \rightarrow \tilde{\mathbf{R}}}$. Examples of informative variations and noise when population responses are represented as continuous variables are given in the last section of Results.

We can now formally state when and why noise correlations are important in neural decoding. During the decoding process, the population response \mathbf{R} is first transformed into \mathbf{R}^{NIL} immediately after the NI assumption takes place (Fig. 3). This transformation induces an information loss when (and only when) the three following conditions are fulfilled:

- (24a) The mapping $\mathbf{R} \rightarrow \mathbf{R}^{\text{NIL}}$ is not injective, and, as a consequence, there are two or more responses $\{\mathbf{R}_1, \dots, \mathbf{R}_Q\}$ mapped onto the same \mathbf{R}^{NIL} ,
- (24b) Equation 23 is not fulfilled for at least two responses complying with condition 24a and one stimulus S_k , and
- (24c) The probability of the set of population responses fulfilling both conditions 24a and 24b is nonzero.

Noise correlations are important for decoding specifically those responses satisfying these three conditions: the cause of the loss (the “why”) relies on the fact that the NI assumption no longer allows the decoder to take into account the differences in their information content. The losses can be linked to specific stimulus and response features by interpreting \mathbf{R}^{NIL} as a reduced representation of the population response (Eyherabide and Samengo, 2010). In such a paradigm, population responses are

represented as a vector of response features, each conveying information about specific stimulus features. Only some of those response features (and their information content) are preserved after the NI assumption (first stage). The analysis of the preserved and lost features allows one to determine the effect of the NI assumption on the neural code.

With conditions 24a, 24b, and 24c in mind, we can provide an approximate estimation of the likelihood that correlations be relevant. For discrete responses, noise correlations are often irrelevant because condition 24a is often violated. This condition requires that at least two different responses, \mathbf{R}_A and \mathbf{R}_B , be mapped onto the same vector, \mathbf{R}^{NIL} . This is unlikely, though, because the mapping from \mathbf{R} to \mathbf{R}^{NIL} goes from a discrete set to a continuous space (Eq. 19). For continuous responses, the mapping from \mathbf{R} to \mathbf{R}^{NIL} goes from a continuous space of dimension N (where N is the number of neurons) to a continuous space of dimension K (where K is the number of stimuli). Intuitively, one would therefore expect that correlations would be relevant more often in the case of continuous responses than in the case of discrete responses and that the importance of correlations should tend to increase with N and decrease with K .

This approximate estimation, however, should not be taken as a hard rule. One can construct an infinite number of counterexamples in which noise correlations are crucial for discrete responses (Figs. 2, 4) or in which the importance of noise correlations, for discrete and continuous responses, decreases with N and increases with K , exactly opposite to the approximate estimation mentioned above. Nevertheless, one can prove that, for examples with a finite number of discrete responses, these counterexamples, though infinite in number, constitute a set of measure zero in the space of all possible examples, at least when using a counting measure (i.e., a measure that simply counts the number of cases regardless of how likely they occur in nature; Tao, 2011). Unfortunately, for examples with continuous responses or with an infinite number of discrete responses, such proof remains elusive.

Furthermore, one should observe that, in experimental conditions, neither responses nor response probabilities can be measured with infinite precision. Experimental errors and limited sampling may both produce broad distributions of responses, each associated with distributions of response probabilities. Mathematically, each population response \mathbf{R} is associated not with a probability $P(\mathbf{R}|S)$, as would occur if probabilities could be estimated with infinite precision, but with a distribution of probabilities $Q[P(\mathbf{R}|S)]$ for each stimulus S . The distribution $Q[P(\mathbf{R}|S)]$ can be estimated either using Bayesian approaches or resampling methods (Bishop, 2006; Hastie et al., 2009). In other words, due to experimental errors and limited sampling, $P(\mathbf{R}|S)$ becomes a random variable with probability $Q[P(\mathbf{R}|S)]$. Furthermore, the mapping from \mathbf{R} to \mathbf{R}^{NIL} (Eq. 19) becomes a probabilistic one-to-many mapping as opposed to the deterministic one-to-one mapping that would be obtained if response probabilities were measured with infinite precision. The change in the nature of the mapping $\mathbf{R} \rightarrow \mathbf{R}^{\text{NIL}}$ should be taken into account when evaluating $\Delta I_{\text{NI}}^{\text{NIL}}$, resulting in a distribution of information losses rather than in a unique deterministic value. Moreover, the equalities in conditions 24a and 24b should be interpreted in statistical terms (i.e., equalities should be assessed through hypothesis testing). In these circumstances, NI decoders become lossy more frequently than predicted by the approximate prediction above. Therefore, the relevance of correlations and the cer-

tainty with which such relevance is assessed depend on the quality of the measured data.

Finally, notice that, in many applications, discrete responses arise as quantizations of continuous responses. It would be desirable to recover, for increasingly small bins, the results obtained with the original continuous responses. The typical procedure is to assign a single probability to each bin as, for example, the mean value of the original continuous probability distribution inside the bin. This approach leads to a purely discrete model that represents poorly the importance of noise correlations in the underlying continuous responses. To solve this problem, a different quantization procedure should be used. One possibility consists in associating each bin R^{Bin} with a probability distribution $P(R^{NII}|R^{Bin})$, representing the spread of the conditional response probabilities of the continuous case and thereby including some uncertainty in the value of R^{NII} . The probabilistic mapping between R^{Bin} and R^{NII} connects two continuous spaces and the results obtained with such quantization procedure becomes consistent with those obtained with the original continuous responses.

Impact of the choice of NI decoder

Within the family of canonical NI decoders (Fig. 3), specific NI decoders differ in the estimation criteria used to decode the stimulus. For example, in classical NI decoders, the estimation process involves the calculation of the probability of each stimulus given the population response (using Bayes’ rule) as if neurons were noise independent, and then the selection of the most likely stimulus. In this section, we show in detail why this estimation strategy need not be optimal when applied after the NI assumption.

Classical NI decoders can be modeled as a three-stage process (Fig. 5A). In the first stage, the population response R is transformed into a vector of NI likelihoods R^{NII} (Eq. 19). In the second stage, R^{NII} is further transformed into a vector of NI posterior probabilities as follows:

$$R^{NIP} = [P_{NI}(S_1|R), \dots, P_{NI}(S_k|R)], \tag{25}$$

through Bayes’ rule (Eq. 5). The final stage is the estimation of the stimulus from R^{NIP} through the maximum-posterior criterion (Wu et al., 2001; Nirenberg and Latham, 2003; Latham and Nirenberg, 2005; Ince et al., 2010; Eq. 7).

In this model, Bayes’ rule acts as a deterministic mapping that transforms the vector of NI likelihoods R^{NII} into another vector of NI posteriors, R^{NIP} . This mapping can be injective or not. If it is injective, then Bayes’ rule is obviously lossless. Otherwise, it may cause an information loss, as we show below

