Supplementary Mathematical Appendix

Contents

§1. Introduction

The main purpose of this Appendix is to delineate a simple theoretical framework that can reconcile two seemingly contradictory observations: (1) the existence of substantial variablity in single neuron coding properties from day to day (as evidenced, for example, by substantial day-to-day changes in single neuron d' values shown in **Extended Data Figs 2f,k**); and (2) the existence of a single decoder for each mouse that can stably decode across many days with a high d' value (as demonstrated in **Fig. 3c**). A key finding that can potentially reconcile the apparent contradiction is the high correlation between two very different kinds of noise fluctuations: (1) within-day, trial-to-trial noise fluctuations about the mean response to each stimulus; and (2) between-day changes in the mean stimulus tuning vector. Indeed, we have found empirically that the change in the mean stimulus tuning vector *between* days is closely aligned with the directions of the largest *within*-day, trial-to-trial noise fluctuations (**Fig. 3f**). Before discussing our main results, in **§2** we first provide a self-contained introduction to optimal linear decoders and their relationship to stimulus statistics. In **§3** we will explain theoretically how this important empirical observation enables decoders optimized for a single day to also maintain high performance across days, despite substantial day-to-day changes in

single neuron coding properties. Further, in **§4** we showcase a simple proof-of-principle neural network model that: (1) exhibits substantial day-to-day variability in single neuron coding properties (analogous to **Extended Data Figs 2f,k**); but nevertheless (2) exhibits a tight relation between the within-day fluctuations about neurons' mean responses and their between-day changes in mean responses (analogous to **Fig. 3f**); and (3) exhibits day-to-day robustness of optimal decoders that are trained on data from a single day.

§2. Discrimination of two stimuli using the signals of a noisy neural population

Consider two stimuli $\mathbf{s}^A$ and $\mathbf{s}^B$, each of which elicits a conditional distribution of cortical neural ensemble activity, namely $P_A(\mathbf{r}|\mathbf{s}^A)$ and $P_B(\mathbf{r}|\mathbf{s}^B)$. We wish to decode stimulus identity from the population activity using a decision variable that reads out the activity in a linear manner, $v = \hat{\mathbf{w}} \cdot \mathbf{r}$. The two conditional distributions for the ensemble activity lead to conditional distributions for the decision variable, $P_A(v|\mathbf{s}^A)$ and $P_B(v|\mathbf{s}^B)$. The ease with which we can discriminate the two stimuli depends on how well separated these two distributions are.

When the readout vector $\hat{\mathbf{w}}$ samples from many neurons, and the neural populations are weakly correlated, the distributions over v will be approximately Gaussian and thus well characterized by their mean and variance. More generally, a convenient measure of the separation between the two distributions is given by the signal-to-noise ratio (SNR), also known as $(d')^2$. This measure is the squared difference in the means of the two distributions, normalized to the variance:

$$d'(\hat{\mathbf{w}})^2 = \frac{\left[\langle v|\mathbf{s}^A\rangle - \langle v|\mathbf{s}^B\rangle\right]^2}{\frac{1}{2}\langle(\delta v)^2|\mathbf{s}^A\rangle + \frac{1}{2}\langle(\delta v)^2|\mathbf{s}^B\rangle} = \frac{[\hat{\mathbf{w}}\cdot\Delta\boldsymbol{\mu}]^2}{\hat{\mathbf{w}}^T\boldsymbol{\Sigma}\hat{\mathbf{w}}}, \quad (1)$$

where $\Delta\boldsymbol{\mu} = \boldsymbol{\mu}_A - \boldsymbol{\mu}_B$ and $\boldsymbol{\mu}_A$ and $\boldsymbol{\mu}_B$ are the mean neural population patterns for each stimulus, and $\boldsymbol{\Sigma} = \frac{1}{2}\boldsymbol{\Sigma}^A + \frac{1}{2}\boldsymbol{\Sigma}^B$ is the average noise covariance matrix of the two conditional distributions of neural activity patterns. This measure of discriminability depends on the statistical structure of the two conditional distributions of neural population activity in response to the two stimuli, and on the linear readout direction $\widehat{\mathbf{w}}$. One can maximize (1) over the choice of readout $\widehat{\mathbf{w}}$ to obtain the optimal readout

$$\mathbf{w}_{\text{opt}} = \boldsymbol{\Sigma}^{-1}\Delta\boldsymbol{\mu}, \tag{2}$$

and its associated optimal signal-to-noise ratio

$$\left(d'_{\text{opt}}\right)^2 = \Delta\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\Delta\boldsymbol{\mu}. \tag{3}$$

This optimal SNR depends on the stimulus conditioned neural distributions only through the difference of means $\Delta\boldsymbol{\mu}$ (which we henceforth call the stimulus tuning vector) and the average covariance $\boldsymbol{\Sigma}$. When the two distributions are exactly Gaussian with the same covariance matrix, this optimal SNR is the Kullback-Leibler divergence between the two distributions. This divergence is a statistical measure of how different the two distributions are from each other and governs the error rate of a hypothesis test attempting to distinguish between them[1]. Furthermore, when the two means $\boldsymbol{\mu}_A$ and $\boldsymbol{\mu}_B$ are close to each other, this Kullback-Leibler divergence becomes proportional to the Fisher information[2] conveyed about the stimulus identity by a single neural activity pattern.

