When and why noise correlations are important in neural decoding

Supplementary material

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

[1] Centro Atómico Bariloche and Instituto Balseiro, San Carlos de Bariloche, Argentina.

[2] Department of Computer Science and Helsinki Institute for Information Technology, University of Helsinki, Finland.

Summary: The material shown in this document is associated with the publication

Eyherabide HG, Samengo I, When and why noise correlations are important in neural decoding, *J Neurosci* (2013), 33(45): 17921-17936; doi: http://dx.doi.org/10.1523/JNEUROSCI.0357-13.2013

It contains additional demonstrations, examples, and the codes for making the figures. Should you use this material, we kindly request you to cite the aforementioned publication. This material has not been peer reviewed.

Last update: 27/12/2013

Corresponding author: Hugo Gabriel Eyherabide - Department of Computer Science and Helsinki Institute for Information Technology - University of Helsinki - Gustaf Hällströmin katu 2b - 00560 Helsinki - Finland - Tel: +358919151237 - Fax: +358919151120 - Email: Hugo.Eyherabide@helsinki.fi

Contents

A Discrepancy between estimators of ΔI_{NI}^{Min}

The minimum information loss ΔI_{NI}^{Min} attainable by NI decoders has been estimated in several different ways. In Eyherabide and Samengo (2013), we analysed four different estimators and showed that all of them overestimate ΔI_{NI}^{Min} in a context dependent manner, and none of them constitutes a universal bound neither to ΔI_{NI}^{Min} nor to the importance of noise correlations in neural decoding. Here we show a detailed analysis of the divergence between these approaches using the example shown in Figure 1A of Eyherabide and Samengo (2013). This example shows the responses $\mathbf{R} = [R_1, R_2]$ of a population of two neurons elicited by each of two stimuli S_1 and S_2. The joint probabilities $P(R_1, R_2, S_1)$ and $P(R_1, R_2, S_2)$ are given by Table S-1, where $\hat{\alpha} = 1 - \alpha$, $\hat{\beta} = 1 - \beta$, and $\hat{p} = 1 - p$, being α, β and p constants with arbitrary real values between 0 and 1.

$P(R_1, R_2, S_1)$		R_1		
		L	M	H
R_2	H	0	0	0
	M	$\alpha\, p$	0	0
	L	0	$\hat{\alpha}\, p$	0

$P(R_1, R_2, S_2)$		R_1		
		L	M	H
R_2	H	0	0	$\hat{\beta}\,\hat{p}$
	M	0	$\beta\,\hat{p}$	0
	L	0	0	0

Table S-1. Joint stimulus-response probabilities $P(R_1, R_2, S)$ for the example shown in Figure 1A of Eyherabide and Samengo (2013).

The posterior probability $P(S_k|R_1, R_2)$ (k being 1 or 2) can be calculated using Bayes' rule

$$P(S_k|R_1, R_2) = \frac{P(R_1, R_2, S_k)}{\sum_{\tilde{k}} P(R_1, R_2, S_{\tilde{k}})}\,, \tag{S-1}$$

The values of $P(S_k|R_1, R_2)$ for the example analysed here are given in Table S-2.

$P(S_1|R_1, R_2)$		R_1		
		L	M	H
R_2	H	0	0	0
	M	1	0	0
	L	0	1	0

$P(S_2|R_1, R_2)$		R_1		
		L	M	H
R_2	H	0	0	1
	M	0	1	0
	L	0	0	0

Table S-2. Posterior probabilities $P(S|R_1, R_2)$ for the example given in Figure 1A and the joint stimulus-response probabilities $P(R_1, R_2, S)$ of Table S-1.

The noise-independent (NI) posterior distribution $P_{NI}(S|R_1,R_2)$ can be obtained using Eq. 5 in Eyherabide and Samengo (2013). The values of $P_{NI}(S|R_1,R_2)$ for the example analysed here are given in Table S-4, where $\gamma = \left(1 + \Phi \frac{p}{\hat{p}}\right)^{-1}$, $\Phi = \frac{\alpha\,\hat{\alpha}}{\beta^2}$ and $\hat{\gamma} = 1 - \gamma$.

$P_{NI}(S_1\|R_1,R_2)$		R_1		
		L	M	H
R_2	H	0	0	0
	M	1	$\hat{\gamma}$	0
	L	1	1	0

$P_{NI}(S_2\|R_1,R_2)$		R_1		
		L	M	H
R_2	H	0	1	1
	M	0	γ	1
	L	0	0	0

Table S-3. Noise-independent posterior probabilities $P_{NI}(S|R_1,R_2)$ for the example given in Figure 1A and the joint stimulus-response probabilities $P(R_1,R_2,S)$ of Table S-1.

From Table S-2 it is clear that a decoder based on $P(S_k|R_1,R_2)$ and using a maximum posterior criterion (that is, a decoder which output is the most likely stimulus given the population response) is capable of decoding without error. In other words, knowledge of noise correlations are sufficient to decode without error. Yet they may not be necessary. A decoder based on the NI posterior probability $P_{NI}(S|R_1,R_2)$, and hence without knowledge of noise correlations, may decode without error as well. In that case, noise correlations are not necessary for optimal decoding. The NI posterior probability $P_{NI}(S|R_1,R_2)$, however, differs from the real posterior distribution $P(S|R_1,R_2)$ for four different population responses $[R_1,R_2]$, namely: $[L,L]$, $[M,M]$, $[H,M]$, and $[M,H]$. Out of these four population responses, only response $[M,M]$ actually occurs with non-zero probability ($P(M,M) > 0$), and therefore is the only one that can induce a discrepancy between decoders constructed with and without knowledge of noise correlations.

A.1 Estimation the minimum information loss using ΔI^D_{NI}

Here we provide the exact formula of the estimator ΔI^D_{NI} (Nirenberg et al., 2001; Nirenberg and Latham, 2003) for the example shown in Figure 1A of Eyherabide and Samengo (2013). To that end, recall the estimators of the minimum information loss induced by NI decoders defined in Eq. 8 of Eyherabide and Samengo (2013). Using the probabilities given in Tables S-1, S-2 and S-3, the

value of ΔI^D_{NI} is given by

$$\Delta I^D_{NI} = \mathbf{D}\left[P(R_1,R_2,S)||P_{NI}(S|R_1,R_2)\right] \tag{S-2a}$$

$$= \sum_{S,R_1,R_2} P(S,R_1,R_2)\,\log_2\frac{P(S|R_1,R_2)}{P_{NI}(S|R_1,R_2)} \tag{S-2b}$$

$$= \overbrace{P(M,L,S_1)\,\log_2\frac{\overbrace{P(S_1|M,L)}^{1}}{P_{NI}(S_1|M,L)}}^{0} + \overbrace{P(L,M,S_1)\,\log_2\frac{\overbrace{P(S_1|L,M)}^{1}}{P_{NI}(S_1|L,M)}}^{0} + \\ + P(M,M,S_2)\,\log_2\frac{P(S_2|M,M)}{\underbrace{P_{NI}(S_2|M,M)}_{\left(1+\Phi\frac{p}{\hat{p}}\right)}} + \underbrace{P(H,H,S_2)\,\log_2\frac{\underbrace{P(S_2|H,H)}_{1}}{P_{NI}(S_2|H,H)}}_{0} \tag{S-2c}$$

$$= \beta\,\hat{p}\,\log_2\left(1+\Phi\frac{p}{\hat{p}}\right) \tag{S-2d}$$

where $\mathbf{D}$ is the Kullback-Leibler divergence (Cover and Thomas, 1991). ΔI^D_{NI} is almost always greater than zero, except when any of p, α and β are 0 or 1 (when p is 0 or 1, the amount of encoded information is zero).

A.2 Estimation the minimum information loss using ΔI^{DL}_{NI}

Here we provide the exact formula of the estimator ΔI^{DL}_{NI} (Latham and Nirenberg, 2005; Oizumi et al., 2010) for the example shown in Figure 1A of Eyherabide and Samengo (2013). The calculation of ΔI^{DL}_{NI} is somewhat tricky, because in their original form, the quantities involved in its calculation are not well defined. In this section, we show how to resolve these issues for the example of Figure 1A. The correct definitions of the quantities involved in the calculation of ΔI^{DL}_{NI} are derived in Section B.

The first step in the calculation of ΔI^{DL}_{NI} is to calculate $\Delta \tilde{I}^{DL}_{NI}(\theta)$ in a similar manner to ΔI^D_{NI}. Next, one needs to minimize $\Delta \tilde{I}^{DL}_{NI}(\theta)$ over the parameter θ (θ is a real number). The quantity $\Delta \tilde{I}^{DL}_{NI}(\theta)$ is

defined as

$$\Delta\tilde{I}_{NI}^{DL}(\theta) = \mathbf{D}[P(S|R_1,R_2)||\tilde{P}(S|R_1,R_2,\theta)]\,, \tag{S-3}$$

where $\tilde{P}(S_k|R_1,R_2,\theta)$ is given by

$$\tilde{P}(S_k|R_1,R_2,\theta) = \frac{P(S_k)\left(P(R_1|S_k)\,P(R_2|S_k)\right)^\theta}{\sum_{\tilde{k}} P(S_{\tilde{k}})\left(P(R_1|S_{\tilde{k}})\,P(R_2|S_{\tilde{k}})\right)^\theta}\,. \tag{S-4}$$

Unfortunately, this definition of $\tilde{P}(S_k|R_1,R_2,\theta)$ is invalid when θ is zero or negative and some $P(R_n|S_k)$ is zero for responses that occur with $P(R_1,R_2)$ greater than zero (n is the index of the neuron in the population and k is the index of the stimulus). This is the case of responses $[M,L]$, $[L,M]$ and $[H,H]$. Luckily, a more general definition can be derived, as we do in the next section (Eqs. S-22 and S-23). In the example under study, this results in

$$\tilde{P}(S_1|M,L,\theta) = \frac{p\,\hat{\alpha}^{2\theta}}{p\,\hat{\alpha}^{2\theta}} = 1 \tag{S-5a}$$

$$\tilde{P}(S_1|L,M,\theta) = \frac{p\,\alpha^{2\theta}}{p\,\alpha^{2\theta}} = 1 \tag{S-5b}$$

$$\tilde{P}(S_2|M,M,\theta) = \frac{\hat{p}\,\beta^{2\theta}}{p\,\hat{\alpha}\,\alpha^{\theta} + \hat{p}\,\beta^{2\theta}} = \left(1+\Phi^\theta\,\frac{p}{\hat{p}}\right)^{-1} \tag{S-5c}$$

$$\tilde{P}(S_2|H,H,\theta) = \frac{\hat{p}\,\hat{\beta}^{2\theta}}{p\,\hat{\beta}^{2\theta}} = 1\,. \tag{S-5d}$$

$\Delta\tilde{I}_{NI}^{DL}(\theta)$ can be calculated as follows

$$\Delta\tilde{I}_{NI}^{DL}(\theta) = \mathbf{D}\left[P(R_1,R_2,S)||\tilde{P}(S|R_1,R_2,\theta)\right] \tag{S-6a}$$

$$= \sum_{S,R_1,R_2} P(S,R_1,R_2)\,\log_2\frac{P(S|R_1,R_2)}{\tilde{P}(S|R_1,R_2,\theta)} \tag{S-6b}$$

$$= P(M,L,S_1)\,\log_2\frac{\overbrace{1}^{0}}{\tilde{P}(S_1|M,L,\theta)} + P(L,M,S_1)\,\log_2\frac{\overbrace{1}^{0}}{\tilde{P}(S_1|L,M,\theta)} +$$

$$+ P(M,M,S_2)\,\log_2 \frac{P(S_2|M,M)}{\underbrace{\tilde{P}(S_2|M,M,\theta)}_{\left(1+\Phi^\theta\,\frac{p}{\hat{p}}\right)}} + P(H,H,S_2)\,\log_2 \underbrace{\frac{P(S_2|H,H)}{\tilde{P}(S_2|H,H,\theta)}}_{1} \qquad \text{(S-6c)}$$

$$= \beta\,\hat{p}\,\log_2\left(1+\Phi^\theta\,\frac{p}{\hat{p}}\right). \qquad \text{(S-6d)}$$

The estimator ΔI_{NI}^{DL} is obtained by minimizing $\Delta\tilde{I}_{NI}^{DL}(\theta)$ with respect to θ (Eq. 8d). There are three possible cases depending on the value of Φ:

i) $\Phi < 1$, in which case $\Delta\tilde{I}_{NI}^{DL}(\theta)$ tends to zero as θ tends to infinity;

ii) $\Phi = 1$, in which case $\Delta\tilde{I}_{NI}^{DL}(\theta)$ is constant and equal to ΔI_{NI}^{D}; and

iii) $\Phi > 1$, in which case $\Delta\tilde{I}_{NI}^{DL}(\theta)$ tends to zero as θ tends to minus infinity.

As a results, ΔI_{NI}^{DL} is given by

$$\Delta I_{NI}^{DL} = \begin{cases} 0 & \Phi \neq 1 \\ \Delta I_{NI}^{D} & \Phi = 1 \end{cases}. \qquad \text{(S-7)}$$

A.3 Estimation the minimum information loss using MAP NI decoders

Consider a specific implementation of the noise-independent (NI) decoder using the maximum a-posteriori criterion (MAP). This NI decoder, here called MAP NI decoder, was defined in Eq. 7 of Eyherabide and Samengo (2013), and for each population response, selects the stimulus with the greatest NI posterior probability $P_{NI}(S|R_1,R_2)$, that is

$$S_{NI}^{MAP} = \arg\max_S P_{NI}(S|R_1,R_2)\,. \qquad \text{(S-8)}$$

The NI posterior probabilities $P_{NI}(S|R_1,R_2)$ were calculated in Table S-3. The only population response for which the MAP NI decoder can make an error is $[R_1,R_2] = [M,M]$, and this will occur

only when $P_{NI}(S_2|M,M) < P_{NI}(S_1|M,M)$. Therefore, when

$$P_{NI}(S_2|M,M) \geq P_{NI}(S_1|M,M)\,, \tag{S-9}$$

the MAP NI decoder is optimal ($\Delta I_{NI} = 0$; also $\Delta I_{NI}^{LS} = 0$, defined in Eq. 8b of Eyherabide and Samengo, 2013) and decodes with no error. It may not be obvious why the MAP NI decoder can be optimal when $P_{NI}(S_2|M,M) = P_{NI}(S_1|M,M)$. Indeed, in this case there is no preference for either stimuli, but the decoder must still choose a stimulus as its output. Therefore, there are several possible MAP NI decoders, depending on how this choice is made. One possibility is to select randomly one of the two stimuli every time $P_{NI}(S_2|R_1,R_2) = P_{NI}(S_2|R_1,R_2)$ for any population response, and this strategy of course results in a suboptimal decoder. Another possibility is that, whenever $P_{NI}(S_2|R_1,R_2) = P_{NI}(S_2|R_1,R_2)$, the MAP NI decoder chooses stimulus S_2. Such MAP NI decoder is optimal.