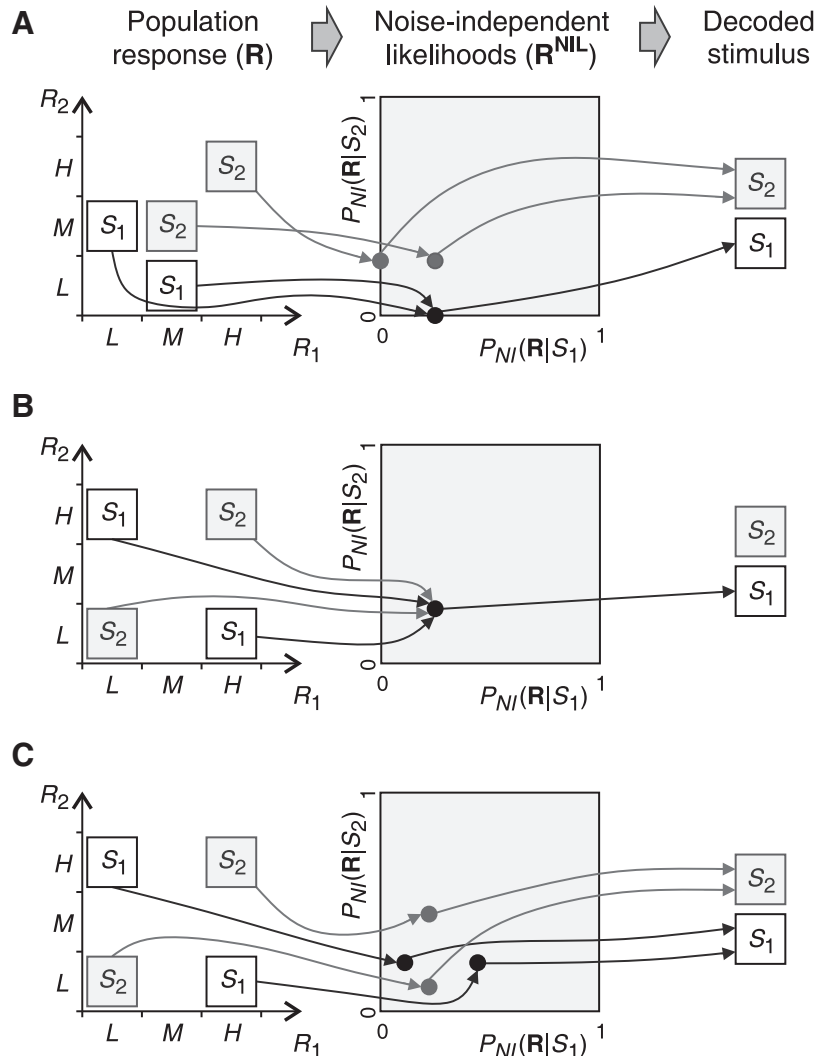


Figure 4. Examples of population activities decoded using the canonical NI decoder. The arrows show the transformation of population responses into vectors of NI likelihoods (R^{NII}) induced by the NI assumption (left to middle) and optimal estimation algorithms (middle to right). In **A**, $P(S_1)$ is set to 0.75, and $P(M, L|S_1)$ and $P(H, H|S_2)$ are both set to 0.5. In **B** and **C**, stimuli are equally likely and $P(H, L|S_1)$ and $P(H, H|S_2)$ are both set to 0.5 in **B** and to 0.66 in **C**. After the NI assumption, the distinction between responses elicited by different stimuli is preserved (middle panel). Therefore, ΔI_{NI}^{NII} is zero and noise correlations are unimportant for decoding. **B**, After the NI assumption, all responses are identical. No information about the stimulus remains and noise correlations are crucial for decoding. **C**, However, whenever population responses are not equally likely given each stimulus, the NI assumption preserves all the encoded information and noise correlations are unimportant for decoding. The case shown in **B**, in which noise correlations are important, therefore constitutes an isolated example.

when discussing the example of Figure 5C. Notice, however that, when using the real stimulus-response probabilities describing the data, Bayes’ rule is always lossless because the responses that are confounded constitute noise, as shown in the previous section.

The actual information loss ΔI_{NI} induced by the classical NI decoder can be separated as follows:

$$\Delta I_{NI} = \underbrace{\Delta I_{NI}^{NII}}_{\Delta I_{NI}^{NIP}} + \underbrace{\Delta I_{NI}^{NIB}}_{\Delta I_{NI}^{NIP}} + \underbrace{\Delta I_{NI}^{Est}}_{\Delta I_{NI}^{Est}} \tag{26}$$

This equation shows that the actual information loss induced by the classical NI decoder contains three different contributions:

(1) the information loss ΔI_{NI}^{NII} induced by the NI assumption (Eq. 21);

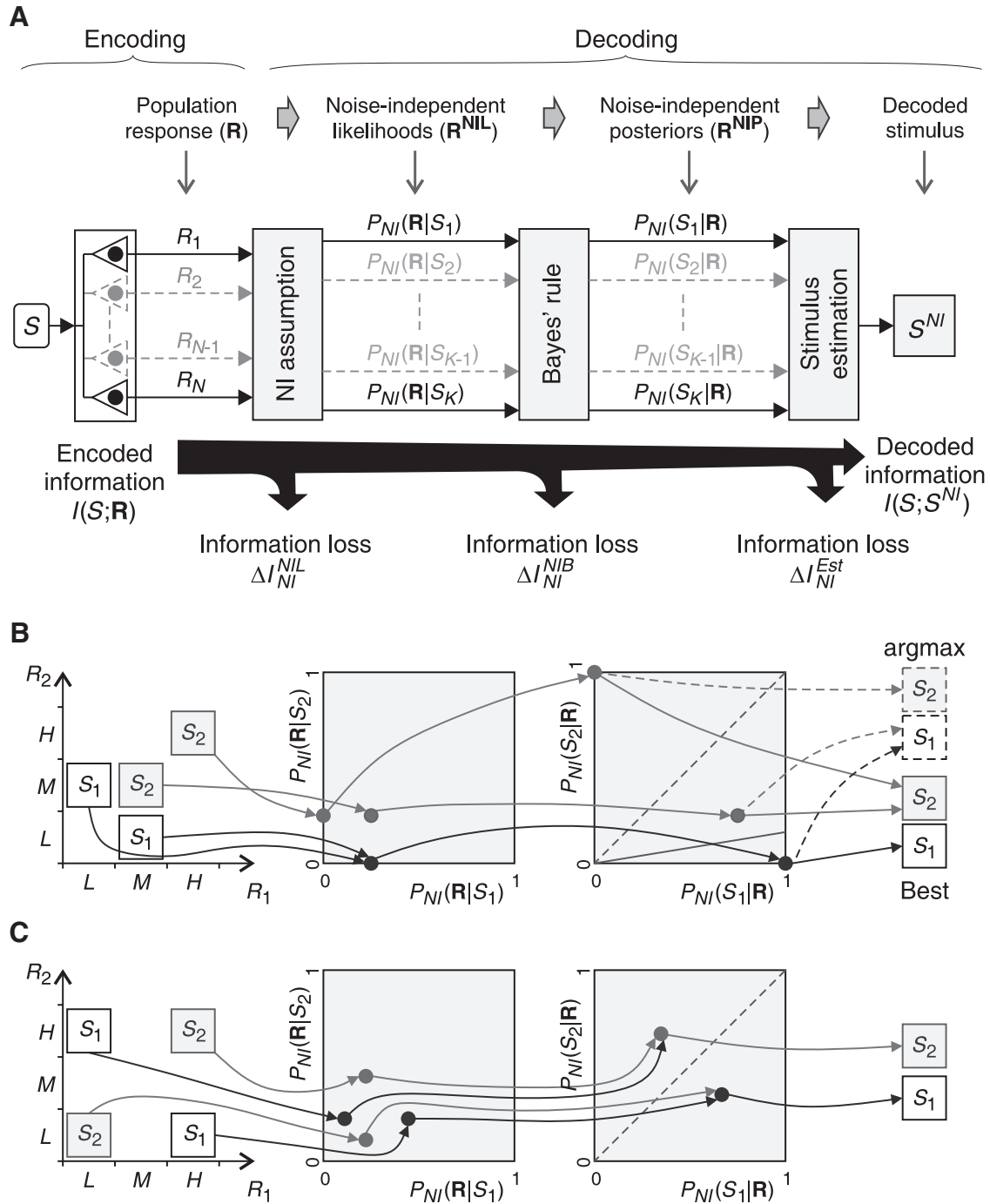


Figure 5. The impact of the choice of a NI decoder. **A**, The classical NI decoder is modeled as a three-stage process involving: the NI assumption (first stage), in which the population response is transformed into a vector of NI likelihoods (\mathbf{R}^{NIL}); Bayes' rule (second stage), in which \mathbf{R}^{NIL} is transformed into a vector of NI posteriors (\mathbf{R}^{NIP}); and the estimation criterion (third stage), in which the decoded stimulus is inferred from \mathbf{R}^{NIP} . Each stage may induce an information loss. **B**, **C**, Population responses of Figure 4, **A** and **C**, decoded with the classical NI decoder. **B**, Both \mathbf{R}^{NIL} and \mathbf{R}^{NIP} keep responses elicited by different stimuli segregated and therefore all information is preserved. However, the stimulus cannot always be correctly inferred by simply choosing the one corresponding to the maximum NI posterior (argmax criterion, dotted line, and arrows; right). Nevertheless, there is an estimation criterion capable of correctly estimating the stimulus (continuous lines and arrows; right). **C**, Although the NI assumption preserves all the encoded information, after Bayes' rule, responses associated with different stimuli are merged, and thus some (but not all) information is lost (ΔI_{NI}^{NIP} is greater than zero). As a result, no estimation criterion is capable of perfectly decoding the stimulus. However, other NI decoders may still be optimal for decoding (Fig. 4C).

- (2) the information loss ΔI_{NI}^{NIP} induced by Bayes' rule; and
- (3) the information loss ΔI_{NI}^{Est} induced by the chosen stimulus-estimation criterion (in this case, the maximum posterior).

