Optimal decoding of noisy neuronal populations using recurrent networks

In the main text we considered an example in which the goal was to estimate two quantities, the presentation angle and contrast, given the response of a population of noisy neurons. With a minor change in notation this can be generalized to the problem of estimating M quantities; the only real difference is that instead of labeling the neurons with two indices, i and j, we need M indices, $i_1, ..., i_M$, where the m^{th} index runs from 1 to N_m. In this case the response of the neurons is denoted $a_{i_1,...,i_M}$, and there are M "angles" to estimate, $\theta_1, ..., \theta_M$. We will let $\mathbf{a}$ denote the complete set of responses, θ the set of angles, and $P(\mathbf{a}|\theta)$ the conditional probability of observing $\mathbf{a}$ given the presentation angles, θ.

The network presented in the main text can be viewed as an algorithm for computing the presentation angles: given a set of responses, $\mathbf{a}$, the network finds a set of angles, $\hat{\theta}_k(\mathbf{a})$, $k = 1, ..., M$, that are estimates of the true angles, θ_k. How good is that algorithm? The answer depends on what we mean by "good". Here we use the determinant of the covariance matrix, denoted $\langle \delta\hat{\theta}_k \delta\hat{\theta}_l \rangle$ (defined immediately below), to measure the quality of the estimate. Our motivation for using this quantity is that it determines, to a large extent, the mutual information between the noisy neuronal responses, $\mathbf{a}$, and the presentation angles, θ: the smaller the log of the determinant of the covariance matrix, the larger the mutual information [1]. (Strictly speaking, this result applies only when the neurons are uncorrelated. We believe it applies also to correlated neurons, as long as the covariance matrix is small. In any case, it is a good starting point.) The components of the covariance matrix are given by

$$\langle \delta\hat{\theta}_k \delta\hat{\theta}_l \rangle \equiv \langle (\hat{\theta}_k(\mathbf{a}) - \bar{\theta}_k)(\hat{\theta}_l(\mathbf{a}) - \bar{\theta}_l) \rangle$$

where the angle brackets indicate an average with respect to the probability distribution $P(\mathbf{a}|\theta)$, and $\bar{\theta}_k \equiv \langle \hat{\theta}_k(\mathbf{a}) \rangle$ is the mean value of the estimate.

We now have two tasks: 1) compute the determinant of the covariance matrix, and 2) compare that determinant to the optimal one. Fortunately, one does not have to consider all possible estimators to find the optimal one; for unbiased estimators, the lower bound on the determinant is given by the inverse of the Fisher Information [2].

$$\det \langle \delta\hat{\theta} \delta\hat{\theta} \rangle \geq \frac{1}{\det I} \tag{1}$$

where I is the Fisher information,

$$I_{kl} = \left\langle -\frac{\partial^2}{\partial\theta_k \partial\theta_l} \log P(\mathbf{a}|\theta) \right\rangle . \tag{2}$$

Equation (1) is the multi-dimensional analog of the Cramér-Rao bound [2].

Our program now is to compute the covariance matrix associated with the network estimate, then compare that to the Cramér-Rao bound, Eq. (1). The networks we consider are those that asymptote to a smooth M-dimensional attractor, with the property that each point on the attractor is neutrally stable (i.e., we exclude such behavior as limit cycles). These networks are estimators in the following sense: Let the initial condition of the network be $\mathbf{a}$, which is generated from the conditional probability distribution, $P(\mathbf{a}|\theta)$. Given that initial condition, the network evolves in time and relaxes onto its M-dimensional attractor. The position on the attractor provides an estimate of the θ_k.