The condition for the optimality of the MAP NI decoder (Eq. S-9) can also be written as

$$\Phi \leq \frac{\hat{p}}{p}\,. \tag{S-10}$$

Comparing with Eq. S-7, we find that in the region

$$p \leq \hat{p} \qquad \text{and} \qquad \alpha\,\hat{\alpha} = \beta^2\,, \tag{S-11}$$

the classical NI decoder is optimal, even though $\Delta I_{NI}^{DL} > 0$.

A.4 Estimation the minimum information loss using ML NI decoders

One could also construct a specific implementation of the noise-independent (NI) decoder using the maximum likelihood criterion (ML), as opposed to the maximum a-posteriori (MAP). This NI decoder, here called ML NI decoder, is one among many other canonical NI decoders defined in Eq. 4 of Eyherabide and Samengo (2013), and for each population response, selects the stimulus with the

highest NI likelihood $P_{NI}(R_1, R_2|S)$, that is

$$S_{NI}^{ML} = \arg\max_S P_{NI}(R_1, R_2|S)\,. \tag{S-12}$$

The NI likelihoods $P_{NI}(R_1, R_2|S)$ for the example analysed in this section are given in Table S-3. Analogously to the MAP NI decoder, the only population response for which the ML NI decoder can

$P_{NI}(R_1, R_2\|S_1)$		R_1		
		L	M	H
R_2	H	0	0	0
	M	α^2	$\alpha\,\hat{\alpha}$	0
	L	$\alpha\,\hat{\alpha}$	$\hat{\alpha}^2$	0

$P_{NI}(R_1, R_2\|S_2)$		R_1		
		L	M	H
R_2	H	0	$\beta\hat{\beta}$	$\hat{\beta}^2$
	M	0	β^2	$\beta\hat{\beta}$
	L	0	0	0

Table S-4. Noise-independent likelihoods $P_{NI}(R_1, R_2|S)$ for the example given in Figure 1A and the joint stimulus-response probabilities $P(R_1, R_2, S)$ of Table S-1.

make an error is $[R_1, R_2] = [M, M]$, and this will occur only when $P_{NI}(M, M|S_2) < P_{NI}(M, M|S_1)$. Therefore, when

$$P_{NI}(M, M|S_2) \geq P_{NI}(M, M|S_1)\,, \tag{S-13}$$

the ML NI decoder is optimal ($\Delta I_{NI} = 0$; also $\Delta I_{NI}^{LS} = 0$) and decodes with no error. The condition for the optimality of the ML NI decoder (Eq. S-13) can also be written as

$$\Phi \leq 1\,. \tag{S-14}$$

Comparing with Eq. S-7, we find that the ML NI decoder is optimal even though $\Delta I_{NI}^{DL} > 0$.

Comparing Eqs. S-10 and S-14, we find that whenever

$$\Phi \leq \max\left(1, \frac{\hat{p}}{p}\right), \tag{S-15}$$

the MAP NI decoder or ML NI decoder is optimal (or both). These two implementations, however, are only two among all possible implementations of canonical NI decoders. Different implementations with optimal performance may exist for cases in which Eq. S-15 does not hold.

To answer whether optimal implementations of the canonical NI decoder exist, we quantified in Eyherabide and Samengo (2013) the exact value of the minimum information loss ΔI_{NI}^{Min} attainable by NI decoders. We proved that the value of ΔI_{NI}^{Min} is equal to the value of ΔI_{NI}^{NIL}, defined in Eq. 21. In Section C of this document we show that, the minimum information loss ΔI_{NI}^{NIL} is zero for any values of the joint stimulus-response probabilities $P(R_1, R_2, S)$ given in Table S-1. Therefore, all four criteria described in Eq. 8 of Eyherabide and Samengo (2013) overestimate ΔI_{NI}^{Min} and, ultimately, the importance of noise correlations.

B Characterization of the estimation ΔI_{NI}^{DL}

Latham and Nirenberg (2005) introduced ΔI_{NI}^{DL} as an estimator of the minimum information loss ΔI_{NI}^{Min}. As it was defined, however, this estimator and its putative properties are only valid for the subset of all possible population responses where marginal probabilities are greater than zero for all stimuli. This is the case of the example shown in Figure 1A of Eyherabide and Samengo (2013). In this section, we extend its definition to all possible population responses and reassessed the validity of its properties.

B.1 Shortcomings in the estimation of ΔI_{NI}^{DL}

The estimator ΔI_{NI}^{DL} (Latham and Nirenberg, 2005; Oizumi et al., 2010) is obtained by minimizing $\Delta \tilde{I}_{NI}^{DL}(\theta)$ over the parameter θ (Eq. 8d of Eyherabide and Samengo, 2013)

$$\Delta I_{NI}^{DL} = \min_{\theta} \Delta \tilde{I}_{NI}^{DL}(\theta) \,, \tag{S-16}$$

where θ can take any real value. The quantity $\Delta \tilde{I}_{NI}^{DL}(\theta)$ (Eq. 9a of Eyherabide and Samengo, 2013) is given by

$$\Delta \tilde{I}_{NI}^{DL}(\theta) = \mathbf{D}[P(S|\mathbf{R}) || \tilde{P}(S|\mathbf{R}, \theta)] \,, \tag{S-17}$$

$\mathbf{R} = [R_1, \ldots, R_N]$ representing the response of a population of N neurons, and $\tilde{P}(S|\mathbf{R}, \theta)$ (Eq. 9b of Eyherabide and Samengo, 2013) given by

$$\tilde{P}(S|\mathbf{R}, \theta) = \frac{P(S)\left(\prod_n P(R_n|S)\right)^\theta}{\sum_{\hat{S}} P(\hat{S})\left(\prod_n P(R_n|\hat{S})\right)^\theta}. \tag{S-18}$$

Unfortunately, this definition of $\tilde{P}(S|\mathbf{R}, \theta)$, given in Latham and Nirenberg (2005) and Oizumi et al. (2010), is not valid when θ is not positive and the population response $\mathbf{R}$ involved in the calculation gives rise to marginal probabilities $P(R_n|S_k)$ that vanish for at least one neuron of index n and one stimulus S_k.

Recall the example shown in Figure 1A of Eyherabide and Samengo (2013). Response $[R_1, R_2] = [H, H]$ is associated with the following marginal probabilities

$$P(R_1 = H|S_1) = 0 \qquad P(R_2 = H|S_1) = 0 \tag{S-19a}$$

$$P(R_1 = H|S_2) = \hat{\beta} \qquad P(R_2 = H|S_2) = \hat{\beta}, \tag{S-19b}$$

which are derived using Table S-1. According to Eq. S-18, $\tilde{P}(S_1|H, H, \theta)$ is given by

$$\tilde{P}(S_1|H, H, \theta) = \frac{P(S_1)\left(P(R_1|S_1)\,P(R_2|S_1)\right)^\theta}{P(S_1)\left(P(R_1|S_1)\,P(R_2|S_1)\right)^\theta + P(S_2)\left(P(R_1|S_2)\,P(R_2|S_2)\right)^\theta} \tag{S-20a}$$

$$= \frac{P(S_1)\,0^\theta}{P(S_1)\,0^\theta + P(S_2)\hat{\beta}^{2\theta}}. \tag{S-20b}$$

When $\theta = 0$ and when $\theta < 0$, this equation is indeterminate (indeterminations of type 0^0 and $\frac{\infty}{\infty}$, respectively).

To resolve this problem, we took a look at the derivation of Eq. S-18 in Latham and Nirenberg

(2005). There, Eq. B13a states that $\tilde{P}(S|\mathbf{R},\theta)$ must satisfy the following constraint

$$\mathop{\mathbf{E}}_{\tilde{P}(R_1,R_2,S_k,\theta)}\left[\log_2\Big(P(R_1|S_k)\,P(R_2|S_k)\Big)\right] = \mathop{\mathbf{E}}_{P(R_1,R_2,S_k)}\left[\log_2\Big(P(R_1|S_k)\,P(R_2|S_k)\Big)\right]. \tag{S-21}$$

This constraint can be satisfied if and only if $\tilde{P}(S_k|R_1,R_2,\theta)$ is zero whenever $P(R_1|S_k)$ or $P(R_2|S_k)$ are zero. More generally, $\tilde{P}(S_k|\mathbf{R},\theta)$ must be zero whenever the marginal probability $P(R_n|S_k)$ is zero for some neuron R_n in the population. A more general definition of $\tilde{P}(S_k|\mathbf{R},\theta)$ is therefore given by

$$\tilde{P}(S|\mathbf{R},\theta=0) = \begin{cases} \dfrac{P(S)\Big(\prod_n P(R_n|S)\Big)^{\theta}}{\sum_{\hat{S}} P(\hat{S})\Big(\prod_n P(R_n|\hat{S})\Big)^{\theta}} & \text{if } \prod_n P(R_n|S) > 0 \qquad \text{(S-22)} \\ 0 & \text{if } \prod_n P(R_n|S) = 0 \qquad \text{(S-23)} \end{cases}$$

where the sum extends over all $\hat{S}$ for which $\prod_n P(R_n|\hat{S}) > 0$. Notice that Eq. S-23 is not redundant with Eq. S-22 because the latter is not valid when $\prod_n P(R_n|S) = 0$ and $\theta \leq 0$.

B.2 Putative stimulus-response independence when θ is zero

Latham and Nirenberg (2005) argued that when $\theta = 0$, stimuli and responses become independent, and thus $\Delta\tilde{I}^{DL}_{NI}(\theta)$ (Eq. S-17 and Eq. 9a in Eyherabide and Samengo, 2013) approaches is maximum value, i.e. the encoded information. This argument, however, does not take into account that $\tilde{P}(S|\mathbf{R},\theta)$ is not well defined by Eq. S-18 when $\prod_n P(R_n|S) = 0$ and $\theta \leq 0$. Using Eq. B.1, we find that

$$\tilde{P}(S|\mathbf{R},\theta=0) = \begin{cases} \dfrac{P(S)}{\sum_{\hat{S}} P(\hat{S})} & \text{if } \prod_n P(R_n|S) > 0 \qquad \text{(S-24)} \\ 0 & \text{if } \prod_n P(R_n|S) = 0 \qquad \text{(S-25)} \end{cases}$$

where the sum extends over all $\hat{S}$ for which $\prod_n P(R_n|\hat{S}) > 0$. Whenever a stimulus $\hat{S}$ exists for which $\prod_n P(R_n|\hat{S}) = 0$, the denominator in Eq. S-24 is less than unity, and $\tilde{P}(S|\mathbf{R},\theta=0) > P(S)$. In other words, stimuli and responses do not become independent, and $\Delta\tilde{I}^{DL}_{NI}(\theta)$ does not equal the encoded

information, but the following quantity

$$\Delta\tilde{I}_{NI}^{DL}(\theta=0) = \mathbf{D}[P(S|\mathbf{R})||\tilde{P}(S|\mathbf{R},\theta=0)] \tag{S-26a}$$

$$= \mathop{\mathbf{E}}_{P(S,\mathbf{R})}\left[\log_2 \frac{P(S|\mathbf{R})}{\tilde{P}(S|\mathbf{R},\theta=0)}\right] \tag{S-26b}$$

$$= \underbrace{\mathop{\mathbf{E}}_{P(S,\mathbf{R})}\left[\log_2 P(S|\mathbf{R})\right]}_{-H(S|R)} - \mathop{\mathbf{E}}_{P(S,\mathbf{R})}\log_2\left[\tilde{P}(S|\mathbf{R},\theta=0)\right] \tag{S-26c}$$

$$= -H(S|R) - \mathop{\mathbf{E}}_{P(S,\mathbf{R})}\left[\log_2 \frac{P(S)}{\sum_{\hat{S}} P(\hat{S})}\right] \tag{S-26d}$$

$$= -H(S|R) - \underbrace{\mathop{\mathbf{E}}_{P(S,\mathbf{R})}\left[\log_2 P(S)\right]}_{-H(S)} + \mathop{\mathbf{E}}_{P(S,\mathbf{R})}\left[\log_2 \sum_{\hat{S}} P(\hat{S})\right] \tag{S-26e}$$

$$= \underbrace{-H(S|R) + H(S)}_{I(\mathbf{R},S)} + \mathop{\mathbf{E}}_{P(S,\mathbf{R})}\left[\log_2 \sum_{\hat{S}} P(\hat{S})\right] \tag{S-26f}$$

$$= I(\mathbf{R},S) + \underbrace{\mathop{\mathbf{E}}_{P(S,\mathbf{R})}\left[\log_2 \sum_{\hat{S}} P(\hat{S})\right]}_{<0} \tag{S-26g}$$

$$< I(\mathbf{R},S)\,, \tag{S-26h}$$

where $I(\mathbf{R},S)$ is the encoded information. In consequence, the putative independence between stimuli and responses for $\theta = 0$ ought to be handled with caution, as its validity depends on whether or not a stimulus S exists for which $\prod_n P(R_n|S) > 0$ for some population response $\mathbf{R}$ that occurs with nonzero probability $P(\mathbf{R}) > 0$.

B.3 Uniqueness of the minimum of $\Delta\tilde{I}_{NI}^{DL}(\theta)$

Latham and Nirenberg (2005) also stated that $\Delta\tilde{I}_{NI}^{DL}(\theta)$ has a single minimum, except when population responses are deterministic. However, in section A we showed that, when $\Phi = 1$, $\Delta\tilde{I}_{NI}^{DL}(\theta) = \Delta I_{NI}^{D}$ for responses that are not deterministic (i.e. a constant for all θ; page 7). In other words, we showed

that the minimum of $\Delta\tilde{I}_{NI}^{DL}(\theta)$ may not be unique for stochastic stimulus-response mappings as well. Indeed, when $P_{NI}(\mathbf{R}|S)$ is constant (like in Figure 1A of Eyherabide and Samengo, 2013 with equal response probabilities), then

$$\frac{\partial\Delta\tilde{I}_{NI}^{DL}(\theta)}{\partial\theta} = \underset{\tilde{P}(S,\mathbf{R},\theta)}{\mathbf{E}} \log_2 P_{NI}(\mathbf{R}|S) - \underset{P(S,\mathbf{R})}{\mathbf{E}} \log_2 P_{NI}(\mathbf{R}|S) = 0\,. \tag{S-27}$$

Thus, $\Delta\tilde{I}_{NI}^{DL}(\theta)$ is constant and independent of θ, even though responses were stochastic.