To understand how choosing the NI decoder affects the decoded information, consider the example shown in Figure 5B

(previously analyzed in Fig. 4A). $P(S_1)$ is set to 0.75 and both $P(M, L|S_1)$ and $P(H, H|S_2)$ are set to 0.5. Throughout the NI-decoding process, the population responses $\mathbf{R} = [R_1, R_2]$ are first transformed through Equation 19 into the representations $\mathbf{R}^{NIL} = [P_{NI}(\mathbf{R}|S_1), P_{NI}(\mathbf{R}|S_2)]$ and then through Equation 25 into $\mathbf{R}^{NIP} = [P_{NI}(S_1|\mathbf{R}), P_{NI}(S_2|\mathbf{R})]$, resulting in:

$$\begin{array}{lcl}
 \mathbf{R} & \rightarrow \mathbf{R}^{\text{NIL}} & \rightarrow \mathbf{R}^{\text{NIP}} \\
 [L, M] & \rightarrow [0.25, 0] & \rightarrow [1, 0] \\
 [M, L] & \rightarrow [0.25, 0] & \rightarrow [1, 0] \\
 [M, M] & \rightarrow [0.25, 0.25] & \rightarrow [0.75, 0.25] \\
 [H, H] & \rightarrow [0, 0.25] & \rightarrow [0, 1]
 \end{array}$$

These transformations are also shown in Figure 5B. The first stage only merges the population responses $[L, M]$ and $[M, L]$, but their distinction only constitutes noise ($\Delta I_{R \rightarrow R^{\text{NIL}}}$ is zero; Eq. 22), and thus no information is lost (i.e., $\Delta I_{NI}^{\text{NIL}}$ is zero). The second stage is an injective mapping so it also does not affect the decoded information ($\Delta I_{NI}^{\text{NIP}}$ is zero).

Using Equations 19 and 25 in an analogous manner, we can generalize the previous results to arbitrary values of the stimulus and response probabilities: Responses associated with different stimuli are always represented in a different manner, both after the first stage and after the second stage. As a result, any information loss (Fig. 1G,J) is due to the estimation criterion. Nevertheless, among all possible estimation criteria, there is at least one capable of extracting all the information remaining in \mathbf{R}^{NIP} , which is equal to the encoded information. This optimal estimation criterion, however, differs from the maximum-posterior criterion (Fig. 5B, right).

Another example is analyzed in Figure 5C (previously analyzed in Fig. 4C), in which losses are distributed differently throughout the decoding process. Stimuli are equally likely and both $P(H, L|S_1)$ and $P(H, H|S_2)$ are set to 0.66. Throughout the NI decoding process, the population responses $\mathbf{R} = [R_1, R_2]$ are first transformed (recall Eqs. 19 and 25) as follows:

$$\begin{array}{lcl}
 \mathbf{R} & \rightarrow \mathbf{R}^{\text{NIL}} & \rightarrow \mathbf{R}^{\text{NIP}} \\
 [L, L] & \rightarrow [0.22, 0.11] & \rightarrow [0.66, 0.33] \\
 [L, H] & \rightarrow [0.11, 0.22] & \rightarrow [0.33, 0.66] \\
 [H, L] & \rightarrow [0.44, 0.22] & \rightarrow [0.66, 0.33] \\
 [H, H] & \rightarrow [0.22, 0.44] & \rightarrow [0.33, 0.66]
 \end{array}$$

These transformations are also shown in Figure 5C. The first transformation merges no population responses. Therefore, the distinction between population responses is preserved after the NI assumption and $\Delta I_{NI}^{\text{NIL}}$ is zero. The second transformation, however, merges response $[L, H]$ with $[H, H]$ and $[H, L]$ with $[L, L]$. Unlike \mathbf{R}^{NIL} , the representation \mathbf{R}^{NIP} carries less information about the stimulus than the encoded information and $\Delta I_{NI}^{\text{NIP}}$ is greater than zero (Fig. 6A). Therefore, although there exists a canonical NI decoder capable of decoding without error (Fig. 4C), classical NI decoders are unable to extract all of the information preserved after the NI assumption.

For other values of stimulus and response probabilities, almost always a canonical NI decoder exists capable of decoding without error (Fig. 4C), except in the isolated case shown in Figure 4B. However, classical NI decoders may still be incapable of extracting all the information preserved after the NI assumption. Unlike \mathbf{R}^{NIL} , population responses associated with different stimuli are not always mapped after Bayes' rule onto different \mathbf{R}^{NIP} . There are two cases in which responses are merged: (1) when $P(L, H|S_1) = P(H, H|S_2)$, in which case response $[L, H]$ is merged with $[L, L]$ and $[H, L]$ with $[H, H]$; and (2) when $P(L, H|S_1) = P(L, L|S_2)$, in which case response $[L, H]$ is merged with $[H, H]$ and $[H, L]$ with $[L, L]$ (this case is shown in Fig. 5C). Therefore, the representation \mathbf{R}^{NIP} carries less information about the stimulus than the encoded information and $\Delta I_{NI}^{\text{NIP}}$ is greater than zero (Fig. 6A). These cases constitute examples in which NI decoders can be optimal, but for achieving optimality, the esti-

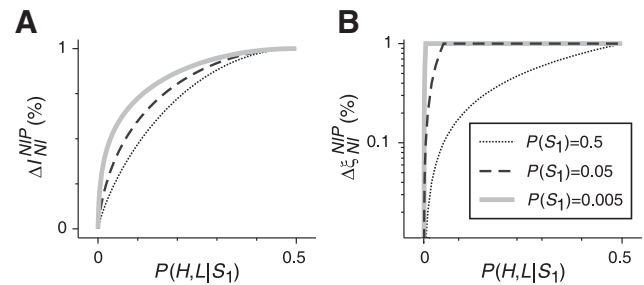


Figure 6. Difference between assessing the role of noise correlations using mutual information or decoding error. The population response shown in Figure 5C is decoded using the classical NI decoder. Response probabilities are set according to $P(H, L|S_1) = P(L, L|S_2)$. For different stimulus probabilities $P(S_1)$, **A** shows the variation of the minimum information loss $\Delta I_{NI}^{\text{NIP}}$ relative to the encoded information $I(\mathbf{R}; S)$, and **B** shows the variation of the increment in the minimum decoding error $\Delta \xi_{NI}^{\text{NIP}}$ relative to the minimum decoding error $\xi^{\text{Min}}(\mathbf{R}; S)$. The decoding error is here measured as decoding-error probability. The curves for $P(S_1) = p$ are identical to the curves for $P(S_1) = 1 - p$ ($0 \leq p \leq 1$). **A**, Unlike the case shown in Figure 4B, here, information is only partially lost and the loss depends on the stimulus and response probabilities. The maximum loss, however, only occurs when $P(H, L|S_1)$ reaches 0.5 regardless of the stimulus probability. **B**, Unlike $\Delta I_{NI}^{\text{NIP}}$, $\Delta \xi_{NI}^{\text{NIP}}$ approaches its maximum value when $P(H, L|S_1)$ is greater or equal to $P(S_1)$.

mation must be based purely on the NI assumption (i.e., on \mathbf{R}^{NIL}).

Using Equations 21, 23, and 26 we can now prove why estimators based on ΔI_{NI} (criterion 8a) and $\Delta I_{NI}^{\text{LS}}$ (criterion 8b) overestimate the minimum information loss $\Delta I_{NI}^{\text{Min}}$. Both methods have two occasions to include unnecessary losses. The first occasion appears when transforming \mathbf{R}^{NIL} into \mathbf{R}^{NIP} (Bayes' rule). When neurons are truly noise independent, $\Delta I_{NI}^{\text{NIL}}$ and $\Delta I_{NI}^{\text{NIP}}$ coincide; otherwise, as a result of Bayes' rule, some responses for which a distinction is informative may be merged (Fig. 5C), and therefore $\Delta I_{NI}^{\text{NIP}} > \Delta I_{NI}^{\text{NIL}}$. The second occasion appears when passing from \mathbf{R}^{NIP} to the decoded stimulus S^{NI} . In the case of ΔI_{NI} , the estimation criterion usually coincides with the maximum posterior, which, as shown above, may be suboptimal when used under the NI assumption. In the case of $\Delta I_{NI}^{\text{LS}}$, \mathbf{R}^{NIP} is transformed into a ranking of stimuli. This stage, although more finely grained than the purely maximum-posterior criterion, may still lump distinct representations \mathbf{R}^{NIP} into one single ranking and thereby perhaps lose information.