To see how this works in practice, we consider a set of network evolution equations of the form

$$\mathbf{o}(t+1) = \mathbf{H}(\mathbf{o}(t)) \tag{3}$$

where $\mathbf{o}$ and $\mathbf{H}$ have components $o_{i_1,\ldots,i_M}$ and $H_{i_1,\ldots,i_M}$, $\mathbf{H}$ is a nonlinear function of $\mathbf{o}$, and we interpret $\mathbf{o}$ as a set of firing rates. This equation is initialized via the relation

$$\mathbf{o}(0) = \mathbf{a}(\theta)\,. \tag{4}$$

The existence of the attractor implies that there is some smooth function, $\mathbf{g}(\phi)$, satisfying

$$\mathbf{g}(\phi) = \mathbf{H}(\mathbf{g}(\phi)) \tag{5}$$

where the ϕ are a set of M generalized angles. In the limit $t \to \infty$, $\mathbf{o}$ approaches the attractor; i.e., $\lim_{t\to\infty} \mathbf{o}(t) = \mathbf{g}(\hat{\theta})$. The point on the attractor, $\hat{\theta}$, is the network estimate of the set of presentation angles, θ. Note that the network we use in the main text can be cast into the form given in Eq. (3) by expressing u_{ij} in terms of o_{ij} via the first equation in the main text, and inserting the resulting expression into the second.

Because the initial condition is generated probabilistically, the estimate will be different on each trial. As just discussed, the quality of the estimate is related to its covariance matrix. To compute that covariance matrix, we take a perturbative approach: we iterate a linearized version of the full nonlinear network, Eq. (3), which allows us to determine analytically the approximate final position on the attractor given the initial condition. A difficulty arises because, unlike point (0-dimensional) attractors, it is not obvious *which* point on the M-dimensional attractor to linearize around. It turns out that we may determine the appropriate point by going past linear order in our perturbation expansion. Let us for now perturb around an arbitrary point, say $\mathbf{o} = \mathbf{g}(\phi)$, and later determine ϕ. Letting

$$\mathbf{o}(t) = \mathbf{g}(\phi) + \delta\mathbf{o}(t)\,, \tag{6}$$

through second order in $\delta\mathbf{o}(t)$, Eq. (3) becomes

$$\delta\mathbf{o}(t+1) = \mathbf{J}(\phi)\cdot\delta\mathbf{o}(t) + \frac{1}{2}\mathbf{H}''(\phi) : \delta\mathbf{o}(t)\delta\mathbf{o}(t) \tag{7}$$

where $\mathbf{J}$ is the Jacobian evaluated on the attractor,

$$J_{\mathbf{ij}}(\phi) \equiv \frac{\partial H_{\mathbf{i}}(\mathbf{g}(\phi))}{\partial g_{\mathbf{j}}(\phi)}\,,$$

the ":" notation is shorthand for a sum on two indices,

$$\mathbf{H}''(\phi) : \delta\mathbf{o}(t)\delta\mathbf{o}(t) \equiv \sum_{\mathbf{jk}} \frac{\partial^2 \mathbf{H}(\mathbf{g}(\phi))}{\partial g_{\mathbf{j}}(\phi)\partial g_{\mathbf{k}}(\phi)} \delta o_{\mathbf{j}} \delta o_{\mathbf{k}}\,,$$

and the indices have multiple components, $\mathbf{i} \equiv i_1, ..., i_M$. In Eq. (7), and in the remainder of the Appendix, we use standard dot-product notation; e.g., the $\mathbf{i}^{th}$ component of $\mathbf{J} \cdot \delta\mathbf{o}$ is $\sum_{\mathbf{j}} J_{\mathbf{ij}} \delta o_{\mathbf{j}}$.

The expansion, Eq. (7), is valid as long as the higher order nonlinearities are small compared to the quadratic term. In that case, Eq. (7) may be rewritten as

$$\delta\mathbf{o}(t) = \mathbf{J}^t \cdot \delta\mathbf{o}(0) + \frac{1}{2}\sum_{s=0}^{t-1} \mathbf{J}^{t-s-1} \cdot \mathbf{H}'' : (\mathbf{J}^s \cdot \delta\mathbf{o}(0))(\mathbf{J}^s \cdot \delta\mathbf{o}(0))\,. \tag{8}$$

Here the superscript t means multiply $\mathbf{J}$ by itself t times; it does *not* mean transpose. To avoid secularity — the last term in Eq. (8) going to infinity as t goes to infinity — we require that

$$\lim_{t\to\infty} \mathbf{J}^t(\phi) \cdot \delta\mathbf{o}(0) = 0\,. \tag{9}$$

Equation (9) is the condition that tells us what angle, ϕ, to linearize around. It also tells us that $\delta\mathbf{o}(\infty) = 0$, which in turn implies, using Eq. (6), that ϕ is the network estimate of the presentation angles, θ.