C Representations $\mathbf{R}^{\mathbf{NIL}}$ and $\mathbf{R}^{\mathbf{NIP}}$

In this section, we extend the analysis of the examples shown in Figures 4 and 5 of Eyherabide and Samengo (2013) to all possible values of the joint probabilities $P(R_1, R_2, S)$. Recall that the representation $\mathbf{R}^{\mathbf{NIL}}$ and $\mathbf{R}^{\mathbf{NIP}}$ of the population responses $\mathbf{R} = [R_1, R_2]$ are obtained using Eqs. 19 and 25 of Eyherabide and Samengo (2013) as follows

$$\mathbf{R}^{\mathbf{NIL}} = \Big[P_{NI}(\mathbf{R}|S_1),\ P_{NI}(\mathbf{R}|S_2)\Big] \tag{S-28a}$$

$$\mathbf{R}^{\mathbf{NIP}} = \Big[P_{NI}(S_1|\mathbf{R}),\ P_{NI}(S_2|\mathbf{R})\Big]\,. \tag{S-28b}$$

In Figures 4 and 5, these representations are depicted in the shaded panels.

C.1 Example shown in Figures 4A and 5B

Consider the stimulus and response probabilities like those given in Table S-1. The representations $\mathbf{R}^{\mathbf{NIL}}$ and $\mathbf{R}^{\mathbf{NIP}}$ are given by Table S-5. As a result, responses associated with different stimuli are always represented in a different manner, both by $\mathbf{R}^{\mathbf{NIL}}$ and $\mathbf{R}^{\mathbf{NIP}}$.

$\mathbf{R}$	$\longrightarrow$	$\mathbf{R}^{\mathbf{NIL}}$	$\longrightarrow$	$\mathbf{R}^{\mathbf{NIP}}$
$[L,M]$	$\longrightarrow$	$[\alpha^2, 0]$	$\longrightarrow$	$[1, 0]$
$[M,L]$	$\longrightarrow$	$[\hat{\alpha}^2, 0]$	$\longrightarrow$	$[1, 0]$
$[M,M]$	$\longrightarrow$	$[\alpha\,\hat{\alpha}, \beta^2]$	$\longrightarrow$	$\dfrac{[\alpha\,\hat{\alpha}\,p,\ \beta^2\,\hat{p}\,]}{\alpha\,\hat{\alpha}\,p+\beta^2\,\hat{p}}$
$[H,H]$	$\longrightarrow$	$[0, \hat{\beta}^2]$	$\longrightarrow$	$[0, 1]$

Table S-5. Representations $\mathbf{R}^{NIL}$ and $\mathbf{R}^{NIP}$ of the population response $\mathbf{R}$ for example of Figures 4A and 5B.

C.2 Example shown in Figures 4B, 4C and 5C

Consider that the joint probabilities $P(R_1, R_2, S)$ for each stimulus S_1 and S_2 and each population response $\mathbf{R} = [R_1, R_2]$ are given by Table S-6, where $\hat{\alpha} = 1-\alpha$, $\hat{\beta} = 1-\beta$, and $\hat{p} = 1-p$. The

$P(\mathbf{R}, S_1)$		R_1	
		L	H
R_2	H	$\alpha\,p$	0
	L	0	$\hat{\alpha}\,p$

$P(\mathbf{R}, S_2)$		R_1	
		L	H
R_2	H	0	$\beta\,\hat{p}$
	L	$\hat{\beta}\,\hat{p}$	0

Table S-6. Joint stimulus-response probabilities $P(R_1, R_2, S)$ for the example shown in Figures 4B, 4C and 5B of Eyherabide and Samengo (2013).

representations $\mathbf{R}^{\mathbf{NIL}}$ and $\mathbf{R}^{\mathbf{NIP}}$ are given by Table S-7. Population responses associated with different stimuli are always mapped onto different $\mathbf{R}^{\mathbf{NIL}}$, and thus are always represented in a different manner after the NI assumption, except for $\alpha = \beta = 0.5$. This case can be regarded as unique, but if it indeed occurs, then information is completely lost.

The representation $\mathbf{R}^{\mathbf{NIP}}$, however, sometimes merges population responses associated with different stimuli. There are two cases where responses are merged: (a) when $\alpha = \beta$, in which case, response $[R_1, R_2] = [L, H]$ is merged with $[L, L]$, and $[H, L]$ with $[H, H]$; and (b) when $\alpha = \hat{\beta}$, in which case, response $[L, H]$ is merged with $[H, H]$, and $[H, L]$ with $[L, L]$ (Figure 5C in Eyherabide and Samengo, 2013) . In both cases, the minimum decoding error $\xi^{Min}(\mathbf{R}^{\mathbf{NIP}}, S)$ (Eq. 32 of Eyherabide and Samengo, 2013) and the minimum information loss ΔI_{NI}^{NIP} (Eq. 26 of Eyherabide and Samengo,

$\mathbf{R}$	$\longrightarrow$	$\mathbf{R}^{\mathbf{NIL}}$	$\longrightarrow$	$\mathbf{R}^{\mathbf{NIP}}$
$[\,L\,,H\,]$	$\longrightarrow$	$[\,\alpha^2\,,\beta\hat{\beta}]$	$\longrightarrow$	$\dfrac{[\;\alpha^2\,p\;,\;\beta\hat{\beta}\,\hat{p}\;]}{\alpha^2\,p+\beta\hat{\beta}\,\hat{p}}$
$[\,H\,,L\,]$	$\longrightarrow$	$[\,\hat{\alpha}^2\,,\beta\hat{\beta}]$	$\longrightarrow$	$\dfrac{[\;\hat{\alpha}^2\,p\;,\;\beta\hat{\beta}\,\hat{p}\;]}{\hat{\alpha}^2\,p+\beta\hat{\beta}\,\hat{p}}$
$[\,L\,,L\,]$	$\longrightarrow$	$[\alpha\,\hat{\alpha},\hat{\beta}^2\,]$	$\longrightarrow$	$\dfrac{[\;\alpha\,\hat{\alpha}\,p\,,\;\hat{\beta}^2\,\hat{p}\;]}{\alpha\,\hat{\alpha}\,p+\hat{\beta}^2\,\hat{p}}$
$[\,H\,,H\,]$	$\longrightarrow$	$[\alpha\,\hat{\alpha},\beta^2\,]$	$\longrightarrow$	$\dfrac{[\;\alpha\,\hat{\alpha}\,p\,,\;\beta^2\,\hat{p}\;]}{\alpha\,\hat{\alpha}\,p+\beta^2\,\hat{p}}$

Table S-7. Representations $\mathbf{R}^{NIL}$ and $\mathbf{R}^{NIP}$ of the population response $\mathbf{R}$ for example of Figures 4B, 4C and 5C.

2013) achieved by classical NI decoders take intermediate values depending on α and p (Figure 6 in Eyherabide and Samengo, 2013). The minimum decoding error (measured as the decoding-error probability) is given by

$$\Delta\xi_{NI}^{NIP} = \min\{\alpha, \hat{\alpha}, p, \hat{p}\}\ . \tag{S-29}$$

The information loss was estimated numerically using the codes provided in the Supplementary Material. These cases constitute examples where NI decoders can be optimal, but for achieving optimality, the estimation must be based purely on the NI assumption, that is, on $\mathbf{R}^{\mathbf{NIL}}$.

D Data processing inequality for the minimum decoding error

Here we provide more details about the proof of the data processing inequality for the minimum decoding error $\xi^{Min}(\mathbf{R};S)$ (Eq. 28 in Eyherabide and Samengo, 2013), which states that $\xi^{Min}(\mathbf{R};S)$ increases with transformations of the population response $\mathbf{R}$. Consider a transformation $\tilde{\mathbf{R}} = g(\mathbf{R})$. The minimum decoding error $\xi^{Min}(\mathbf{R};S)$ is given by Eq. 11 of Eyherabide and Samengo (2013),

namely

$$\xi^{Min}(\mathbf{R};S) = \sum_{\mathbf{R}} P(\mathbf{R}) \min_{S^{Dec}} \left\{ \sum_{S} P(S|\mathbf{R})\, \mathcal{L}(S, S^{Dec}) \right\}, \tag{S-30}$$

where S^{Dec} is the decoded stimulus, and $\mathcal{L}(S, S^{Dec})$ is non-negative and represents the cost of decoding S^{Dec} when the encoded stimulus was S. The probability $P(\mathbf{R})$ can be written as

$$P(\mathbf{R}) = \sum_{\tilde{\mathbf{R}}} P(\mathbf{R}, \tilde{\mathbf{R}}). \tag{S-31}$$

Replacing in Eq. S-30 we obtain

$$\xi^{Min}(\mathbf{R};S) = \sum_{\mathbf{R}} \sum_{\tilde{\mathbf{R}}} P(\mathbf{R}, \tilde{\mathbf{R}}) \min_{S^{Dec}} \left\{ \sum_{S} P(S|\mathbf{R})\, \mathcal{L}(S, S^{Dec}) \right\} \tag{S-32a}$$

$$= \sum_{\tilde{\mathbf{R}}} P(\tilde{\mathbf{R}}) \sum_{\mathbf{R}} P(\mathbf{R}|\tilde{\mathbf{R}}) \min_{S^{Dec}} \left\{ \sum_{S} P(S|\mathbf{R})\, \mathcal{L}(S, S^{Dec}) \right\}. \tag{S-32b}$$

This step was taken when passing from Eq. 28a to Eq. 28b. Then, we used the fact that the function $\min_{X} \{f(X,Y)\}$ is concave, that is

$$\min_{X} \left\{ \sum_{Y} \lambda_Y f(X,Y) \right\} \geq \sum_{Y} \lambda_Y \min_{X} \{f(X,Y)\}, \tag{S-33}$$

for all $\lambda_Y \geq 0$. This can be proved for two terms as follows

$$\min_{X} \left\{ \lambda_{Y_1} f(X,Y_1) + \lambda_{Y_2} f(X,Y_2) \right\} = \lambda_{Y_1} \overbrace{f(X_0,Y_1)}^{\geq \min_X \{f(X,Y_1)\}} + \lambda_{Y_2} \overbrace{f(X_0,Y_2)}^{\geq \min_X \{f(X,Y_2)\}} \quad \text{for some } X_0 \tag{S-34a}$$

$$\geq \lambda_{Y_1} \min_{X} \{f(X,Y_1)\} + \lambda_{Y_2} \min_{X} \{f(X,Y_2)\}, \tag{S-34b}$$

and, by induction, the proof can be extended to any number of terms. In Eq. S-32b, the concavity of $\min_{X} \{f(X,Y)\}$ implies that

$$\sum_{\mathbf{R}} P(\mathbf{R}|\tilde{\mathbf{R}}) \min_{S^{Dec}} \left\{ \sum_{S} P(S|\mathbf{R})\, \mathcal{L}(S, S^{Dec}) \right\} \leq \min_{S^{Dec}} \left\{ \sum_{\mathbf{R}} P(\mathbf{R}|\tilde{\mathbf{R}}) \sum_{S} P(S|\mathbf{R})\, \mathcal{L}(S, S^{Dec}) \right\}. \tag{S-35}$$

Replacing in Eq. S-32b, we obtain

$$\xi^{Min}(\mathbf{R};S) \leq \sum_{\tilde{\mathbf{R}}} P(\tilde{\mathbf{R}}) \min_{S^{Dec}} \left\{ \sum_{\mathbf{R}} P(\mathbf{R}|\tilde{\mathbf{R}}) \sum_{S} P(S|\mathbf{R})\, \mathcal{L}(S, S^{Dec}) \right\} \tag{S-36a}$$

$$= \underbrace{\sum_{\tilde{\mathbf{R}}} P(\tilde{\mathbf{R}}) \min_{S^{Dec}} \left\{ \sum_{S} P(S|\tilde{\mathbf{R}})\, \mathcal{L}(S, S^{Dec}) \right\}}_{=\xi^{Min}(\tilde{\mathbf{R}};S)} . \tag{S-36b}$$

Comparing the right term with Eq. S-30 we find that this is the minimum decoding error $\xi^{Min}(\tilde{\mathbf{R}};S)$ that can be attained after the transformation $\tilde{\mathbf{R}} = g(\mathbf{R})$, and therefore

$$\xi^{Min}(\mathbf{R};S) \leq \xi^{Min}(\tilde{\mathbf{R}};S) . \tag{S-37}$$

This result coincides with Eq. 28d of Eyherabide and Samengo (2013).

E The importance of noise correlations when the response distributions are Gaussian

Consider that the responses of two neurons R_1 and R_2 elicited by two stimuli S_1 and S_2 have a two-dimensional Gaussian distribution $\mathcal{N}$ given by

$$P(\mathbf{R}|S_k) = \frac{1}{\sqrt{\det\left(2\pi\,\mathbf{C}_k\right)}}\, \mathbf{e}^{-\frac{1}{2}\,(\mathbf{R}-\boldsymbol{\mu}_k)^{\mathrm{T}}\,\mathbf{C}_k^{-1}\,(\mathbf{R}-\boldsymbol{\mu}_k)} , \tag{S-38}$$

where the population response $\mathbf{R}$, the mean value $\boldsymbol{\mu}_k$, and the covariance matrices $\mathbf{C}_k$ are given by

$$\mathbf{R} = \begin{bmatrix} R_1 \\ R_2 \end{bmatrix} \qquad \boldsymbol{\mu}_k = \begin{bmatrix} \mu_{1k} \\ \mu_{2k} \end{bmatrix} \qquad \mathbf{C}_k = \begin{bmatrix} \sigma_{1k}^2 & \tilde{\rho}_k \\ \tilde{\rho}_k & \sigma_{2k}^2 \end{bmatrix} . \tag{S-39}$$

Here, μ_{nk} and σ_{nk} represent the mean value and the variance of the responses of neuron R_n to stimulus S_k, and $\tilde{\rho}_k = \rho_k\,\sigma_{1k}\,\sigma_{2k}$, being ρ_k the correlation coefficient of the responses of both neurons elicited by stimulus S_k.