Minimum decoding error

The analysis of the effects of ignoring noise correlations in neural decoding was here performed using mutual information (Cover and Thomas, 1991; Borst and Theunissen, 1999). This quantity sets a limit to the decoding performance (e.g., in the number of stimulus categories that can be distinguished with negligible error probability), but this limit may only be achievable when decoding long sequences of population responses (i.e., comprising several consecutive population responses, also known as block coding; Cover and Thomas, 1991). Long sequences of responses inevitably have a long duration. To produce timely behavioral reactions, however, neural systems must process information in short time windows (tens or hundreds of milliseconds) (Hari and Kujala, 2009; Panzeri et al., 2010). Long sequences of responses may therefore be inconsistent with the fast behavioral responses observed in nature. In addition, mutual information may not adequately represent the cost of wrongly estimating the stimulus (Nirenberg and Latham, 2003). To overcome these issues,

in this section, we also bound the inefficiency of NI decoders using the minimum decoding error.

The minimum decoding error $\xi^{Min}(S; \mathbf{R})$ (Eq. 11) can be defined using different cost functions (Simoncelli, 2009), allowing one to assess the importance of noise correlations when decoding is performed on a single-response basis. Like mutual information, it is non-negative and depends on the representation \mathbf{R} of the population response. Furthermore, $\xi^{Min}(S; \mathbf{R})$ also follows the data-processing inequality (Cover and Thomas, 1991; Quiñero and Panzeri, 2009), but it actually increases with transformations of \mathbf{R} . Let $\tilde{\mathbf{R}} = g(\mathbf{R})$ be one of such transformations (deterministic or stochastic); then:

$$\xi^{Min}(\mathbf{R}; S) \leq \xi^{Min}(\tilde{\mathbf{R}}; S). \quad (27)$$

To prove this, recall that $\mathbb{E}[X]$ represents the weighted mean of X with weights Y . We derive Equation 27 as follows:

$$\xi^{Min}(\mathbf{R}, S) = \mathbf{E}_{P(\mathbf{R})} \left[\min_{S^{Dec}} \left\{ \mathbf{E}_{P(S|\mathbf{R})} [\mathcal{L}(S, S^{Dec})] \right\} \right] \quad (28a)$$

$$= \mathbf{E}_{P(\tilde{\mathbf{R}})} \left[\mathbf{E}_{P(\mathbf{R}|\tilde{\mathbf{R}})} \left[\min_{S^{Dec}} \left\{ \mathbf{E}_{P(S|\mathbf{R})} [\mathcal{L}(S, S^{Dec})] \right\} \right] \right] \quad (28b)$$

$$\leq \mathbf{E}_{P(\tilde{\mathbf{R}})} \left[\min_{S^{Dec}} \left\{ \mathbf{E}_{P(\mathbf{R}|\tilde{\mathbf{R}})} \left[\mathbf{E}_{P(S|\mathbf{R})} [\mathcal{L}(S, S^{Dec})] \right] \right\} \right] \quad (28c)$$

$$= \xi^{Min}(\tilde{\mathbf{R}}; S). \quad (28d)$$

Because of these similarities, the mathematical framework and the interpretations obtained from the minimum decoding error are almost identical to those of mutual information, taking care of the change in the sign of the data-processing inequality (as shown below). However, when applied to experimental data, the results obtained using mutual information or minimum decoding error may differ both quantitatively and qualitatively depending on the case under study (as shown in the next section).

The increment in the decoding error $\Delta\xi_{NI}$ induced by a canonical NI decoder (Fig. 3) with respect to the minimum decoding error $\xi^{Min}(\mathbf{R}; S)$ (that would be achievable if noise correlations were taken into account) is given as follows:

$$\Delta\xi_{NI} = \xi_{NI} - \xi^{Min}(\mathbf{R}; S), \quad (29)$$

where ξ_{NI} is the actual decoding error induced by the specific implementation of the canonical NI decoder. Analogously to Equation 20, $\Delta\xi_{NI}$ can be separated as follows:

$$\Delta\xi_{NI} = \underbrace{\Delta\xi_{NI}^{NIP}}_{\xi^{Min}(S; \mathbf{R}^{NIP}) - \xi^{Min}(S; \mathbf{R})} + \underbrace{\Delta\xi_{NI}^{Est}}_{\xi_{NI} - \xi^{Min}(S; \mathbf{R}^{NIP})}. \quad (30)$$

$\Delta\xi_{NI}^{NIP}$ and $\Delta\xi_{NI}^{Est}$ represent the increment in the minimum decoding error induced by the NI assumption and the estimation criterion, respectively. Among all mappings between \mathbf{R}^{NIP} and the decoded stimulus S^{NI} , there is one for which:

$$S^{NI} = \arg \min_{\tilde{S}} \left\{ \mathbf{E}_{P(S|\mathbf{R}^{NIP})} [\mathcal{L}(S, \tilde{S})] \right\}. \quad (31)$$

Such a decoder induces no additional increment in the minimum decoding error (Eq. 12). Therefore, $\Delta\xi_{NI}^{NIP}$ constitutes the minimum increment in the minimum decoding error attainable by at least one canonical NI decoder (i.e., those purely based on the NI assumption). Whenever $\Delta\xi_{NI}^{NIP}$ is zero, the NI assumption does not increase the minimum decoding error and noise correlations can be safely ignored.

Similarly to Equation 26, the increment in the decoding error induced by classical NI decoders (Fig. 5A) can be written as follows:

$$\Delta\xi_{NI} = \underbrace{\Delta\xi_{NI}^{NIP}}_{\xi^{Min}(S; \mathbf{R}^{NIP}) - \xi^{Min}(S; \mathbf{R}^{NIP})} + \underbrace{\Delta\xi_{NI}^{NIP}}_{\xi_{NI} - \xi^{Min}(S; \mathbf{R}^{NIP})} + \underbrace{\Delta\xi_{NI}^{Est}}_{\xi_{NI} - \xi^{Min}(S; \mathbf{R}^{NIP})}, \quad (32)$$

where $\Delta\xi_{NI}^{NIP}$ was defined in Equation 30. Here, $\Delta\xi_{NI}^{NIP}$ represents the increment in the minimum decoding error due to Bayes' rule. $\Delta\xi_{NI}^{NIP}$ is the minimum increment in the minimum decoding error attainable by all classical NI decoders and may be greater or equal to $\Delta\xi_{NI}^{NIP}$ (Eq. 27).

Any increment in the minimum decoding error occurs because, during the decoding process, some population responses are treated as identical. The importance of the distinction between two (or more) population responses, \mathbf{R}_A and \mathbf{R}_B , can be tested by first constructing a representation $\tilde{\mathbf{R}}$ that treats them as identical (but keeps the distinction between all other responses) and then computing as follows:

$$\Delta\xi_{\mathbf{R} \rightarrow \tilde{\mathbf{R}}} = \xi^{Min}(\tilde{\mathbf{R}}; S) - \xi^{Min}(\mathbf{R}; S) \geq 0. \quad (33)$$

Whenever $\Delta\xi_{\mathbf{R} \rightarrow \tilde{\mathbf{R}}}$ is zero, the distinction between \mathbf{R}_A and \mathbf{R}_B does not increment the minimum decoding error and can be safely ignored. Otherwise, $\Delta\xi_{\mathbf{R} \rightarrow \tilde{\mathbf{R}}}$ shows the increment in the minimum decoding error due to ignoring their distinction.

As an example, consider the population response shown in Figure 4, B and C. When responses are equally likely (Fig. 4B), $\Delta\xi_{NI}^{NIP}$ reaches its maximum value as follows:

$$\Delta\xi_{NI}^{NIP} = \min_{S^{Dec}} \left\{ \mathbf{E}_{P(S)} [\mathcal{L}(S, S^{Dec})] \right\}. \quad (34)$$

Indeed, after the NI assumption, all population responses are represented in the same way and the NI decoder performs at chance level. However, in all cases in which responses are not equally likely (Fig. 4C), $\Delta\xi_{NI}^{NIP}$ is zero. In other words, if responses are not equally likely, a canonical NI decoder exists that is capable of decoding the stimulus with the same accuracy as if noise correlations were taken into account (such a decoder is shown in Fig. 4C). Classical NI decoders operate substantially worse (Fig. 5C) because population responses become indistinguishable before the estimation process and thus perform at chance level for a wide range of response probabilities (Fig. 6B). Even though some information still remains in \mathbf{R}^{NIP} (Fig. 6A), it cannot be extracted using single responses.