We can recast Eq. (9) into a form suitable for calculation by noting, via Eq. (5), that $\mathbf{J}(\phi)$ has M eigenvectors with eigenvalue 1; those eigenvectors are, $\mathbf{v}_k \equiv \partial_{\phi_k}\mathbf{g}, k = 1, ..., M$. Since we are assuming that $\mathbf{H}$ admits an attractor, all the other eigenvalues of $\mathbf{J}$ must be less than 1. Thus, in the limit that $t \to \infty$, $\mathbf{J}^t$ takes on a very simple form:

$$\lim_{t\to\infty} \mathbf{J}^t(\phi) = \sum_k \mathbf{v}_k(\phi)\mathbf{v}_k^\dagger(\phi)$$

where the $\mathbf{v}_k^\dagger(\phi)$ are the adjoint eigenvectors of the Jacobian. Using the orthogonality condition $\mathbf{v}_k^\dagger \cdot \mathbf{v}_l = \delta_{kl}$ (δ_{kl} is the Kronecker delta), Eq. (9) breaks into M equations, one for each k,

$$\mathbf{v}_k^\dagger(\phi) \cdot [\mathbf{a}(\theta) - \mathbf{g}(\phi)] = 0 \tag{10}$$

where we used Eqs. (4) and (6) to express $\delta\mathbf{o}(0)$ in terms of $\mathbf{a}(\theta)$ and $\mathbf{g}(\phi)$. The value of ϕ that satisfies Eq. (10) corresponds to the point on the attractor that we linearized around, and also to the network estimate of θ. Letting $\phi = \theta + \delta\theta$, which, term by term, means $\phi_k = \theta_k + \delta\theta_k$, and expanding Eq. (10) to first order in $\delta\theta$, we arrive at the set of equations

$$\mathbf{v}_k^\dagger(\theta) \cdot [\mathbf{a}(\theta) - \mathbf{g}(\theta)] + \sum_l \delta\theta_l \partial_{\phi_l} \left(\mathbf{v}_k^\dagger(\phi) \cdot [\mathbf{a}(\theta) - \mathbf{g}(\phi)] \right)_{\phi=\theta} = 0\,. \tag{11}$$

To proceed we define the noise, $\mathbf{N}(\theta)$, through the relationship

$$\mathbf{a}(\theta) = \mathbf{f}(\theta) + \mathbf{N}(\theta) \tag{12}$$

where

$$\mathbf{f}(\theta) \equiv \langle \mathbf{a} \rangle$$

is the mean value of the neuronal response given the set of presentation angles, θ. Inserting Eq. (12) into (11) yields

$$\mathbf{v}_k^\dagger(\theta) \cdot \mathbf{N}(\theta) + \mathbf{v}_k^\dagger(\theta) \cdot [\mathbf{f}(\theta) - \mathbf{g}(\theta)] + \sum_l \delta\theta_l \partial_{\phi_l} \left(\mathbf{v}_k^\dagger(\phi) \cdot [\mathbf{f}(\theta) + \mathbf{N}(\theta) - \mathbf{g}(\phi)] \right)_{\phi=\theta} = 0\,. \tag{13}$$