The NI likelihood (Eq. 3 in Eyherabide and Samengo, 2013) can be written as

$$P_{NI}(\mathbf{R}|S_k) = \frac{1}{\sqrt{\det\left(2\pi\,\mathbf{C}_k^{NI}\right)}}\,\mathbf{e}^{-\frac{1}{2}\,(\mathbf{R}-\mu_\mathbf{k})^{\mathrm{T}}\left(\mathbf{C}_k^{NI}\right)^{-1}(\mathbf{R}-\mu_\mathbf{k})}, \tag{S-40}$$

where the noise-independent covariance matrices $\mathbf{C}_k^{NI}$ are given by

$$\mathbf{C}_k^{NI} = \begin{bmatrix} \sigma_{1k}^2 & 0 \\ 0 & \sigma_{2k}^2 \end{bmatrix}. \tag{S-41}$$

$\mathbf{C}_k^{NI}$ only differs from $\mathbf{C}_k$ in the non-diagonal elements.

Recall conditions 24a-c for the optimality of an NI decoder derived in the third section of Results. Noise correlations are irrelevant for decoding if and only if any two population responses merged after the NI assumption (i.e. having the same NI likelihoods) are non-informative (and thus, comply with Eq.23). In the example analysed here, population responses have Gaussian distributions, and hence two population responses $\mathbf{R}_A = [R_{A1}, R_{A2}]$ and $\mathbf{R}_B = [R_{B1}, R_{B2}]$ are merged after the NI assumption if they comply with

$$P_{NI}(\mathbf{R}_A|S_k) = P_{NI}(\mathbf{R}_B|S_k) \tag{S-42}$$

for all k (that is, for k being 1 and 2 in this example). Replacing by Eq. S-40 we obtain

$$(\mathbf{R}_A-\boldsymbol{\mu}_\mathbf{k})^{\mathrm{T}}\left(\mathbf{C}_k^{NI}\right)^{-1}(\mathbf{R}_A-\boldsymbol{\mu}_\mathbf{k}) = (\mathbf{R}_B-\boldsymbol{\mu}_\mathbf{k})^{\mathrm{T}}\left(\mathbf{C}_k^{NI}\right)^{-1}(\mathbf{R}_B-\boldsymbol{\mu}_\mathbf{k}) \tag{S-43}$$

which can be further simplified as follows

$$\frac{(R_{A1}-\mu_{1k})^2}{\sigma_{1k}^2} + \frac{(R_{A2}-\mu_{2k})^2}{\sigma_{2k}^2} = \frac{(R_{B1}-\mu_{1k})^2}{\sigma_{1k}^2} + \frac{(R_{B2}-\mu_{2k})^2}{\sigma_{2k}^2}. \tag{S-44}$$

Pairs of responses complying with Eq. S-44 for both $k = 1$ and $k = 2$ are merged after the NI assumption. This transformation, however, is lossless if the responses are non-informative. Pairs of

responses are non-informative if they comply with Eq. 23 of Eyherabide and Samengo (2013), that is

$$P(S_k|\mathbf{R}_A) = P(S_k|\mathbf{R}_B)\,, \tag{S-45}$$

for both $k = 1$ and $k = 2$. Because the posterior probabilities $P(S|\mathbf{R}_A)$ and $P(S|\mathbf{R}_B)$ are normalised to unity, if two responses comply with Eq. S-45 for stimulus S_1, they also comply with Eq. S-45 for stimulus S_2. Hence, from now on we focus on Eq. S-45 for stimulus S_1. This equation can be written as

$$\frac{P(\mathbf{R}_A|S_1)\,P(S_1)}{P(\mathbf{R}_A|S_1)\,P(S_1) + P(\mathbf{R}_A|S_2)\,P(S_2)} = \frac{P(\mathbf{R}_B|S_1)\,P(S_1)}{P(\mathbf{R}_B|S_1)\,P(S_1) + P(\mathbf{R}_B|S_2)\,P(S_2)}\,, \tag{S-46}$$

and replacing by Eq. S-38 we obtain the following

$$\left[1 + \frac{\dfrac{P(S_2)}{\sqrt{\det\left(2\pi\,\mathbf{C}_2\right)}}\,\mathbf{e}^{-\frac{1}{2}\,\tilde{\mathbf{R}}_A^{2\,\mathrm{T}}\,\mathbf{C}_2^{\ -1}\,\tilde{\mathbf{R}}_A^2}}{\dfrac{P(S_1)}{\sqrt{\det\left(2\pi\,\mathbf{C}_1\right)}}\,\mathbf{e}^{-\frac{1}{2}\,\tilde{\mathbf{R}}_A^{1\,\mathrm{T}}\,\mathbf{C}_1^{\ -1}\,\tilde{\mathbf{R}}_A^1}}\right]^{-1} = \left[1 + \frac{\dfrac{P(S_2)}{\sqrt{\det\left(2\pi\,\mathbf{C}_2\right)}}\,\mathbf{e}^{-\frac{1}{2}\,\tilde{\mathbf{R}}_B^{2\,\mathrm{T}}\,\mathbf{C}_2^{\ -1}\,\tilde{\mathbf{R}}_B^2}}{\dfrac{P(S_1)}{\sqrt{\det\left(2\pi\,\mathbf{C}_1\right)}}\,\mathbf{e}^{-\frac{1}{2}\,\tilde{\mathbf{R}}_B^{1\,\mathrm{T}}\,\mathbf{C}_1^{\ -1}\,\tilde{\mathbf{R}}_B^1}}\right]^{-1} \tag{S-47}$$

where $\tilde{\mathbf{R}}_A^k = \mathbf{R}_A - \boldsymbol{\mu}_k$ and $\tilde{\mathbf{R}}_B^k = \mathbf{R}_B - \boldsymbol{\mu}_k$. After simplifications and rearrangements of the terms it becomes

$$\tilde{\mathbf{R}}_A^{2\,\mathrm{T}}\,\mathbf{C}_2^{\ -1}\,\tilde{\mathbf{R}}_A^2 - \tilde{\mathbf{R}}_B^{2\,\mathrm{T}}\,\mathbf{C}_2^{\ -1}\,\tilde{\mathbf{R}}_B^2 = \tilde{\mathbf{R}}_A^{1\,\mathrm{T}}\,\mathbf{C}_1^{\ -1}\,\tilde{\mathbf{R}}_A^1 - \tilde{\mathbf{R}}_B^{1\,\mathrm{T}}\,\mathbf{C}_1^{\ -1}\,\tilde{\mathbf{R}}_B^1\,. \tag{S-48}$$

Further simplification can be obtained by noticing that

$$\mathbf{C}_k^{\ -1} = \frac{1}{\det\left(\mathbf{C}_k\right)}\begin{bmatrix}\sigma_{2k}^2 & -\tilde{\rho}_k\\ -\tilde{\rho}_k & \sigma_{1k}^2\end{bmatrix} \tag{S-49a}$$

$$= \frac{1}{\det\left(\mathbf{C}_k\right)}\left(\underbrace{\begin{bmatrix}\sigma_{2k}^2 & 0\\ 0 & \sigma_{1k}^2\end{bmatrix}}_{\det\left(\mathbf{C}_k^{NI}\right)\left(\mathbf{C}_k^{NI}\right)^{-1}} + \begin{bmatrix}0 & -\tilde{\rho}_k\\ -\tilde{\rho}_k & 0\end{bmatrix}\right) \tag{S-49b}$$

$$= \frac{\det\left(\mathbf{C}_k^{NI}\right)}{\det\left(\mathbf{C}_k\right)}\left(\mathbf{C}_k^{NI}\right)^{-1} + \frac{1}{\det\left(\mathbf{C}_k\right)}\begin{bmatrix}0 & -\tilde{\rho}_k\\ -\tilde{\rho}_k & 0\end{bmatrix}. \tag{S-49c}$$

Replacing into Eq. S-48 we find that the left side becomes

$$\frac{\det\left(\mathbf{C}_2^{NI}\right)}{\det\left(\mathbf{C}_2\right)} \overbrace{\left(\tilde{\mathbf{R}}_A^{2\,\mathrm{T}}\,\mathbf{C}_2^{NI^{-1}}\,\tilde{\mathbf{R}}_A^2 - \tilde{\mathbf{R}}_B^{2\,\mathrm{T}}\,\mathbf{C}_2^{NI^{-1}}\,\tilde{\mathbf{R}}_B^2\right)}^{=\,0\ \text{(Eq. S-43)}} + \frac{1}{\det\left(\mathbf{C}_2\right)} \underbrace{\left(\tilde{\mathbf{R}}_A^{2\,\mathrm{T}}\begin{bmatrix}0 & -\tilde{\rho}_2\\ -\tilde{\rho}_2 & 0\end{bmatrix}\tilde{\mathbf{R}}_A^2 - \tilde{\mathbf{R}}_B^{2\,\mathrm{T}}\begin{bmatrix}0 & -\tilde{\rho}_2\\ -\tilde{\rho}_2 & 0\end{bmatrix}\tilde{\mathbf{R}}_B^2\right)}_{2\,\tilde{\rho}_2\left(\tilde{R}_{B1}^2\,\tilde{R}_{B2}^2 - \tilde{R}_{A1}^2\,\tilde{R}_{A2}^2\right)}, \tag{S-50}$$

Applying an analogous transformation to the right side of Eq. S-48 we arrive at

$$\frac{\tilde{\rho}_2\ \det\left(\mathbf{C}_1\right)}{\tilde{\rho}_1\ \det\left(\mathbf{C}_2\right)} = \frac{(R_{B1}-\mu_{11})\,(R_{B2}-\mu_{21}) - (R_{A1}-\mu_{11})\,(R_{A2}-\mu_{21})}{(R_{B1}-\mu_{12})\,(R_{B2}-\mu_{22}) - (R_{A1}-\mu_{12})\,(R_{A2}-\mu_{22})}\,. \tag{S-51}$$

Whenever pairs of responses satisfying simultaneously Eq. S-44 for both stimuli and Eq. S-51, they are non-informative, and merging them after the NI assumption induces no information loss.

In order to simplify the analysis of the relevance of noise correlations in decoding, we study separately the following three cases:

Case I: Response distributions with equal mean values, that is $\mu_{11} = \mu_{12}$ and $\mu_{21} = \mu_{22}$.

Case II: Response distributions parallel to one axis, i.e. $\mu_{11} = \mu_{12}$ or $\mu_{21} = \mu_{22}$.

Case III: Response distributions with other mean values, i.e. $\mu_{11} \neq \mu_{12}$ and $\mu_{21} \neq \mu_{22}$.

E.1 Case I: $\mu_{11} = \mu_{12}$ and $\mu_{21} = \mu_{22}$

Because the mean values of the response distributions coincide, Eq. S-51 becomes

$$\frac{\tilde{\rho}_2\ \det\left(\mathbf{C}_1\right)}{\tilde{\rho}_1\ \det\left(\mathbf{C}_2\right)} = 1 \tag{S-52a}$$

$$\frac{\rho_2\,\sigma_{11}\,\sigma_{21}\,(1-{\rho_1}^2)}{\rho_1\,\sigma_{12}\,\sigma_{22}\,(1-{\rho_2}^2)} = 1 \tag{S-52b}$$

$$\frac{\sigma_{12}\,\sigma_{22}}{\sigma_{11}\,\sigma_{21}} = \frac{\rho_2\,(1-{\rho_1}^2)}{\rho_1\,(1-{\rho_2}^2)}\,. \tag{S-52c}$$

Therefore, noise correlations are almost always crucial for decoding except in those cases where the variances and the correlation coefficients fulfill Eq. S-52c, which is the same as Eq. 43a in Eyherabide and Samengo (2013). Since the left side is always positive, this condition is fulfilled only if both correlation coefficients have the same sign.

An example in which noise correlations are irrelevant for decoding is shown in Figure S-1. In this example, the parameters of the response distributions comply with Eq. S-52c. Each contour curve (a curve along which a function remains constant) of the NI likelihood associated with stimulus S_1 (blue curves) may intersect each contour of the NI likelihood curve associated with stimulus S_2 (orange curves) in up to four population responses. Responses lying in the intersections of a pair of contour curves are symmetric with respect to the origin of coordinates (panel A). Because these responses have the same NI likelihood, they are merged after the NI assumption, and consequently information may be lost. Such an information loss does not occur, however, because these merged responses are non-informative. Analogously to the responses merged after the NI assumption, the contour curves of the posterior probabilities $P(S_1|R_1,R_2)$ and $P(S_2|R_1,R_2)$ are also symmetric with respect to the origin of coordinates (panel B), and therefore, the posterior probabilities of the merged responses comply with Eq. S-45.

Noise correlations are important for decoding whenever the parameters of the response distributions do not comply with Eq. S-52c. An example is shown in Figure S-2. This example differs from the one shown in Figure S-1 solely in the value of the correlation coefficient ρ_1. NI likelihoods are therefore the same, and so is the symmetry of responses merged after the NI assumption (panel A). The contour curves of the posterior probabilities $P(S_1|R_1,R_2)$ and $P(S_2|R_1,R_2)$, however, are not symmetric with respect to the origin of coordinates (panel B). Except for those lying on the axes of coordinates, all responses merged after the NI assumption do not comply with Eq. S-45 (they have different posterior probabilities). These responses are informative, and merging them after the NI assumption induces an information loss.