Although, in general, ΔI_{NI}^{NIP} and $\Delta\xi_{NI}^{NIP}$ are not related deterministically (Thomson and Kristan, 2005), here we show some useful relations between these two quantities in specific cases as follows:

(35a) If ΔI_{NI}^{NIP} is zero, then $\Delta\xi_{NI}^{NIP}$ is zero. *Proof:* If $\Delta I_{NI}^{NIP} = 0$, then $I(S; \mathbf{R}|\mathbf{R}^{NIP}) = 0$ and hence \mathbf{R} can be written as a transformation (deterministic or stochastic) of \mathbf{R}^{NIP} . Following the data-processing inequality (Eq. 27), $\Delta\xi_{NI}^{NIP} = 0$.

(35b) *Corollary:* if $\Delta\xi_{NI}^{NIP} > 0$, then $\Delta I_{NI}^{NIP} > 0$.

(35c) If $\xi^{Min}(\mathbf{R}^{NIP}, S)$ is zero, then ΔI_{NI}^{NIP} and $\Delta\xi_{NI}^{NIP}$ are zero. *Proof:* If $\xi^{Min}(\mathbf{R}^{NIP}, S) = 0$, the data-processing inequality ensures that $\Delta\xi_{NI}^{NIP} = 0$. In the absence of errors, such NI decoder extracts all the encoded information, and thus $\Delta I_{NI}^{NIP} = 0$.

(35d) If ΔI_{NI}^{NIP} is equal to the encoded information $I(S; \mathbf{R})$, then $\xi^{Min}(\mathbf{R}^{NIP}, S)$ is given by Equation 34. *Proof:* In this case, S and \mathbf{R}^{NIP} are independent. The result follows from introducing independence into Equation 11.

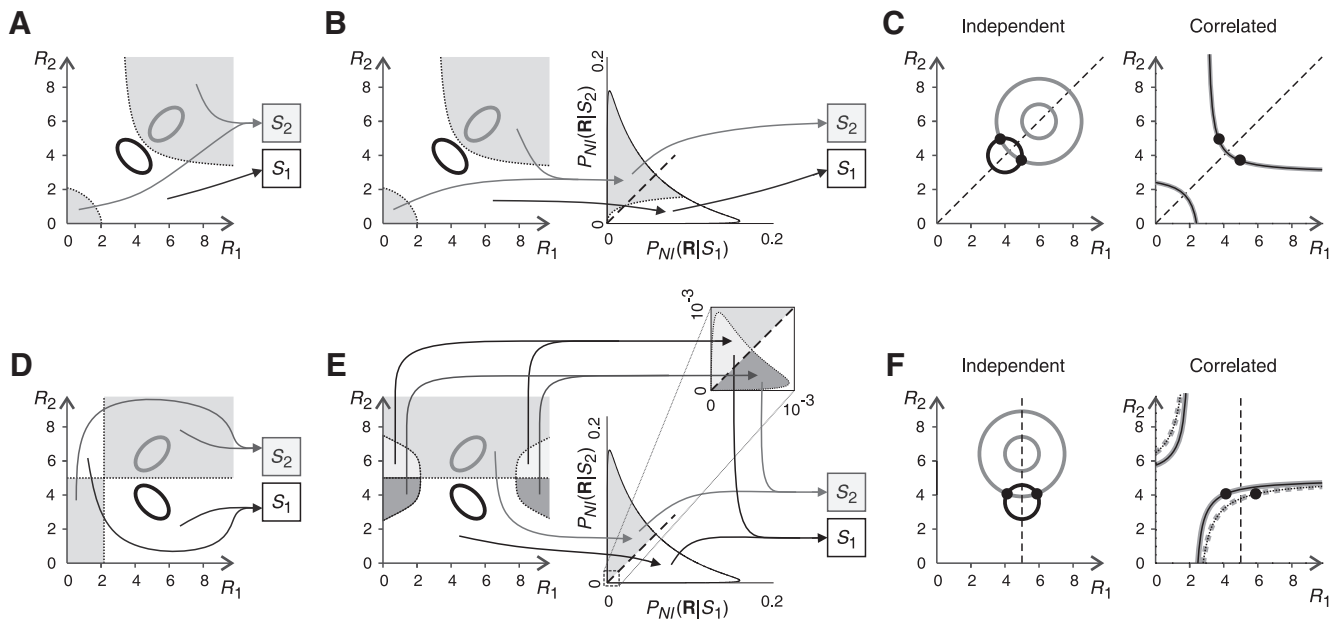


Figure 7. Impact of ignoring noise correlations when decoding population responses with Gaussian distributions. In both examples, black and gray ellipses represent the contour curves of the response distributions elicited by two stimuli, S_1 and S_2 , respectively. Stimuli are equally likely. Response parameters are as follows: $\mu_{nk} = 5 + (-1)^k$ (A–C) and $\mu_{nk} = 5 + \sqrt{2}(1 - n)^k$ (D–F); $\sigma_{nk} = 1$, $\rho_1 = -0.5$, and $\rho_2 = 0.5$ (Eq. 42). **A, D**, Optimal decoding strategy (minimization of the decoding-error probability) when correlations are known (Eq. 13). White regions are decoded as S_1 and gray regions as S_2 . **B, E**, Optimal decoding strategy when correlations are ignored (Eq. 31). The arrows show the transformation of the population response \mathbf{R} throughout the decoding process, as described in Figure 3. Population responses (left) are first transformed into vectors of NI likelihoods \mathbf{R}^{NI} (Eq. 19; middle). The distinction between regions filled with different gray levels is preserved. **B**, Optimal NI decoder maps the white region in the middle panel onto S_1 and the gray region onto S_2 , thus decoding population responses in the same way as in **A**. **E**, The first transformation merges population responses that are decoded differently in **D** (compare the regions), thus incrementing the minimum decoding error. The optimal NI decoder maps white and light-gray regions in the middle panel onto S_1 and gray and dark-gray regions onto S_2 . In both **B** and **E**, the optimal estimation criterion differs from the maximum likelihood (or maximum posterior) criterion, which maps regions above and below the diagonal (middle, dashed line) into S_2 and S_1 , respectively. **C, F**, Analysis of the responses \mathbf{R} that can be merged, after the NI assumption, without information loss. Left, Contour curves of the NI likelihoods $P_{NI}(\mathbf{R}|S_1)$ (black) and $P_{NI}(\mathbf{R}|S_2)$ (gray). Black dots represent two responses that, for each stimulus, have the same NI likelihoods (and are thus mapped onto the same \mathbf{R}^{NI}). Right, Contour curves of $P(S_1|\mathbf{R})$ (black) and $P(S_2|\mathbf{R})$ (gray) passing through the responses denoted in the left panel (continuous and dashed lines, respectively). **C**, The NI assumption merges pairs of responses \mathbf{R} that are symmetric with respect to the diagonal (left, dashed line). Because these pairs also have the same posterior probabilities (right), Equation 23 is fulfilled and no information is lost. **F**, The NI assumption merges pairs of responses \mathbf{R} that are symmetric with respect to the line $R_1 = 5$ (left, dashed line). These pairs have different posterior probabilities (right). Therefore, Equation 23 is not fulfilled and some information is lost.

The analysis can also be generalized to other measures of transmitted information, such as those defined by Victor and Nirenberg (2008), with the condition that they comply with the data-processing inequality (Cover and Thomas, 1991). Moreover, it can be extended to other probabilistic mismatched decoders, that is, to decoders constructed using stimulus-response probability distributions that differ from the real ones (Quiroga and Panzeri, 2009; Oizumi et al., 2010). One simply replaces the NI likelihoods with those corresponding to the probabilistic model under consideration. Decoding in the real brain may be subjected to additional constraints imposed by biophysics, connection length, metabolic cost, or robustness, which cannot simply be represented by a generalized measure of information. Our approach can be extended to these cases by computing the difference between the information transmitted by the optimal NI decoder that additionally satisfies these constraints with that of optimal decoders constructed with knowledge of noise correlations that operate under the same constraints.

Role of noise correlations in biologically plausible models

So far, we have discussed the role of noise correlations in examples in which the response space is discrete. In those examples, the minimum information loss and the minimum decoding error lead to the same conclusions about the role of noise correlations. However, more interesting models from the biological point of view generally involve responses varying in a continuum. In this

section, we apply our theoretical framework to continuum extensions of the discrete examples mentioned above and, on the way, compare what can be learned about the role of noise correlations when using the minimum information loss (ΔI_{NI}^{NL}) and the minimum increment in the minimum decoding error ($\Delta \xi_{NI}^{NL}$).