If the term $\mathbf{v}_k^\dagger(\theta) \cdot [\mathbf{f}(\theta) - \mathbf{g}(\theta)]$ does *not* vanish for all θ, then, for some θ, $\delta\theta$ will be nonzero in the limit that the noise goes to zero, and the network will produce biased estimates. Conversely, if it does vanish, then the network will be unbiased. We thus make the assumption that $\mathbf{v}_k^\dagger(\theta) \cdot [\mathbf{f}(\theta) - \mathbf{g}(\theta)] = 0, k = 1, ..., M\ \forall\theta$. If this condition is satisfied (and it is for the divisive normalization used in the main text), then Eq. (13) implies that, for small $\mathbf{N}$, $\delta\theta \sim \mathbf{N}$. Thus, the term $\delta\theta\partial_\phi \mathbf{v}_k^\dagger \cdot \mathbf{N}$ that appears in Eq. (13) is $\mathcal{O}(\mathbf{N}^2)$ and can be ignored. With this simplification, we find that $\delta\theta$ is given by

$$\delta\theta_k = \sum_l [\mathbf{v}_k^\dagger(\theta) \cdot \partial_{\theta_l} \mathbf{f}(\theta)]^{-1} \mathbf{v}_l^\dagger(\theta) \cdot \mathbf{N}(\theta)\,. \tag{14}$$

In this expression, and in what follows, we are using a shorthand notation for the inverse of a matrix: $[A_{kl}]^{-1} \equiv [A^{-1}]_{kl}$. Thus, $[\mathbf{v}_k^\dagger(\theta) \cdot \partial_{\theta_l} \mathbf{f}(\theta)]^{-1}$ is the kl^{th} component of the inverse of the matrix $\mathbf{v}_k^\dagger(\theta) \cdot \partial_{\theta_l} \mathbf{f}(\theta)$.

Using Eq. (14), it is now straightforward to compute the covariance matrix that determines the error in the estimate of the angles, and we find that

$$\langle \delta\theta_k \delta\theta_l \rangle = \left[\partial_{\theta_k} \mathbf{f}(\theta) \cdot \left(\sum_{ij} \mathbf{v}_i^\dagger(\theta) \, [\mathbf{v}_i^\dagger(\theta) \cdot \mathbf{R}(\theta) \cdot \mathbf{v}_j^\dagger(\theta)]^{-1} \, \mathbf{v}_j^\dagger(\theta) \right) \cdot \partial_{\theta_l} \mathbf{f}(\theta) \right]^{-1}$$

where $\mathbf{R}(\theta)$ is the noise covariance matrix,

$$\mathbf{R}(\theta) \equiv \langle \mathbf{N}(\theta)\mathbf{N}(\theta) \rangle \,.$$

Because we now have two covariance matrices, $\mathbf{R}$ and $\langle \delta\theta\delta\theta \rangle$, we will consistently refer to $\mathbf{R}$ as the *noise* covariance matrix and $\langle \delta\theta\delta\theta \rangle$ simply as the covariance matrix.

As discussed, above, decreasing the determinant of the covariance matrix increases the mutual information between the the neuronal responses and the presentation angles. A straightforward, but somewhat tedious, calculation shows that the determinant is minimized when

$$\mathbf{v}_k^\dagger \propto \mathbf{R}^{-1} \cdot \partial_{\theta_k} \mathbf{f} \,, \tag{15}$$

at which value of $\mathbf{v}_k^\dagger$ the covariance matrix simplifies to

$$\langle \delta\theta_k \delta\theta_l \rangle = \left[\partial_{\theta_k} \mathbf{f} \cdot \mathbf{R}^{-1} \cdot \partial_{\theta_l} \mathbf{f} \right]^{-1} \,.$$

Thus, whenever Eq. (15) is satisfied, the nonlinear recurrent network given in Eq. (3) leads to a covariance matrix such that

$$\det \langle \delta\theta\delta\theta \rangle = \frac{1}{\det \left[\partial_\theta \mathbf{f} \cdot \mathbf{R}^{-1} \cdot \partial_\theta \mathbf{f} \right]} \,. \tag{16}$$