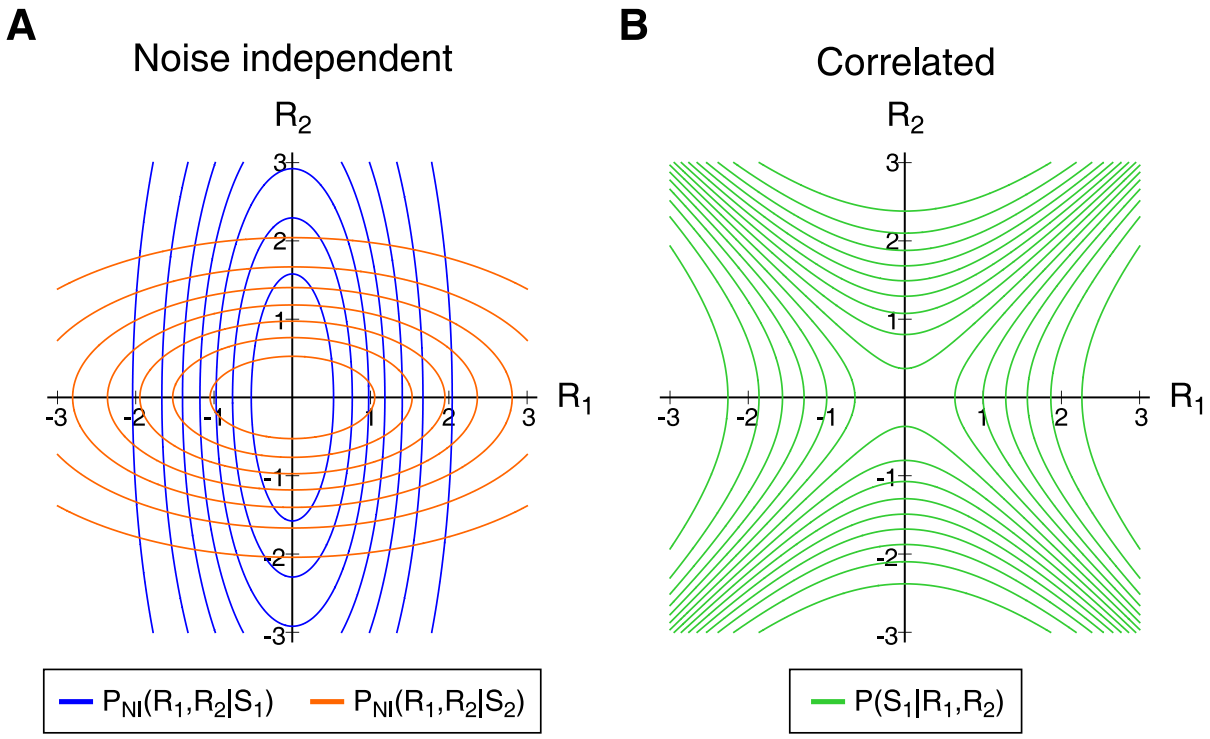

Figure S-1. Example in which noise correlations are irrelevant for decoding. *A*, Contour curves of noise-independent likelihoods $P_{NI}(R_1, R_2|S_1)$ and $P_{NI}(R_1, R_2|S_2)$. *B*, Contour curves of posterior probability $P(S_1|R_1, R_2)$ (which coincide with the contour curves of $P(S_2|R_1, R_2)$ defined by $P(S_2|R_1, R_2) = 1 - P(S_1|R_1, R_2)$). Response parameters are: $\boldsymbol{\mu}_1 = [0, 0]$, $\boldsymbol{\mu}_2 = [0, 0]$, $\sigma_{11} = \sigma_{22} = 1$, $\sigma_{21} = 3$, $\sigma_{12} = 2$, $\rho_1 \approx 0.41$ and $\rho_2 = 0.3$. These response parameters comply with Eq. S-52c.

E.2 Case II: $\mu_{11} = \mu_{12}$ or $\mu_{21} = \mu_{22}$

In this section we study the case in which $\mu_{11} = \mu_{12}$ (that is, the mean values of the responses of R_1 elicited by either stimuli are identical). The same results apply to the case $\mu_{21} = \mu_{22}$ after interchanging neurons 1 and 2. The analysis can be simplified by centering and scaling the distributions through the following change of variables

$$\tilde{R}_1 = \frac{R_1 - \mu_{11}}{\sigma_{11}} \qquad \text{and} \qquad \tilde{R}_2 = \frac{R_2 - (\mu_{21} + \mu_{22})/2}{(\mu_{21} - \mu_{22})/2}\,, \tag{S-53}$$

In the new variables, the response distributions are located at $\tilde{\boldsymbol{\mu}}_1 = [0, 1]$ and $\tilde{\boldsymbol{\mu}}_2 = [0, -1]$ with variances $\tilde{\sigma}_{1k} = \sigma_{1k}/\sigma_{11}$ and $\tilde{\sigma}_{2k} = 2\,\sigma_{2k}/(\mu_{21} - \mu_{22})$ for each stimulus S_k. As a result, Eq. S-44 for both stimuli becomes

$$\left(\tilde{R}_{A1}\right)^2 + \left(\frac{\tilde{R}_{A2} - 1}{\tilde{\sigma}_{21}}\right)^2 = \left(\tilde{R}_{B1}\right)^2 + \left(\frac{\tilde{R}_{B2} - 1}{\tilde{\sigma}_{21}}\right)^2 \qquad \text{for } S_1 \tag{S-54}$$

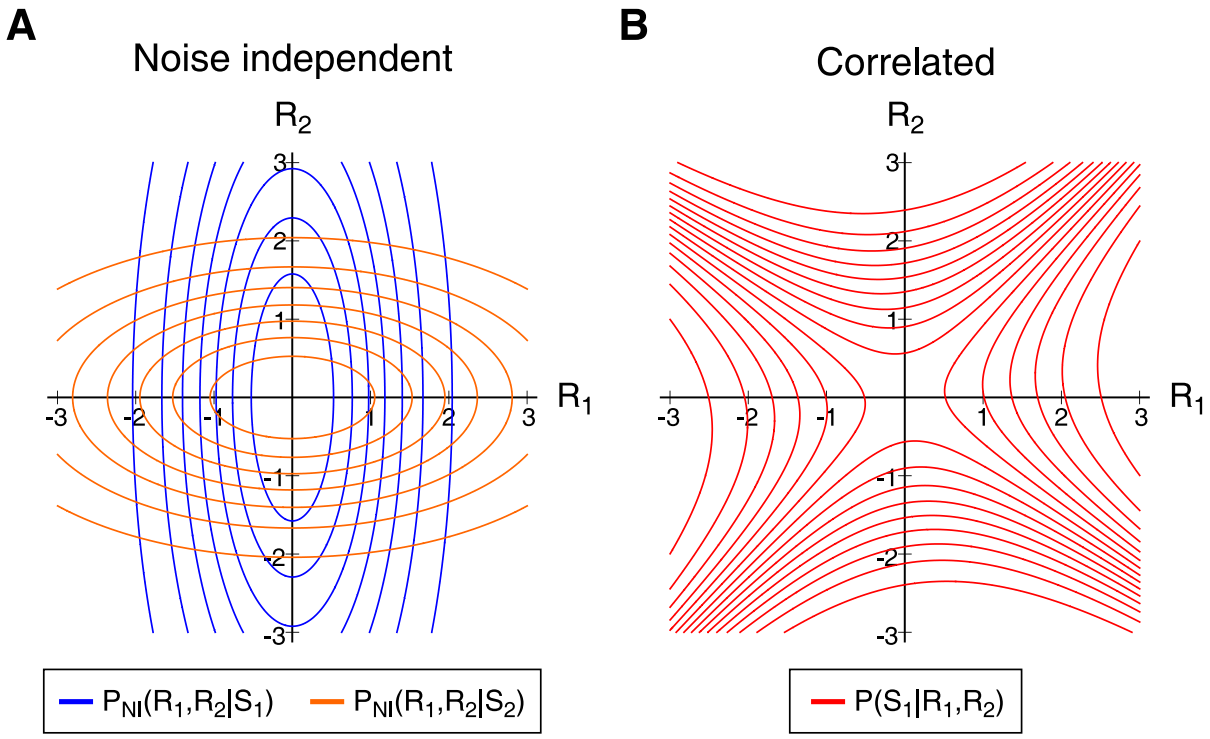

Figure S-2. Example in which noise correlations are important for decoding. *A*, Contour curves of noise-independent likelihoods $P_{NI}(R_1, R_2|S_1)$ and $P_{NI}(R_1, R_2|S_2)$. *B*, Contour curves of posterior probability $P(S_1|R_1, R_2)$ (which coincide with the contour curves of $P(S_2|R_1, R_2)$ defined by $P(S_2|R_1, R_2) = 1 - P(S_1|R_1, R_2)$). Response parameters are: $\boldsymbol{\mu}_1 = [0, 0]$, $\boldsymbol{\mu}_2 = [0, 0]$, $\sigma_{11} = 1$, $\sigma_{21} = 3$, $\sigma_{12} = 2$, $\sigma_{22} = 1$, $\rho_1 = 0$ and $\rho_2 = 0.3$. These response parameters do not comply with Eq. S-52c.

$$\left(\frac{\tilde{R}_{A1}}{\tilde{\sigma}_{12}}\right)^2 + \left(\frac{\tilde{R}_{A2}+1}{\tilde{\sigma}_{22}}\right)^2 = \left(\frac{\tilde{R}_{B1}}{\tilde{\sigma}_{12}}\right)^2 + \left(\frac{\tilde{R}_{B2}+1}{\tilde{\sigma}_{22}}\right)^2 \qquad \text{for } S_2\,. \tag{S-55}$$

Subtracting Eq. S-54 from Eq. S-55 and solving for $\tilde{R}_{A2}$ result in

$$\tilde{R}_{A2} = \tilde{R}_{B2} + \frac{{\tilde{R}_{A1}}^2 - {\tilde{R}_{B1}}^2}{4}\left(\frac{{\tilde{\sigma}_{21}}^2}{{\tilde{\sigma}_{11}}^2} - \frac{{\tilde{\sigma}_{22}}^2}{{\tilde{\sigma}_{12}}^2}\right). \tag{S-56}$$

Replacing Eq. S-56 into any of the Eqs. S-54 or S-55 results in a quartic equation in $\tilde{R}_{A1}$, whereas replacing Eq. S-56 into Eq. S-51 leads to a cubic equation in $\tilde{R}_{A1}$. Therefore, to fulfill Eq. S-51 it is necessary (though not sufficient) that the quartic equation has less than four real solutions. This requires that the variances comply with the condition

$$\frac{\tilde{\sigma}_{21}}{\tilde{\sigma}_{11}} = \frac{\tilde{\sigma}_{22}}{\tilde{\sigma}_{12}}, \tag{S-57}$$

which means that the response distributions associated with each stimulus have the same aspect ratio. In this case, the Eqs. S-54 and S-55 have the two following solutions

(1) $\tilde{R}_{A1} = \tilde{R}_{B1}$ and $\tilde{R}_{A2} = \tilde{R}_{B2}$

(2) $\tilde{R}_{A1} = -\tilde{R}_{B1}$ and $\tilde{R}_{A2} = \tilde{R}_{B2}$

which are symmetric with respect to the line $R_1 = 0$. After the change of variables of Eqs. S-53 and replacing by solution (2), Eq. S-51 becomes

$$\frac{\tilde{\rho}_2}{\det\left(\mathbf{C}_2\right)}\left(\tilde{R}_{A2}+1\right) = \frac{\tilde{\rho}_1}{\det\left(\mathbf{C}_1\right)}\left(\tilde{R}_{A2}-1\right), \tag{S-58}$$

and solving for $\tilde{R}_{A2}$ we obtain

$$\tilde{R}_{A2} = \frac{\tilde{\rho}_1\ \det\left(\mathbf{C}_2\right) + \tilde{\rho}_2\ \det\left(\mathbf{C}_1\right)}{\tilde{\rho}_1\ \det\left(\mathbf{C}_2\right) - \tilde{\rho}_2\ \det\left(\mathbf{C}_1\right)}. \tag{S-59}$$

As a result, Eq. S-51 holds only for one among infinite possible values of $\tilde{R}_{A2}$. Noise correlations are therefore always crucial for decoding. Notice that this result differs from Eq. 43b in Eyherabide and Samengo (2013). That result is incorrect, as stated in Section H.3.

E.3 Case III: $\mu_{11} \neq \mu_{12}$ and $\mu_{21} \neq \mu_{22}$

The analysis can be simplified by centering and scaling the distributions through the following change of variables

$$\tilde{R}_n = \frac{R_n - (\mu_{n1}+\mu_{n2})/2}{(\mu_{n1}-\mu_{n2})/2} \tag{S-60}$$

for each neuron n, so that in the new variables, the response distributions are located at $\tilde{\boldsymbol{\mu}}_{\mathbf{1}} = [1,1]$ and $\tilde{\boldsymbol{\mu}}_{\mathbf{2}} = [-1,-1]$ with variances $\tilde{\sigma}_{nk} = 2\,\sigma_{nk}/(\mu_{n1}-\mu_{n2})$ for each neuron n and stimulus S_k. As a result, Eqs. S-44 for both stimuli becomes

$$\left(\frac{\tilde{R}_{A1}-1}{\tilde{\sigma}_{11}}\right)^2 + \left(\frac{\tilde{R}_{A2}-1}{\tilde{\sigma}_{21}}\right)^2 = \left(\frac{\tilde{R}_{B1}-1}{\tilde{\sigma}_{11}}\right)^2 + \left(\frac{\tilde{R}_{B2}-1}{\tilde{\sigma}_{21}}\right)^2 \qquad \text{for } S_1 \tag{S-61}$$

$$\left(\frac{\tilde{R}_{A1}+1}{\tilde{\sigma}_{12}}\right)^2+\left(\frac{\tilde{R}_{A2}+1}{\tilde{\sigma}_{22}}\right)^2 = \left(\frac{\tilde{R}_{B1}+1}{\tilde{\sigma}_{12}}\right)^2+\left(\frac{\tilde{R}_{B2}+1}{\tilde{\sigma}_{22}}\right)^2 \qquad \text{for } S_2\,. \tag{S-62}$$

Analogously to the previous case, by subtracting the first equation from the second and solving for $\tilde{R}_{A2}$ we arrive at

$$\tilde{R}_{A2}=\tilde{R}_{B2}-\frac{\tilde{R}_{A1}-\tilde{R}_{B1}}{2}\left(\frac{\tilde{\sigma}_{21}{}^2}{\tilde{\sigma}_{11}{}^2}+\frac{\tilde{\sigma}_{22}{}^2}{\tilde{\sigma}_{12}{}^2}\right)+\frac{\tilde{R}_{A1}{}^2-\tilde{R}_{B1}{}^2}{4}\left(\frac{\tilde{\sigma}_{21}{}^2}{\tilde{\sigma}_{11}{}^2}-\frac{\tilde{\sigma}_{22}{}^2}{\tilde{\sigma}_{12}{}^2}\right). \tag{S-63}$$

Replacing Eq. S-63 into any of the Eqs. S-61 or S-62 results in a quartic equation in $\tilde{R}_{A1}$, whereas replacing Eq. S-63 into Eq. S-51 leads to a cubic equation in $\tilde{R}_{A1}$. Therefore, to fulfill Eq. S-51 it is necessary (though not sufficient) that the quartic equation has less than four real solutions. This requires that the variances comply with the condition

$$\frac{\tilde{\sigma}_{21}}{\tilde{\sigma}_{11}}=\frac{\tilde{\sigma}_{22}}{\tilde{\sigma}_{12}}\,, \tag{S-64}$$

which means that the response distributions associated with each stimulus have the same aspect ratio. In this case, the Eqs. S-61 and S-62 have the two following solutions

(1) $\tilde{R}_{A1}=\tilde{R}_{B1}$ and $\tilde{R}_{A2}=\tilde{R}_{B2}$

(2) $\tilde{R}_{A1}=\gamma\,\tilde{R}_{B1}+(1-\gamma)\,\tilde{R}_{B2}$ and $\tilde{R}_{A2}=(1+\gamma)\,\tilde{R}_{B1}-\gamma\,\tilde{R}_{B2}$

where $\gamma=\dfrac{\tilde{\sigma}_{22}{}^2-\tilde{\sigma}_{12}{}^2}{\tilde{\sigma}_{22}{}^2+\tilde{\sigma}_{12}{}^2}$.