Consider the two examples shown in Figure 7, in which responses to each stimulus are drawn from Gaussian distributions. Similar examples have been previously studied assuming equal variance and correlations among neurons (Sompolinsky et al., 2001; Wu et al., 2001; Averbeck and Lee, 2006; Averbeck et al., 2006). These studies, however, have also pointed out that those highly homogeneous examples are rather unlikely in nature. More biologically driven examples should include differences in variances, in correlations, and in the symmetry of the distribution of responses among neurons as observed, for example, in the monkey MT area (Huang and Lisberger, 2009) and V1 area (Kohn and Smith, 2005). These differences, even if small, may change dramatically the role of noise correlations in neural decoding (Fig. 1). With these ideas in mind, we have here chosen two examples in which responses elicited by stimulus S_1 are negatively correlated whereas responses elicited by S_2 are positively correlated.

The example of Figure 7A–C is an extension to the continuum of the case studied in Figures 1A, 4A, and 5B. In what follows, we first show that, for the particular values of the parameters used in Figure 7A–C, noise correlations are irrelevant in neural decoding, a conclusion that previous estimators failed to reveal. However, later we show

that the irrelevance of noise correlations is a direct consequence of the specific values of the parameters used in this example and that, contrary to the discrete case, noise correlations are almost always important for arbitrary values of the parameters.

For the example shown in Figure 7A–C, the estimations of the minimum information loss ΔI_{NI}^{Min} are as follows:

$$\begin{aligned} \Delta I_{NI} &= 15.53\% & \Delta I_{NI}^D &= 6.43\% \\ \Delta I_{NI}^{LS} &= 15.53\% & \Delta I_{NI}^{DL} &= 6.32\% \end{aligned} \quad (36)$$

all showing that noise correlations are not dispensable. However,

$$\Delta I_{NI}^{NIL} = 0, \quad (37)$$

indicating that the NI assumption preserves all the encoded information, and that all previous estimations overestimated both ΔI_{NI}^{Min} and the importance of noise correlations.

To understand the discrepancy between our approach and previous estimators, see Figure 7C, left. Responses that are symmetric with respect to the diagonal $R_1 = R_2$ have the same NI likelihoods (Eq. 19). Therefore, after the NI assumption (Fig. 3 and Fig. 7B, middle), these responses are merged. Luckily, the distinction between these responses is not informative. Indeed, in the Figure 7C, right, we show that these responses have the same posterior probabilities and thus comply with Equation 23. Therefore, no information is lost after the NI assumption. In other words, noise correlations can be safely ignored.

Analogously, $\Delta \xi_{NI}^{NIL}$ is zero, and thus the NI assumption does not increment the minimum decoding error. To see this, compare Figure 7, A and B, in which we show the performance of optimal decoders with and without knowledge of correlations, respectively. Population responses associated with different decoded stimuli in Figure 7A are never merged after the NI assumption (Fig. 7B, middle). Therefore, the same mapping from population responses to decoded stimuli can be constructed even after the NI assumption takes place. Notice, nevertheless, that the optimal decision boundary based on the NI likelihoods is curved and differs from the maximum-likelihood criterion (dashed diagonal) or maximum-posterior criterion, which coincides with the maximum-likelihood criterion when stimuli are equally likely.

The situation is different in the example shown in Figure 7, D–F. Here, the estimations of ΔI_{NI}^{Min} are as follows:

$$\begin{aligned} \Delta I_{NI} &= 19.16\% & \Delta I_{NI}^D &= 3.22\% \\ \Delta I_{NI}^{LS} &= 19.16\% & \Delta I_{NI}^{DL} &= 3.22\% \end{aligned} \quad (38)$$

whereas

$$\Delta I_{NI}^{NIL} = 1.32\%, \quad (39)$$

indicating that previous estimations overestimated the importance of noise correlations, although noise correlations are not completely irrelevant. Notice that, as in the previous example, ΔI_{NI} (or ΔI_{NI}^{LS}) is far greater than ΔI_{NI}^{NIL} (and also greater than ΔI_{NI}^{DL} and ΔI_{NI}^D). These results indicate that the maximum rate of stimuli that can be processed without decoding errors decreases in $>1\%$ after ignoring noise correlations. One may wonder how the performance of the NI decoder is affected when decoding single population responses, a situation in which response sequences are short and decoding errors are allowed.

To answer this question, we determine the minimum decoding error (measured as the error probability; see Eq. 10 in Materials and Methods) with and without the NI assumption as follows:

$$\begin{aligned} \xi^{Min}(\mathbf{R}^{NIL}, S) &= 7.847\% \\ \xi^{Min}(\mathbf{R}, S) &= 7.834\% \end{aligned} \quad (40)$$

Their difference (which equals $\Delta \xi_{NI}^{NIL}$; see Eq. 30) represents only 0.166% of $\xi^{Min}(\mathbf{R}, S)$, indicating that noise correlations are almost irrelevant when decoding single responses.

To understand why this increment in the minimum decoding error occurs, compare Figure 7, D and E. Unlike the previous example, here, some population responses associated with different decoded stimuli in Figure 7D are merged after the NI assumption (Fig. 7E, middle). Therefore, the mapping from population responses onto decoded stimuli under the NI assumption is inevitably different from the mapping of an optimal decoder that takes correlations into account. The arrows from the middle to the right panels in Figure 7E indicate the optimal decoding strategy (achieving $\Delta \xi_{NI}^{NIL}$) once the NI assumption is made. The optimal mapping is constructed by first transforming the population response \mathbf{R} into \mathbf{R}^{NIL} (Eq. 19). Then, we decode for each \mathbf{R}^{NIL} the stimulus that most likely elicited all responses mapped into \mathbf{R}^{NIL} as follows:

$$S^{NI} = \begin{cases} S_1 & \text{if } P(S_1|\mathbf{R}^{NIL}) > P(S_2|\mathbf{R}^{NIL}) \\ S_2 & \text{otherwise} \end{cases} \quad (41)$$

where $P(S|\mathbf{R}^{NIL})$ is proportional to the sum of all joint probabilities $P(\mathbf{R}, S)$ whose response \mathbf{R} is mapped onto \mathbf{R}^{NIL} . Although, in general, population responses mapped onto regions above and below the dashed diagonal are decoded as S_2 and S_1 , respectively, for some regions near the origin of coordinates, the situation is reversed.

We can now generalize the results to arbitrary Gaussian distributions and stimulus probabilities, following the same reasoning as in Figure 7, C and F. Consider that the responses of two neurons R_1 and R_2 elicited by two stimuli S_1 and S_2 have a Gaussian distribution \mathcal{N} (Bishop, 2006) given as follows:

$$P(\mathbf{R} | S_k) = \mathcal{N} \left(\begin{bmatrix} R_1 \\ R_2 \end{bmatrix}, \begin{bmatrix} \mu_{1k} \\ \mu_{2k} \end{bmatrix}, \begin{bmatrix} \sigma_{1k}^2 & \tilde{\rho}_k \\ \tilde{\rho}_k & \sigma_{2k}^2 \end{bmatrix} \right), \quad (42)$$

where μ_{nk} , ρ_k , and σ_{nk} represent the mean values, correlation coefficients, and standard deviations, respectively, of the responses of the n^{th} neuron to stimulus S_k , and $\tilde{\rho}_k = \rho_k \sigma_{1k} \sigma_{2k}$. Noise correlations are almost always important for decoding except when the following conditions are met:

$$\frac{\sigma_{12}\sigma_{22}}{\sigma_{11}\sigma_{21}} = \frac{\rho_2(1 - \rho_1^2)}{\rho_1(1 - \rho_2^2)}, \quad \text{if } \mu_{11} = \mu_{12}, \text{ and } \mu_{21} = \mu_{22}; \quad (43a)$$

$$\frac{\sigma_{12}}{\sigma_{11}} = \frac{\sigma_{22}}{\sigma_{21}} = \sqrt{\frac{\rho_2(1 - \rho_1^2)}{\rho_1(1 - \rho_2^2)}}, \quad \text{if } \mu_{11} = \mu_{12}, \text{ or } \mu_{21} = \mu_{22}; \quad (43b)$$

$$\frac{\sigma_{21}}{\sigma_{11}} = \frac{\sigma_{22}}{\sigma_{12}} = \frac{\mu_{11} - \mu_{12}}{\mu_{21} - \mu_{22}}, \quad \text{if } \mu_{11} \neq \mu_{12}, \text{ and } \mu_{21} \neq \mu_{22}; \quad (43c)$$

Conditions 43a, 43b, and 43c establish relations between the mean values μ_{nk} , correlation coefficients ρ_k , and standard deviations σ_{nk} of the responses of the n^{th} neuron to stimulus S_k , respectively. Conditions 43a and 43b hold only when population responses always exhibit the same type of correlations for all stimuli (i.e., they are always positively correlated or always negatively correlated). Condition 43b also requires that all contour curves of the NI response distributions are shifted and/or scaled versions of one another (but not rotated). Finally, condition 43c analogously

constrains the shape of the contour curves, but holds for arbitrary correlation coefficients. Notice the change in the subindexes from condition 43b to condition 43c. For any departure from conditions 43a to 43c, noise correlations are important for decoding: Both ΔI_{NI}^{NIL} and $\Delta \xi_{NI}^{NIL}$ are greater than zero; their values depend on the specific case under study, and can range from ~ 0 to 100% (e.g., when condition 43a holds and variances are equal).