This equations gives us the minimum determinant of the covariance matrix associated with the estimate produced by the network. To determine whether that minimum reaches the Cramér-Rao bound, we need to know the distribution of the noise; i.e., we need to know the explicit form of $P(\mathbf{a}|\theta)$. Let us consider two types of noise: Gaussian with an arbitrary correlation matrix, for which

$$P(\mathbf{a}|\theta) = \frac{\exp\left[\frac{1}{2} (\mathbf{a} - \mathbf{f}(\theta)) \cdot \mathbf{R}^{-1} \cdot (\mathbf{a} - \mathbf{f}(\theta)) \right]}{\sqrt{(2\pi)^N \det \mathbf{R}}} \,,$$

where $N \equiv \sum_m N_m$ is the total number of neurons, and Poisson with uncorrelated noise, for which

$$P(\mathbf{a}|\theta) = \prod_{\mathbf{i}} \frac{f_{\mathbf{i}}(\theta)^{a_{\mathbf{i}}} e^{-f_{\mathbf{i}}(\theta)}}{a_{\mathbf{i}}!} \,. \tag{17}$$

In the second expression, $\mathbf{a}$ and $\mathbf{f}$ now refer to the number of spikes in an interval rather than the firing rate. For the Poisson distribution the mean value of $\mathbf{a}$ is $\mathbf{f}$, and the noise covariance matrix is given by

$$\langle (a_{\mathbf{i}} - f_{\mathbf{i}})(a_{\mathbf{j}} - f_{\mathbf{j}}) \rangle_{\text{Poisson}} = f_{\mathbf{i}} \delta_{\mathbf{ij}} \equiv R_{\mathbf{ij}}$$

where the subscript "Poisson" indicates an average over the probability distribution given in Eq. (17). Note that we are using the symbol $\mathbf{R}$ for the noise covariance matrix of both the Gaussian and Poisson distributions; which distribution we mean should be clear from the context.

It is straightforward to show that the Fisher information, Eq. (2), for the two cases is given by

$$I_{kl,\text{Gaussian}} = \partial_{\theta_k} \mathbf{f} \cdot \mathbf{R}^{-1} \cdot \partial_{\theta_l} \mathbf{f} + \frac{1}{2} \text{tr}\{\mathbf{R}^{-1} \cdot \partial_{\theta_k} \mathbf{R} \cdot \mathbf{R}^{-1} \cdot \partial_{\theta_l} \mathbf{R}\} \quad (18)$$

$$I_{kl,\text{Poisson}} = \partial_{\theta_k} \mathbf{f} \cdot \mathbf{R}^{-1} \cdot \partial_{\theta_l} \mathbf{f} \quad (19)$$

where tr refers to the trace of a matrix. The trace term in Eq. (18) is a non-negative definite matrix with respect to the indices k and l, as is the first term on the right hand side of Eq. (18). Thus,

$$\det[I_{\text{Gaussian}}] \geq \det[\partial_\theta \mathbf{f} \cdot \mathbf{R}^{-1} \cdot \partial_\theta \mathbf{f}] \quad (20)$$

$$\det[I_{\text{Poisson}}] = \det[\partial_\theta \mathbf{f} \cdot \mathbf{R}^{-1} \cdot \partial_\theta \mathbf{f}] \,. \quad (21)$$

Equality is achieved in Eq. (20) only if the trace term vanishes, which happens only when $\mathbf{R}$ is independent of θ. Comparing Eqs. (16) and (20), we arrive at our final result:

1. For Gaussian noise with constant noise covariance matrix, $\mathbf{R}$, the minimum determinant of the covariance matrix achieved by the network is *equal* to the Cramér-Rao bound.

2. For Gaussian noise with a noise covariance matrix that depends on presentation angle, the minimum determinant of the covariance matrix achieved by the network *exceeds* the Cramér-Rao bound. The difference may be calculated by comparing Eqs. (16) and (18).

3. For uncorrelated Poisson noise, the minimum determinant of the covariance matrix achieved by the network is *equal* to the Cramér-Rao bound.

References

[1] N. Brunel and J.P. Nadal. Mutual information, Fisher information and population coding. *Neural Computation*, 10:1731–57, 1998.

[2] T.M. Cover and J.A. Thomas. *Elements of information theory.* John Wiley & Sons, New York, 1991.