After the change of variables of Eq. S-60 and the substitution of the solution (2), Eq. S-51 becomes

$$(\tilde{\sigma}_{12}-\tilde{\sigma}_{22})\,(\tilde{R}_{B1}-\tilde{R}_{B2})\left[\left(\frac{\tilde{R}_{B1}}{\tilde{\sigma}_{12}{}^2}+\frac{\tilde{R}_{B2}}{\tilde{\sigma}_{22}{}^2}\right)(\beta_{12}-\beta_{21})+\left(\frac{1}{\tilde{\sigma}_{12}{}^2}+\frac{1}{\tilde{\sigma}_{22}{}^2}\right)(\beta_{12}+\beta_{21})\right]=0\,, \tag{S-65}$$

where $\beta_{jk}=\rho_j\,(1-\rho_k{}^2)\,\tilde{\sigma}_{1k}\,\tilde{\sigma}_{2k}$. This equation holds for all possible values of $\mathbf{R}_B$ if and only if $\tilde{\sigma}_{22}=\tilde{\sigma}_{12}$. Therefore, in the original variables, noise correlations are almost always crucial for decoding except when the following condition holds

$$\frac{\sigma_{21}}{\sigma_{11}}=\frac{\sigma_{22}}{\sigma_{12}}=\frac{\mu_{11}-\mu_{12}}{\mu_{21}-\mu_{22}}\,. \tag{S-66}$$

An example in which noise correlations are irrelevant for decoding is shown in Figure S-3. In this example, the parameters of the response distributions comply with Eq. S-66. Each contour curve of the NI likelihood associated with stimulus S_1 (blue curves) may intersect each contour of the NI likelihood curve associated with stimulus S_2 (orange curves) in up to two population responses. Responses lying in the intersections of a pair of contour curves are symmetric with respect to a diagonal line passing through the origin of coordinates ($R_2 = R_1$; panel A). Because these responses have the same NI likelihood, they are merged after the NI assumption, and consequently information may be lost. Such an information loss does not occur, however, because these merged responses are non-informative. Analogously to the responses merged after the NI assumption, the contour curves of the posterior probabilities $P(S_1|R_1, R_2)$ and $P(S_2|R_1, R_2)$ are also symmetric with respect to a diagonal line passing through the origin of coordinates (panel B), and therefore, the posterior probabilities of the merged responses comply with Eq. S-45.

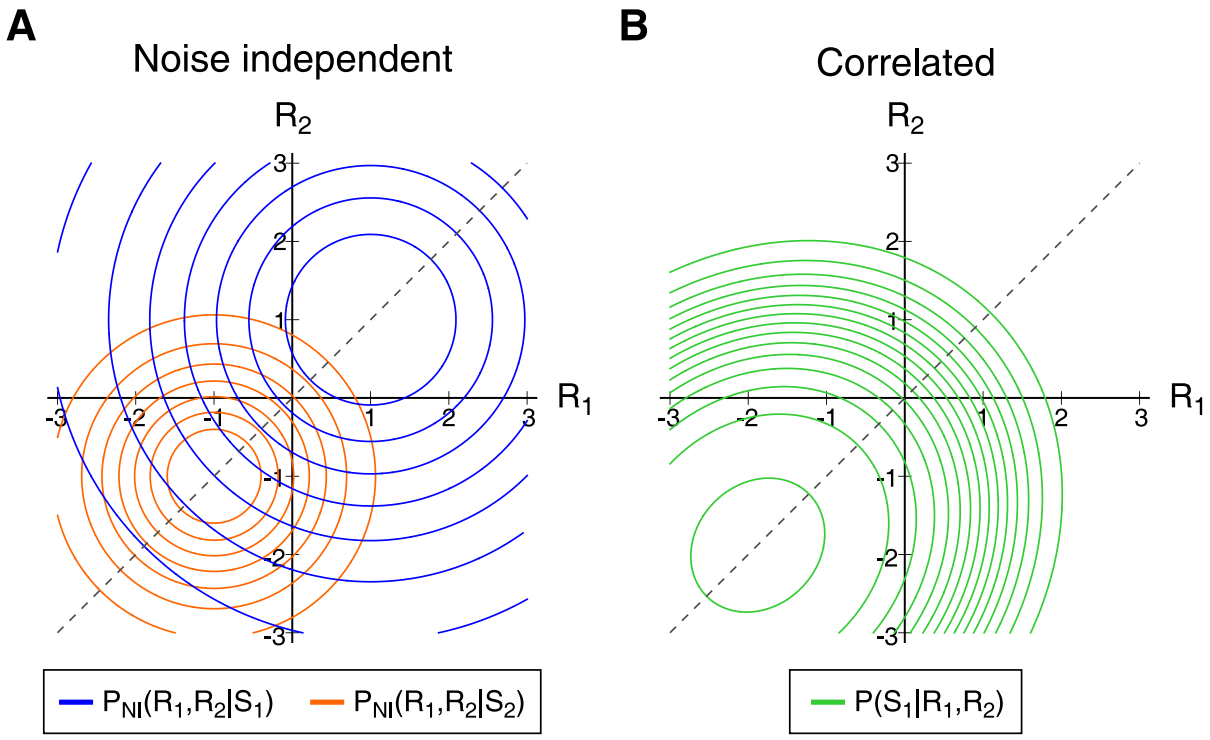

Figure S-3. Example in which noise correlations are irrelevant for decoding. *A*, Contour curves of NI likelihoods $P_{NI}(R_1, R_2|S_1)$ and $P_{NI}(R_1, R_2|S_2)$. *B*, Contour curves of posterior probability $P(S_1|R_1, R_2)$ (which coincide with the contour curves of $P(S_2|R_1, R_2)$ defined by $P(S_2|R_1, R_2) = 1 - P(S_1|R_1, R_2)$). Response parameters are: $\boldsymbol{\mu}_1 = [1, 1]$, $\boldsymbol{\mu}_2 = [-1, -1]$, $\sigma_{11} = \sigma_{12} = \sigma_{21} = \sigma_{22} = 1$, $\rho_1 = -0.1$ and $\rho_2 = 0.1$. These response parameters comply with Eq. S-66.

Noise correlations are important for decoding whenever the parameters of the response distributions do not comply with Eq. S-66. An example is shown in Figure S-4. This example differs from

the one shown in Figure S-3 solely in the value of the standard deviation σ_{11}. The NI likelihood associated with stimulus S_1 is therefore stretched along the horizontal axis, and responses merged after the NI assumption are not symmetric with respect to any straight line (panel A). To show that merged responses are informative, consider the pair of responses denoted by black dots in panel A. These responses have the same NI likelihoods (they lie in the intersections of a pair of contour curves of NI likelihoods), and are therefore merged after the NI assumption. As shown in panel B, these responses (black dots) lie near different contour curves of the posterior probabilities. Most of the responses merged after the NI assumption are informative, and merging them induces an information loss.

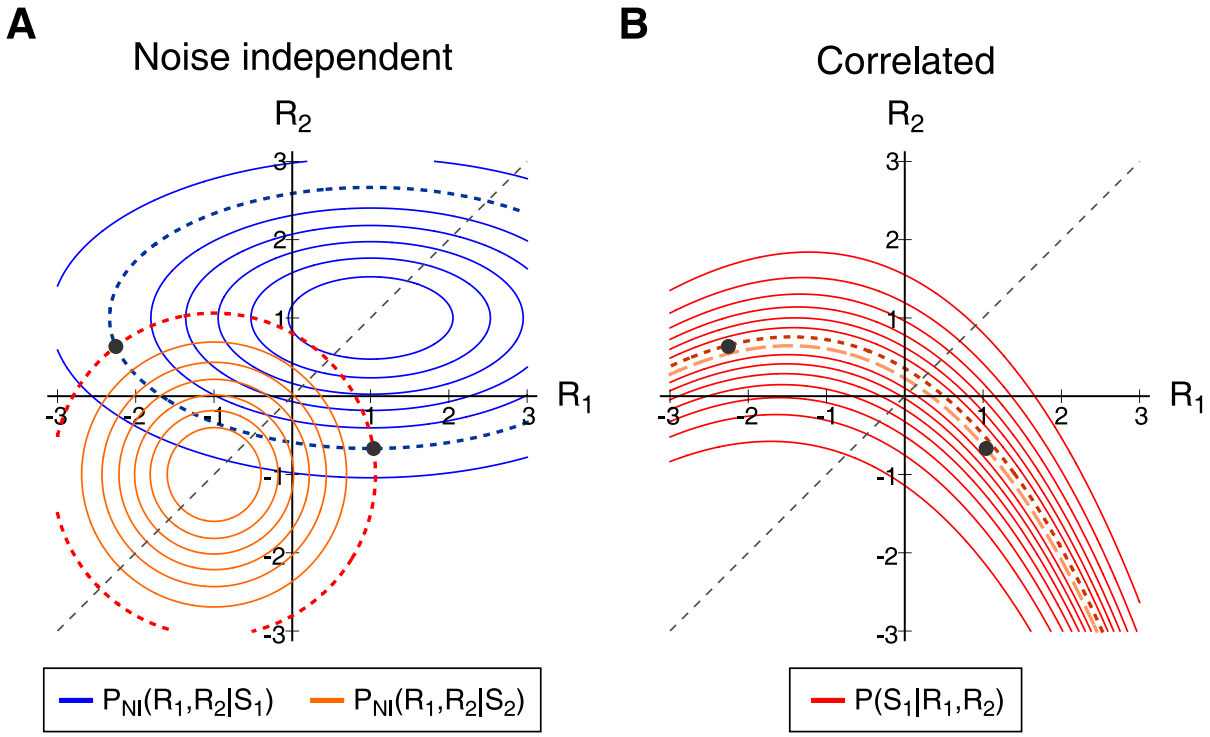

Figure S-4. Example in which noise correlations are irrelevant for decoding. *A*, Contour curves of NI likelihoods $P_{NI}(R_1, R_2|S_1)$ and $P_{NI}(R_1, R_2|S_2)$. *B*, Contour curves of posterior probability $P(S_1|R_1, R_2)$ (which coincide with the contour curves of $P(S_2|R_1, R_2)$ defined by $P(S_2|R_1, R_2) = 1 - P(S_1|R_1, R_2)$). Response parameters are: $\boldsymbol{\mu}_1 = [1, 1]$, $\boldsymbol{\mu}_2 = [-1, -1]$, $\sigma_{11} = 2$, $\sigma_{21} = \sigma_{12} = \sigma_{22} = 1$, $\rho_1 = -0.1$ and $\rho_2 = 0.1$. These response parameters do not comply with Eq. S-66.

An additional example in which noise correlations are irrelevant for decoding is shown in Figure S-5. In this example, the parameters of the response distributions comply with Eq. S-66, but they differ from the response parameters used in Figure S-3 in the mean values $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, and in the standard deviations σ_{11} and σ_{12}. Each contour curve of the NI likelihood associated with stimulus S_1 (blue curves) may intersect each contour of the NI likelihood curve associated with stimulus S_2

(orange curves) in up to two population responses. Responses merged after the NI assumption are not symmetric with respect to any straight line. Merging these responses induces no information loss, however, because these responses are non-informative. Consider for example the responses denoted by black dots in panel A. These responses have the same NI likelihoods and are therefore merged after the NI assumption. Similarly, these responses have the same posterior probabilities (they lie near the same contour curve of the posterior probabilities; panel B). The posterior probabilities of the merged responses comply with Eq. S-45 and therefore are non-informative.