Discussion

In neural decoding, the importance of noise correlations has been linked to the minimum inefficiency of NI decoders. These decoders have been constructed using two different methods. The first one involves training specific types of decoders (generally linear) using surrogate NI responses (Nirenberg et al., 2001; Latham and Nirenberg, 2005; Quiñ Quiroga and Panzeri, 2009; Berens et al., 2012). Here, we showed that the inefficiency of these decoders may, depending on the decoding models and optimization functions, overestimate or underestimate the importance of noise correlations, and may not even be related to the NI assumption (Fig. 2). Therefore, the results obtained with this method ought to be observed with caution. The second method involves probabilistic decoders that explicitly take the NI assumption as part of the decoding algorithm (Nirenberg et al., 2001; Wu et al., 2001; Nirenberg and Latham, 2003; Latham and Nirenberg, 2005; Ince et al., 2010; Oizumi et al., 2010); the consistency with the NI assumption is therefore guaranteed (Nirenberg and Latham, 2003).

The inefficiency of probabilistic NI decoders (hereafter called NI decoders) has been previously assessed either by measuring the information preserved in their output (ΔI_{NI} and ΔI_{NI}^{LS} ; Nirenberg et al., 2001; Ince et al., 2010) or by using information theoretical quantities (ΔI_{NI}^D and ΔI_{NI}^{DL} ; Nirenberg et al., 2001; Nirenberg and Latham, 2003; Latham and Nirenberg, 2005; Oizumi et al., 2010). Here, we compared these estimators for a wide range of population responses and probability distributions. We found that none of them bound the inefficiency of all NI decoders tightly (Figs. 1, 7) and all of them overestimate the importance of noise correlations. Therefore, previous studies concluding that noise correlations are important based on these estimators may require a second evaluation with the methods presented here.

Previous studies have claimed that NI decoders inferring the stimulus from a maximum-posterior criterion (Eq. 7) are optimal. When operating with the true response probabilities, this criterion minimizes the decoding-error probability (Eq. 13). However, neither other definitions of decoding error (Eq. 10) nor the information loss (Eq. 2) are guaranteed to be minimized. When operating with NI response probabilities, not even the minimization of the decoding-error probability is guaranteed (unless ΔI_{NI}^D is zero; Nirenberg and Latham, 2003). As shown in Figures 5 and 7, using a maximum-posterior criterion may result in overestimating the minimum inefficiency of NI decoders and therefore the importance of noise correlations.

To solve this problem, we first modeled NI decoders as series of transformations of the population response, with only the first one embodying the NI assumption and the following ones representing the estimation criterion (Figs. 3, 5A). We then noticed that the information loss ΔI_{NI}^{NIL} induced by the first transformation is common to all NI decoders and, with no restrictions on the stimulus estimation algorithms, constitutes an attainable lower bound to the lost information (the coding theorem; Cover and Thomas, 1991). The computation of ΔI_{NI}^{NIL} (and also its variance and bias) can be done using standard tools for the analysis of

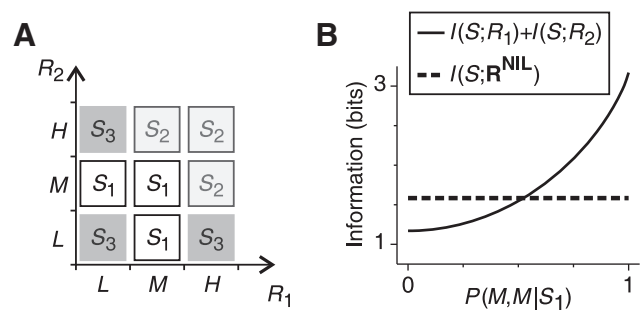


Figure 8. The NI decoder can paradoxically extract more information than that encoded by individual neurons. **A**, Example showing the responses of two neurons R_1 and R_2 elicited by three stimuli S_1 , S_2 , and S_3 . All stimuli are equally likely. Response probabilities $P(L, L | S_3)$, $P(M, M | S_1)$, and $P(H, H | S_2)$ are equal to α and response probabilities $P(L, M | S_1)$, $P(M, L | S_1)$, $P(L, H | S_3)$, $P(H, L | S_3)$, $P(M, H | S_2)$, and $P(H, M | S_2)$ are equal to $0.5 - \alpha/2$, with α varying between 0 and 1. **B**, The NI decoder is capable of extracting more information than the sum of the information encoded by individual neurons for a wide range of response probabilities. This effect is enhanced by the fact that the latter information is only an upper bound of the information conveyed individually by the neurons in the population.

neural codes (Montemurro et al., 2007; Panzeri et al., 2007; Eyherabide and Samengo, 2010).

The interpretation of NI decoders as sequences of processes is fundamental to understanding the role of noise correlations in neural decoding. Using this paradigm, we studied the effect of the NI assumption on later stages of the NI decoder. After the NI assumption, the application of Bayes' rule may give rise to additional information losses (Fig. 5C). Interestingly, information losses may be reduced by using different stimulus prior probabilities than those set in the experiment (Fig. 1G). However, when applied to the true conditional response probabilities (as opposed to the NI response probabilities), Bayes' rule induces no information loss. These results stress the remarkable differences (often overlooked) between decoding algorithms constructed with the real population-response probabilities and mismatched decoders (Oizumi et al., 2010).

Most importantly, we determined when and why information is lost by NI decoders in a single-response basis: information losses occur because some population responses that are informative are transformed in such a way that their distinction is unavailable for subsequent stages. To identify which distinctions are informative and which ones constitute noise, we do not rely on previous definitions of noise as mere variations around a mean (Oram et al., 1998; Sompolinsky et al., 2001; Averbek et al., 2006). Instead, our definition of noise explicitly evaluates the role of response variations in information transmission. Certainly, some variations around the mean are essential to information transmission even in the absence of noise correlations (Fig. 2B).

By analyzing the importance of noise correlations on a single-response basis, we found that, in broad terms, their importance depends on the relation between the number of stimuli (K) and the number of neurons (N) and that, in general, noise correlations are likely to be irrelevant. This approximate picture may explain why previous studies using different values of N and K often differed in the relevance ascribed to noise correlations. Moreover, it may aid the design of future experiments for which the outcomes depend on the importance of noise correlations. To get an accurate assessment of the importance of noise correlations in each individual case, however, one should rely on ΔI_{NI}^{NIL} and not on the approximate argument mentioned above.

The role of noise correlations in neural decoding, however, cannot be completely characterized with quantities solely based on mutual information because mutual information takes into account neither temporal constraints of real neural systems nor behavioral meaning of stimuli (Nirenberg and Latham, 2003). When operating on single responses, higher or lower decoded information may not be reflected directly in the efficiency of a decoder (Thomson and Kristan, 2005). Here, we proposed to additionally assess the role of noise correlations using quantities based on the minimum decoding error. This quantity exhibits many similarities with mutual information and, in addition, can handle short time windows and reflect the biological cost of making specific decoding errors.

Our results contrast with previous studies in three major points. First, we assess the role of correlations using ΔI_{NI}^{NIL} and $\Delta \xi_{NI}^{NIL}$ without explicitly displaying the best NI decoders (i.e., NI decoders that minimize ΔI_{NI}^{NIL} or $\Delta \xi_{NI}^{NIL}$). Our formulation has both benefits and limitations. The benefits are that we can draw conclusions about the importance of correlations with minimal computational cost. The limitation is that, if we do actually need to use a decoder, we have no explicit formula describing the best ones. To our knowledge, an explicit formula only exists for the NI decoder that minimizes the decoding error (Eq. 31), providing that correlations are known. Nevertheless, one should remember that minimizing the decoding error does not translate into maximizing the decoded information (Treves, 1997; Thomson and Kristan, 2005). In general, the best NI decoders must be found by searching among all possible NI decoders. To aid the search, our analysis provides insight into which distinctions between population responses must be preserved throughout the decoding process to achieve optimality.