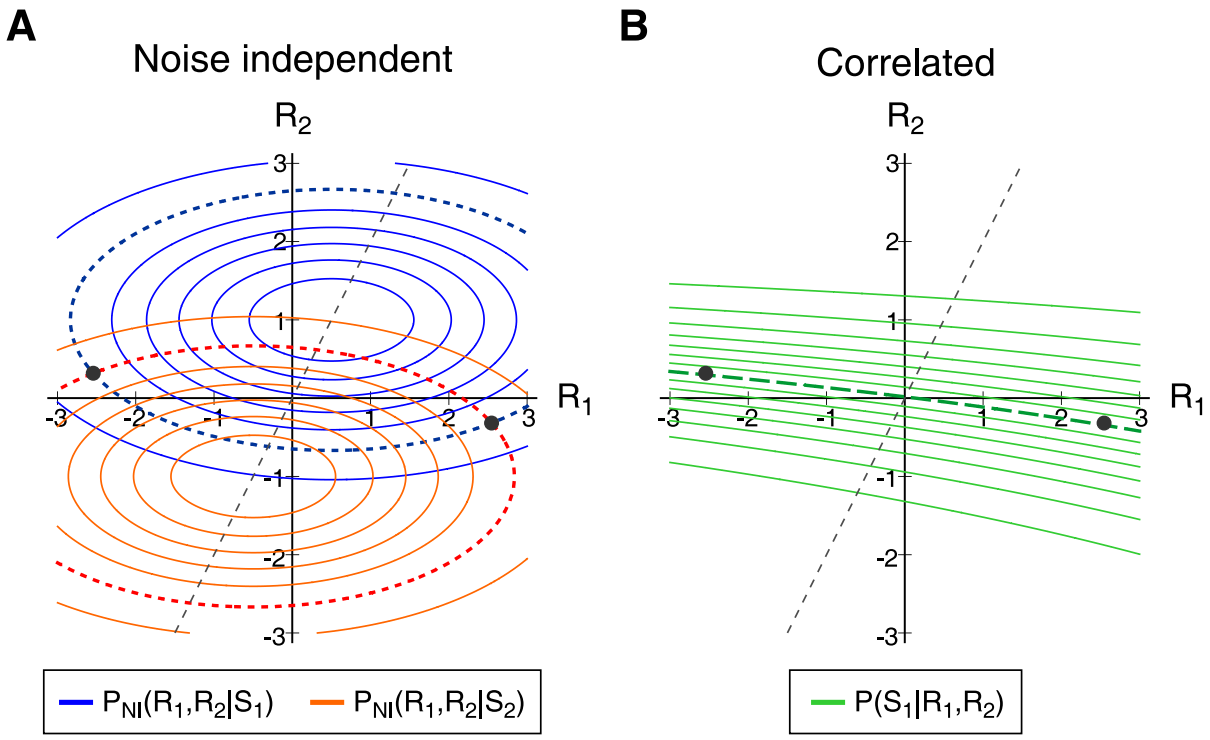

Figure S-5. Example in which noise correlations are irrelevant for decoding. *A*, Contour curves of NI likelihoods $P_{NI}(R_1, R_2|S_1)$ and $P_{NI}(R_1, R_2|S_2)$. *B*, Contour curves of posterior probability $P(S_1|R_1, R_2)$ (which coincide with the contour curves of $P(S_2|R_1, R_2)$ defined by $P(S_2|R_1, R_2) = 1 - P(S_1|R_1, R_2)$). Response parameters are: $\boldsymbol{\mu}_1 = [1, 1]$, $\boldsymbol{\mu}_2 = [-1, -1]$, $\sigma_{11} = \sigma_{12} = 2$, $\sigma_{21} = \sigma_{22} = 1$, $\rho_1 = -0.1$ and $\rho_2 = 0.1$. These response parameters comply with Eq. S-66.

F Noise correlations are almost always irrelevant when decoding discrete responses

Consider a finite population of N (>1) neurons $[R_1, \ldots, R_N]$ and a finite set of K (>1) stimuli $\{S_1, \ldots, S_K\}$. The population response is represented as a vector $\mathbf{R} = [R_1, \ldots, R_N]$. Each stimulus S_k occurs with probability $P(S_k)$ and elicits one population response $\mathbf{R}$ among a finite set of M_k

population responses $\{\mathbf{R}_1^k, \ldots, \mathbf{R}_{M_k}^k\}$ with probability $P(\mathbf{R}|S_k)$. We define a case or example of a stimuli-response mapping with K stimuli and N neurons as a set C_{NK} of probabilities

$$C = \left\{P(S_1), \ldots, P(S_K), P(\mathbf{R}_1^1|S_1), \ldots, P(\mathbf{R}_{M_1}^1|S_1), \ldots, \ldots, P(\mathbf{R}_1^K|S_K), \ldots, P(\mathbf{R}_{M_K}^K|S_K)\right\}. \quad \text{(S-67)}$$

The set C defines the probabilities with which all stimuli and responses occur in a case or example. A valid case must comply with the normalization constraints of the probabilities, that is

$$P(S_K) \geq 0 \qquad \text{and} \qquad \sum_{S_k} P(S_K) \quad = 1 \quad \text{(S-68a)}$$

$$P(\mathbf{R}|S_k) \geq 0 \qquad \text{and} \qquad \sum_{\mathbf{R}} P(\mathbf{R}|S_k) = 1 \text{ for every } k\,. \quad \text{(S-68b)}$$

For example, the stimulus-response mapping shown in Figure S-6 where stimuli are equally likely and responses to each stimulus are equally likely constitutes a valid case.

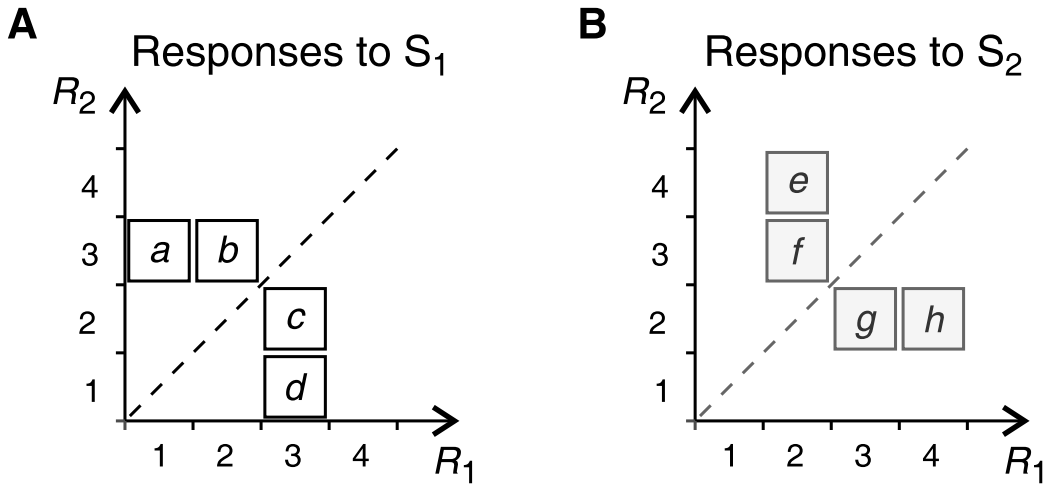

Figure S-6. Example of a population of two neurons R_1 and R_2 in which responses are elicited by two stimuli S_1 (panel A) and S_2 (panel B). Symbols inside each square represent the conditional probability $P(R_1, R_2|S_k)$, k being 1 or 2.

For each stimulus S_k, Eq. S-68a defines a simplex $\mathbf{C}^S$ in $\mathbb{R}^K$ (the convex flat area which vertexes have only one non-zero component equal to unity in real space of dimension K). The simplex $\mathbf{C}^S$ is of dimension $K-1$, and each point of the simplex represents the values the vector of probabilities $[P(S_1, \ldots, P(S_K)]$ can take. Analogously, for each stimulus S_k Eq. S-68b defines a simplex $\mathbf{C}^k$ in $\mathbb{R}^{M_k}$. Each simplex $\mathbf{C}^k$ is of dimension $M_k - 1$, and each point of the simplex represents the values the

vector of probabilities $[P(\mathbf{R}_1^k|S_k), \ldots, P(\mathbf{R}_{M_k}^k|S_k)]$ can take. The set of all valid cases $\mathbf{C}$ is given by

$$\mathbf{C} = \mathbf{C}^S \times \mathbf{C}^1 \times \ldots \times \mathbf{C}^K \,, \tag{S-69}$$

and it also constitutes a simplex in $\mathbb{R}^\eta$, where $\eta = \sum_k M_k$. For example, in Figure S-6, $\mathbf{C}^S$ is a segment connecting the points $[1, 0]$ and $[0, 1]$ in $\mathbb{R}^2$, $\mathbf{C}^1$ and $\mathbf{C}^2$ are regular tetrahedrons in $\mathbb{R}^4$, and $\mathbf{C}$ is a region of dimension 7 in $\mathbb{R}^8$.

For noise correlations to be important for decoding, it is necessary (but not sufficient) that at least one pair of population responses $\mathbf{R}_A$ and $\mathbf{R}_B$, for which $P(\mathbf{R}_A) > 0$ and $P(\mathbf{R}_B) > 0$, have the same NI likelihoods (Condition 24a in Eyherabide and Samengo, 2013). This condition implies that

$$\prod_{n=1}^{N} P(R_{An}|S_k) - \prod_{n=1}^{N} P(R_{Bn}|S_k) = 0 \,. \tag{S-70}$$

Here, $P(R_{An}|S_k)$ is the marginal conditional probability of neuron R_n given stimulus S_k evaluated for the component n of response $\mathbf{R}_A$, and is given by

$$P(R_{An}|S_k) = \sum_{\mathbf{R}\setminus R_n} P(R_1, \ldots, R_{n-1}, R_{An}, R_{n+1}, \ldots, R_N|S_k) \,, \tag{S-71}$$

where the sum extends over all responses of all neurons except neuron R_n. The marginal conditional probability $P(R_{Bn}|S_k)$ is defined in analogous manner. By replacing Eq. S-71 into Eq. S-70 we find that the equality of NI likelihoods can be expressed as a polynomial equation of degree up to N. The degree of the polynomial depends on the responses $\mathbf{R}_A$ and $\mathbf{R}_B$. For example, in Figure S-6, when $\mathbf{R}_A = [R_1 = 1, R_2 = 3]$ and $\mathbf{R}_B = [2, 3]$, $P(R_{A2}|S_1) = P(R_{B2}|S_1)$, and therefore Eq. S-70 is a polynomial of degree 1. The same occurs when $\mathbf{R}_A = [3, 1]$ and $\mathbf{R}_B = [3, 2]$. For any other pair of response, however, Eq. S-70 is a polynomial of degree 2.

Eq. S-70 defines a surface of dimension $M_k - 1$ in $\mathbb{R}^{M_k}$ (the same dimensionality as $\mathbf{C}^k$), but not all the points in this surface correspond to probabilities satisfying Eqs. S-68a and S-68b. For each stimulus S_k, Eqs. S-68b and S-70 are fulfilled simultaneously only for a subset of cases $\mathbf{C}_{NI}^k$ of dimension less than $M_k - 1$. Otherwise, there would exist a region in which the gradient of the surfaces defined by both equations coincide. This cannot occur, however, for the gradient of a simplex is constant

whereas the gradient of the surface defined by of Eq. S-70 is not. When Eq. S-70 is a polynomial of degree >1, the gradient is not constant but depends on the conditional probabilities of responses given stimulus S_k. When Eq. S-70 is a polynomial of degree equal to unity, the gradient is not constant but each component varies in sign depending on whether the derivative in each component is taken with respect to a conditional response probability involving a component of $\mathbf{R}_A$ or $\mathbf{R}_B$, and zero otherwise. Hence, the dimension of $\mathbf{C}_{NI}^k$ must be $M_k - 2$ or less, whereas the dimension of $\mathbf{C}^k$ is $M_k - 1$. This result occurs for all stimuli S_k. Assuming that the system of equations defined by Eq. S-70 for all stimuli has $\tilde{K}$ independent equations ($K \geq \tilde{K} \geq 1$), the set $\mathbf{C}_{NI}$ of all valid cases where noise correlations are irrelevant is a bounded region in $\mathbb{R}^{\eta_{NI}}$, where $\eta_{NI} \leq \sum_k M_k - \tilde{K}$. Compared to the set $\mathbf{C}$ of all valid cases, the set $\mathbf{C}_{NI}$ is therefore of measure zero (using a counting measure). Intuitively, the ratio between the number of cases where noise correlations are important and the total number of valid cases is zero.

Though the example shown in Figure S-6 uses a square lattice for quantization of the response, the demonstration is valid for any type of lattice. The demonstration, however, relies on the assumption that probabilities $P(\mathbf{R}|S_k)$ are only constrained by Eq. S-68b. Additional constraints may reduce the dimensionality of the surface defined by Eq. S-68b and result in a set of cases where noise correlations are important with the same dimensionality as the set of all valid examples. Consider the example shown in Figure S-6. The neural population consists of two neurons R_1 and R_2. Population responses are elicited by two different stimuli S_1 (panel A) and S_2 (panel B) with specific probabilities (symbol inside squares). Consider first that probabilities are only constrained by Eq.S-68b. Noise correlations are important for decoding if responses $[R_1, R_2] = [2, 3]$ and $[3, 2]$ are merged after the NI assumption, which occurs whenever Eq. S-70 is fulfilled, that is, when both of the following equations hold

$$c = \frac{a+b}{1-a-b}\, b \tag{S-72a}$$

$$g = \frac{e+f}{1-e-f}\, f\,. \tag{S-72b}$$

Out of an infinite number of values that c and g can take, only for one are noise correlations important. The ratio between the number of cases in which noise correlations are important and the total number of cases is zero. In other words, the set of cases where noise correlations are important is of measure zero.

Instead, consider now that probabilities are symmetric with respect to the diagonal line ($R_2 = R_1$), that is, $a = d$, $b = c$, $e = h$, and $f = g$. Under this constraint, Eq. S-70 holds for any value of the probabilities, and all possible cases are also cases where noise correlations are important. In other words, the ratio between the number of cases in which noise correlations are important and the total number of cases is unity. This indicates that the frequency of occurrence of cases in which noise correlations are important depends on the type of cases being analysed, where the type is defined by the additional constraints imposed over the probability distributions of the stimuli and the population responses.

We have here assumed that noise correlations are irrelevant if and only if the minimum information loss attainable by NI decoders is strictly zero, and crucial otherwise. In general, noise correlations have been assumed to play a minor role when the information loss was <10% of the encoded information. If reasons exist to set a threshold for the importance of noise correlations higher than the one we used (like 10% instead of 0%) then the number of cases where noise correlations are important are even fewer than those here reported (which is already infinitesimally small). The present discussion, however, does not take into account neither what type of stimuli the information loss represents nor the consequences of making decoding errors from a biological perspective (or any other perspective other than communication).

G Codes

The programs provided here are intended to reproduce the results and figure shown in Eyherabide and Samengo (2013). In order to run the programs, remember to set the current working folder in Matlab to the folder containing the files. Should you use this code, we kindly request you to cite the aforementioned publication. Should you find bugs, please contact either Prof. Inés Samengo (samengo at cab.cnea.gov.ar) or Hugo Gabriel Eyherabide (hugo.eyherabide at helsinki.fi).

G.1 License and copyright

G.2 Codes for constructing figures

The codes for constructing the figures are the following

1) **figure1g.m:** Constructs panek G in Figure 1.

2) **figure1h.m:** Constructs panek H in Figure 1.

3) **figure1i.m:** Constructs panek I in Figure 1.

4) **figure1j.m:** Constructs panek J in Figure 1.

5) **figure1k.m:** Constructs panek K in Figure 1.

6) **figure1l.m:** Constructs panek L in Figure 1.

7) **figure6a.m:** Constructs panek A in Figure 6.

8) **figure6b.m:** Constructs panek B in Figure 6.

The files require no input arguments, and provide as the output the handle of the figure they create.