Second, previous studies have argued that decoding strategies that ignore noise correlations are simpler than those taking noise correlations into account (Nirenberg et al., 2001; Wu et al., 2001; Nirenberg and Latham, 2003; Latham and Nirenberg, 2005; Averbeck and Lee, 2006; Averbeck et al., 2006; Ince et al., 2010; Oizumi et al., 2010). Even though the NI assumption simplifies the probabilistic encoding model, optimal NI decoders may require more complex estimation algorithms than those used in decoders constructed without the NI assumption. Other estimation algorithms may be simpler but less efficient (Fig. 7).

Third, our results do not support directly any qualitative claim about the nature of the decoded information (Nirenberg et al., 2001). The amount of information extracted by NI decoders may paradoxically depend on the noise correlations in the population response. For example, in Figure 7, the amount of information extracted by the optimal NI decoder does depend on the amount of correlation ρ in the neural response. The extracted information may even exceed the sum of the informations encoded individually by each neuron (Schneidman et al., 2003) regardless of whether or not surrogate NI population responses occur in the real data (Fig. 8). The solution to this paradox goes beyond quantitative arguments and requires a comparison of what sort of information (stimulus features) is individually encoded by each neuron with that extracted by NI decoders (Eyherabide and Samengo, 2010 and references therein).

To conclude, our work provides a rigorous framework for understanding, both quantitatively and qualitatively, the role of noise correlations in neural decoding. The quantities defined here allow one to quantify exactly the trade-off between the complexity and optimality of NI decoders, either in natural situations or under artificial conditions (i.e., using long sequences of re-

sponses). This assessment is fundamental for the development of computational algorithms, brain-machine interfaces, and neuro-prosthetics. Our description provides the basis for understanding how the NI assumption (or any other assumption during the decoding process) affects the amount and type of decoded information, establishing for the first time a link between probabilistic decoding models and the neural code. The framework is general enough to analyze the importance of noise correlations not only between neurons in neural populations, but also between neural populations in different cortical areas or, more recently, between cortical areas in different brains (Hari and Kujala, 2009; Babiloni and Astolfi, 2012).

Notes

Supplemental material for this article is available at <http://eyherabidehg.com.ar/>. This material includes additional examples and demonstrations and the codes for making the figures. This material has not been peer reviewed.

References

- Averbeck BB, Lee D (2006) Effects of noise correlations on information encoding and decoding. *J Neurophysiol* 95:3633–3644. CrossRef Medline
- Averbeck BB, Latham PE, Pouget A (2006) Neural correlations, population coding and computation. *Nat Rev Neurosci* 7:358–366. CrossRef Medline
- Babiloni F, Astolfi L (2012) Social neuroscience and hyperscanning techniques: past, present and future. *Neurosci Biobehav Rev*. Advance online publication. doi:10.1016/j.neubiorev.2012.07.006 CrossRef
- Berens P, Ecker AS, Cotton RJ, Ma WJ, Bethge M, Tolias AS (2012) A fast and simple population code for orientation in primate V1. *J Neurosci* 32:10618–10626. CrossRef Medline
- Bishop CM (2006) *Pattern recognition and machine learning*. New York: Springer.
- Borst A, Theunissen FE (1999) Information theory and neural coding. *Nat Neurosci* 2:947–957. CrossRef Medline
- Cover TM, Thomas JA (1991) *Elements of information theory*. New York: Wiley-Interscience.
- Duda RO, Hart PE, Stork DG (2000) *Pattern classification*, Ed 2. New York: Wiley.
- Eyherabide HG, Samengo I (2010) Time and category information in pattern-based codes. *Front Comput Neurosci* 4:145. CrossRef Medline
- Gawne TJ, Richmond BJ (1993) How independent are the messages carried by adjacent inferior temporal cortical neurons? *J Neurosci* 13:2758–2771. Medline
- Graf AB, Kohn A, Jazayeri M, Movshon JA (2011) Decoding the activity of neuronal populations in macaque primary visual cortex. *Nat Neurosci* 14:239–245. CrossRef Medline
- Hari R, Kujala MV (2009) Brain basis of human social interaction: From concepts to brain imaging. *Physiol Rev* 89:453–479. CrossRef Medline
- Hastie T, Tibshirani R, Friedman J (2009) *The elements of statistical learning: data mining, inference and prediction*, Ed 2. New York: Springer.
- Huang X, Lisberger SG (2009) Noise correlations in cortical MT area and their potential impact on trial-by-trial variation in the direction and speed of smooth-pursuit eye movements. *J Neurophysiol* 101:3012–3030. CrossRef Medline
- Ince RA, Senatore R, Arabzadeh E, Montani F, Diamond ME, Panzeri S (2010) Information-theoretic methods for studying population codes. *Neural Netw* 23:713–727. CrossRef Medline
- Klam F, Zemel RS, Pouget A (2008) Population coding with motion energy filters: the impact of correlations. *Neural Comput* 20:146–175. CrossRef Medline
- Knill DC, Pouget A (2004) The bayesian brain: the role of uncertainty in neural coding and computation. *Trends Neurosci* 27:712–719. CrossRef Medline
- Kohn A, Smith MA (2005) Stimulus dependence of neuronal correlation in primary visual cortex of the macaque. *J Neurosci* 25:3661–3673. CrossRef Medline
- Latham PE, Nirenberg S (2005) Synergy, redundancy, and independence in population codes, revisited. *J Neurosci* 25:5195–5206. CrossRef Medline
- Montemurro MA, Senatore R, Panzeri S (2007) Tight data-robust bounds

- to mutual information combining shuffling and model selection techniques. *Neural Comput* 19:2913–2957. [CrossRef Medline](#)
- Nirenberg S, Latham PE (2003) Decoding neuronal spike trains: How important are correlations. *Proc Natl Acad Sci U S A* 100:7348–7353. [CrossRef Medline](#)
- Nirenberg S, Carcieri SM, Jacobs AL, Latham PE (2001) Retinal ganglion cells act largely as independent decoders. *Nature* 411:698–701. [CrossRef Medline](#)
- Oizumi M, Ishii T, Ishibashi K, Hosoya T, Okada M (2010) Mismatched decoding in the brain. *J Neurosci* 30:4815–4826. [CrossRef Medline](#)
- Oram MW, Földiák P, Perrett DI, Sengpiel F (1998) The ‘ideal homunculus’: decoding neural population signals. *Trends Neurosci* 21:259–265. [CrossRef Medline](#)
- Panzeri S, Senatore R, Montemurro MA, Petersen RS (2007) Correcting for the sampling bias problem in spike train information measures. *J Neurophysiol* 98:1064–1072. [CrossRef Medline](#)
- Panzeri S, Brunel N, Logothetis NK, Kayser C (2010) Sensory neural codes using multiplexed temporal scales. *Trends Neurosci* 33:111–120. [CrossRef Medline](#)
- Pita-Almenar JD, Ranganathan GN, Koester HJ (2011) Impact of cortical plasticity on information signaled by populations of neurons in the cerebral cortex. *J Neurophysiol* 106:1118–1124. [CrossRef Medline](#)
- Quiñero R, Panzeri S (2009) Extracting information from neuronal populations: information theory and decoding approaches. *Nat Rev Neurosci* 10:173–185. [CrossRef Medline](#)
- Schneidman E, Bialek W, Berry MJ 2nd (2003) Synergy, redundancy, and independence in population codes. *J Neurosci* 23:11539–11553. [Medline](#)
- Shannon CE (1948) A mathematical theory of communication. *The Bell System Technical Journal* 27:379–423. [CrossRef](#)
- Simoncelli EP (2009) Optimal estimation in sensory systems. In: *The cognitive neurosciences*, Ed 4 (Gazzaniga MS), pp 525–538. Cambridge, MA: MIT.
- Sompolinsky H, Yoon H, Kang K, Shamir M (2001) Population coding in neuronal systems with correlated noise. *Phys Rev E Stat Nonlin Soft Matter Phys* 64:051904. [CrossRef Medline](#)
- Tao T (2011) *An introduction to measure theory*. Providence, RI: American Mathematical Society.
- Thomson EE, Kristan WB (2005) Quantifying stimulus discriminability: A comparison of information theory and ideal observer analysis. *Neural Comput* 17:741–778. [CrossRef Medline](#)
- Treves A (1997) On the perceptual structure of face space. *BioSystems* 40:189–196. [CrossRef Medline](#)
- Victor JD, Nirenberg S (2008) Indices for testing neural codes. *Neural Comput* 20:2895–2936. [CrossRef Medline](#)
- Wu S, Nakahara H, Amari S (2001) Population coding with correlation and an unfaithful model. *Neural Comput* 13:775–797. [CrossRef Medline](#)