G.3 Codes for calculating information losses and decoding errors

The following codes are necessary for constructing the figures and provide additional features for those interested in experimenting with the results shown in Eyherabide and Samengo (2013)

1) **entropy.m:**	Calculates the entropy of a discrete probability distribution.
2) **information.m:**	Calculates the mutual information $I(S;\mathbf{R})$ when responses are *discrete*.
3) **informationgauss.m:**	Calculates the mutual information $I(S;\mathbf{R})$ when responses are *Gaussian distributed*.
4) **infoloss.m:**	Calculates the information loss induced by NI decoders when responses are discrete.
5) **infolossgauss.m:**	Calculates the information loss induced by NI decoders when responses are Gaussian distributed.
6) **decerr.m:**	Calculates the decoding error induced by NI decoders when responses are discrete.
7) **decerrgauss.m:**	Calculates the decoding error induced by NI decoders when responses are Gaussian distributed.
8) **plotf1p1.m:**	Estimates the minimum information loss attainable by NI decoders when responses are discrete as a function of $P(S_1)$ and creates a figure similar to Figure 1G.
9) **plotf1p1g.m:**	Estimates the minimum information loss attainable by NI decoders when responses are Gaussian distributed as a function of $P(S_1)$ and creates a figure similar to Figure 1I.
10) **plotf1plld s2.m:**	Estimates the minimum information loss attainable by NI decoders when responses are discrete as a function of $P(L, L\|S_2)$ and creates a figure similar to Figure 1J.
11) **plotf1prho2g.m:**	Estimates the minimum information loss attainable by NI decoders when responses are Gaussian distributed as a function of ρ_2 and creates a figure similar to Figure 1L.

Additional details including input arguments and examples can be found inside each file.

H CORRIGENDA

H.1 Parameter value in Figure 1L

At the end of the caption of Figure 1 it was stated that, in panel L, the stimulus probability was set to $P(S_1) = 0.2$. This is not correct. The actual stimulus probability was set to $P(S_1) = 0.1$.

H.2 Scales can caption in Figure 6

In Figure 6, the tick labels of the vertical axes are expressed in per-unit, not in percentage as the axes labels indicate. For example, ΔI_{NI}^{NIP} in per-unit and percentage units is defined as follows

$$\underbrace{\Delta I_{NI}^{NIP}\,[{}^{0}/_{1}] = \frac{\Delta I_{NI}^{NIP}}{I(\mathbf{R},S)}}_{\text{In per-unit}} \qquad \underbrace{\Delta I_{NI}^{NIP}\,[\%] = 100\,\frac{\Delta I_{NI}^{NIP}}{I(\mathbf{R},S)}}_{\text{In percentage}} \tag{S-73}$$

In addition, the caption states between the fifth and the sixth line that

> ***B*** shows the variation of the increment in the minimum decoding error $\Delta\xi_{NI}^{NIP}$ relative to the minimum decoding error $\xi^{Min}(\mathbf{R};S)$.

This is not correct. Instead, it should say

> ***B*** shows the variation of the increment in the minimum decoding error $\Delta\xi_{NI}^{NIP}$ relative to the minimum decoding error at chance level.

The minimum decoding error at chance level is here defined as $1 - \max_S\left\{P(S)\right\}$, and represents the minimum decoding error that would be achieved if all population responses were merged before the estimation stage. In that case, the estimation strategy that minimizes the decoding error consists in decoding the stimulus that occurs with the maximum probability. This generalization of chance level reduces to the usual definition of chance level when stimulus categories are balanced, and should also be taken into account when reading the sentence starting in the seventh line after Eq. 34.

All in all, Figure 6 should look like this

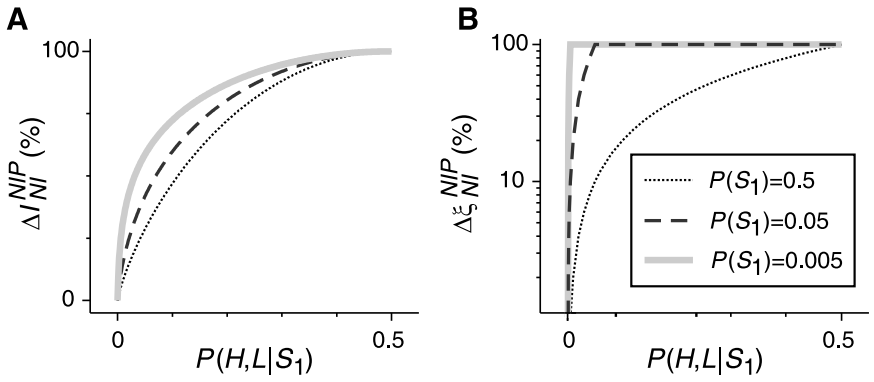

Figure 6. Difference between assessing the role of noise correlations using mutual information or decoding error. The population response shown in Figure 5C is decoded using the classical NI decoder. Response probabilities are set according to $P(H, L|S_1) = P(L, L|S_2)$. For different stimulus probabilities $P(S_1)$, panel ***A*** shows the variation of the minimum information loss ΔI_{NI}^{NIP} (relative to the encoded information $I(\mathbf{R}; S)$), and panel ***B*** shows the variation of the increment in the minimum decoding error $\Delta\xi_{NI}^{NIP}$ relative the minimum decoding error at chance level. 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 ΔI_{NI}^{NIP}, $\Delta\xi_{NI}^{NIP}$ approaches its maximum value when $P(H, L|S_1)$ is greater or equal to $P(S_1)$.

H.3 Correction in Eq. 43b and last paragraph of Results

In the last section of Results of Eyherabide and Samengo (2013), we erroneously concluded that, when the responses of two neurons elicited by two different stimuli have Gaussian distributions, noise correlations are irrelevant for decoding if the parameters of the response distributions comply with Eq. 43b. That is, noise correlations are irrelevant if $\mu_{11} = \mu_{12}$ or $\mu_{21} = \mu_{22}$ and

$$\frac{\sigma_{12}}{\sigma_{11}} = \frac{\sigma_{22}}{\sigma_{21}} = \sqrt{\frac{\rho_2\,(1-\rho_1^2)}{\rho_1\,(1-\rho_2^2)}}\,. \tag{S-74}$$

As we showed in Section E.2, this result is not correct, and not even under this conditions noise correlations become irrelevant. Further corrections in the last paragraph of the Results are necessary to take into account the elimination of Eq. 43b. This changes are stated later in this section.

This correction only strengthens our conclusions that noise correlations are almost always important when decoding the activity of two neurons with Gaussian conditional probability distributions $P(R_1, R_2|S)$ elicited by two different stimuli.

An example in which noise correlations are important for decoding even though response distributions comply with Eq. S-74 is shown in Figure S-7. Responses lying in the intersections between pairs of contour curves of NI likelihoods (panel A) are merged after the NI assumption. Such responses are symmetric with respect to the vertical axis ($R_1 = 0$). This symmetry, however, is not shared by the contour curves of the posterior probability, indicating that merged responses have different posterior probabilities. Therefore, most of the merged responses are informative (they do not comply with Eq. S-45), and noise correlations are important for decoding.

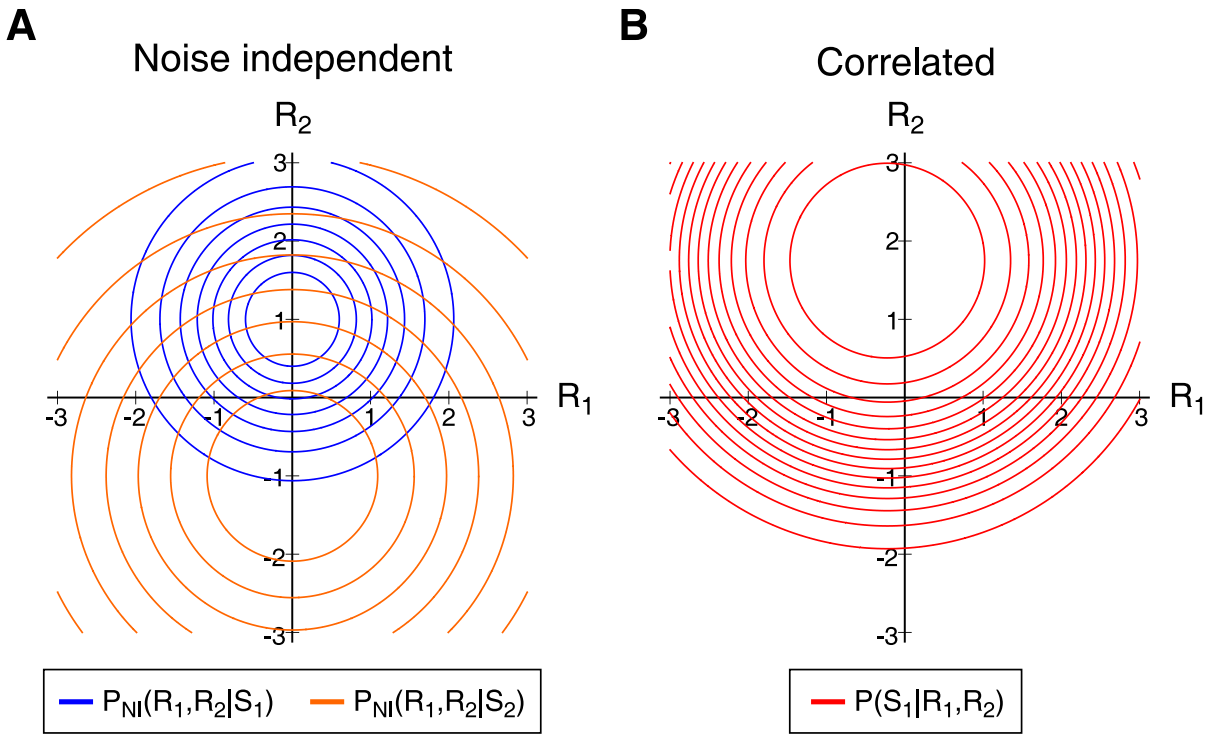

Figure S-7. Example that noise correlations are important for decoding even though response distributions comply with Eq. S-74. *A*, Contour curves of NI likelihoods $P_{NI}(R_1, R_2|S_1)$ and $P_{NI}(R_1, R_2|S_2)$. *B*, Contour curves of posterior probability $P(S_1|R_1, R_2)$ (which coincide with the contour curves of $P(S_2|R_1, R_2)$ defined by $P(S_2|R_1, R_2) = 1 - P(S_1|R_1, R_2)$). Response parameters are: $\boldsymbol{\mu}_1 = [0, 1]$, $\boldsymbol{\mu}_2 = [0, -1]$, $\sigma_{11} = \sigma_{21} = 1$, $\sigma_{12} = \sigma_{22} = 2$, $\rho_1 \approx 0.082$ and $\rho_2 = 0.3$.

In addition, the last sentence of the last paragraph of Results states that

> For any departure from conditions 43a-c, 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 approximately zero to 100 % (for example, when condition 43a holds and variances are equal).

As stated at the beginning of the sentence, we are referring to the cases where conditions 43a-c do not hold, and therefore the text inside the brackets is not correct. Instead, it should say

> For any departure from conditions 43a-c, 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 approximately zero to 100 % (for example, when condition 43a does not hold and variances are equal).

In summary, the original version of the manuscript states that

> Noise correlations are almost always important for decoding except when the following conditions are met
>
> (43a) $$\frac{\sigma_{12}\,\sigma_{22}}{\sigma_{11}\,\sigma_{21}} = \frac{\rho_2\,(1-\rho_1^2)}{\rho_1\,(1-\rho_2^2)}\,, \qquad \begin{array}{l}\text{if} \quad \mu_{11}=\mu_{12},\\ \text{and} \quad \mu_{21}=\mu_{22};\end{array}$$
>
> (43b) $$\frac{\sigma_{12}}{\sigma_{11}} = \frac{\sigma_{22}}{\sigma_{21}} = \sqrt{\frac{\rho_2\,(1-\rho_1^2)}{\rho_1\,(1-\rho_2^2)}}\,, \qquad \begin{array}{l}\text{if} \quad \mu_{11}=\mu_{12};\\ \text{or} \quad \mu_{21}=\mu_{22};\end{array}$$
>
> (43c) $$\frac{\sigma_{21}}{\sigma_{11}} = \frac{\sigma_{22}}{\sigma_{12}} = \frac{\mu_{11}-\mu_{12}}{\mu_{21}-\mu_{22}}\,, \qquad \begin{array}{l}\text{if} \quad \mu_{11}\neq\mu_{12},\\ \text{and} \quad \mu_{21}\neq\mu_{22};\end{array}$$
>
> Condition 43a, 43b, and 43c establishes 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. 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 $\sim$ 0 to 100 % (for example, when condition 43a holds and variances are equal).

This should be replaced by

Noise correlations are almost always important for decoding except when the following conditions are met

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

$$\text{(43b)}\quad \frac{\sigma_{21}}{\sigma_{11}} = \frac{\sigma_{22}}{\sigma_{12}} = \frac{\mu_{11}-\mu_{12}}{\mu_{21}-\mu_{22}}\,, \qquad \begin{array}{ll}\text{if} & \mu_{11}\neq\mu_{12},\\ \text{and} & \mu_{21}\neq\mu_{22};\end{array}$$

Condition 43a and 43b establishes 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. Condition 43a holds 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 constrains the shape of the contour curves, but holds for arbitrary correlation coefficients. For any departure from conditions 43a and 43b, 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 $\sim$ 0 to 100 % (for example, when condition 43a does not hold and variances are equal).

References

Cover TM, Thomas JA (1991) *Elements of information theory* Wiley-Interscience, New York, USA.

Eyherabide HG, Samengo I (2013) When and why noise correlations are important in neural decoding. *J Neurosci* 33:17921–17936.

Latham PE, Nirenberg S (2005) Synergy, redundancy, and independence in population codes, revisited. *J Neurosci* 25:5195–5206.

Nirenberg S, Carcieri SM, Jacobs AL, Latham PE (2001) Retinal ganglion cells act largely as independent decoders. *Nature* 411:698–701.

Nirenberg S, Latham PE (2003) Decoding neuronal spike trains: How important are correlations. *Proc Natl Acad Sci USA* 100:7348–7353.

Oizumi M, Ishii T, Ishibashi K, Hosoya T, Okada M (2010) Mismatched decoding in the brain. *J Neurosci* 30:4815–4826.