§3. Robustness of decoders to changes in day-to-day neural population statistics

We next consider a situation in which we record neural activity on two distinct days (which we term day 1 and day 2). We assume that on day 1 the neural population has a stimulus tuning vector $\Delta\boldsymbol{\mu}_1$ and average within-day, trial-to-trial noise covariance matrix $\boldsymbol{\Sigma}_1$, while on day 2 the

corresponding stimulus statistics have changed to $\Delta\boldsymbol{\mu}_2$ and $\boldsymbol{\Sigma}_2$, respectively. Thus, the two days have two *different* optimal decoders according to Eq. (2): $\mathbf{w}_{\mathrm{opt},1} = \boldsymbol{\Sigma}_1^{-1}\Delta\boldsymbol{\mu}_1$ and $\mathbf{w}_{\mathrm{opt},2} = \boldsymbol{\Sigma}_2^{-1}\Delta\boldsymbol{\mu}_2$. To understand the conditions under which we can obtain robust decoding across days using a single decoder, *despite* changes in stimulus statistics, we consider the $(d')^2$ values obtained by decoding on day 2 using two different decoders: the optimal decoder $\mathbf{w}_{\mathrm{opt},2}$ for day 2, and the alternate decoder $\mathbf{w}_{\mathrm{opt},1}$ that would have been optimal for day 1 but is now suboptimal for day 2. The first optimal $(d')^2$ is given, according to Eq. (3), by

$$\left(d'_{\mathrm{opt},2}\right)^2 = \Delta\boldsymbol{\mu}_2\boldsymbol{\Sigma}_2^{-1}\Delta\boldsymbol{\mu}_2. \tag{4}$$

The second suboptimal $(d')^2$, obtained using the stimulus statistics from day 2 but the optimal decoder from day 1 is given, according to Eq. (2), by

$$\left(d'_{\mathrm{subopt},2}\right)^2 = d'\left(\mathbf{w}_{\mathrm{opt},1}\right)^2 = \frac{\left[\mathbf{w}_{\mathrm{opt},1}\cdot\Delta\boldsymbol{\mu}_2\right]^2}{\mathbf{w}_{\mathrm{opt},1}^T\boldsymbol{\Sigma}_2\,\mathbf{w}_{\mathrm{opt},1}} = \frac{\left[\Delta\boldsymbol{\mu}_1^T\,\boldsymbol{\Sigma}_1^{-1}\Delta\boldsymbol{\mu}_2\right]^2}{\Delta\boldsymbol{\mu}_1^T\,\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1^{-1}\Delta\boldsymbol{\mu}_1}. \tag{5}$$

We define a measure of decoder robustness R as

$$R = \frac{\left(d'_{\mathrm{subopt},2}\right)^2}{\left(d'_{\mathrm{opt},2}\right)^2} = \frac{\left[\Delta\boldsymbol{\mu}_1^T\,\boldsymbol{\Sigma}_1^{-1}\Delta\boldsymbol{\mu}_2\right]^2}{(\Delta\boldsymbol{\mu}_1^T\,\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\Sigma}_2\boldsymbol{\Sigma}_1^{-1}\Delta\boldsymbol{\mu}_1)(\Delta\boldsymbol{\mu}_2\boldsymbol{\Sigma}_2^{-1}\Delta\boldsymbol{\mu}_2)}. \tag{6}$$

The decoder robustness R is the fractional $(d')^2$ attainable on day 2 using the decoder optimal for day 1 (but suboptimal for day 2) relative to that for the optimal decoder for day 2. R has a value between 0 and 1. This measure captures how robustly a decoder optimized solely for day 1 generalizes to day 2, with values near 1 indicating high day-to-day decoder robustness.

We are particularly interested in situations in which the stimulus statistics change substantially from day 1 to day 2, yet the decoder robustness nevertheless remains high. To provide a simple example of such a scenario, we consider the simple case in which the covariance doesn't change (*i.e.,* $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$) but the stimulus tuning vector does (*i.e.,* $\Delta\boldsymbol{\mu}_1 =$

$\Delta\boldsymbol{\mu} - \boldsymbol{\varepsilon}$ and $\Delta\boldsymbol{\mu}_2 = \Delta\boldsymbol{\mu} + \boldsymbol{\varepsilon}$). Here $\Delta\boldsymbol{\mu} = (\Delta\boldsymbol{\mu}_1 + \Delta\boldsymbol{\mu}_2)/2$ is the mean stimulus tuning vector across days, and $\boldsymbol{\varepsilon} = (\Delta\boldsymbol{\mu}_2 - \Delta\boldsymbol{\mu}_1)/2$ reflects the change in stimulus tuning across days. To better understand the decoder robustness R, we can take advantage of the common covariance $\boldsymbol{\Sigma}$ across days to express $\Delta\boldsymbol{\mu}$ and $\boldsymbol{\varepsilon}$ in the eigenbasis of the within-day noise covariance (note there is nothing essential about the assumption of identical covariance across days; it just provides a particularly simple example, and in the next section we treat another example in which the covariance changes across days). Let λ_α and $\mathbf{v}^\alpha$ denote the α'th eigenvalue and eigenvector, respectively, of $\boldsymbol{\Sigma}$. Furthermore, let $\Delta\mu_\alpha$ and ε_α denote the components of $\Delta\boldsymbol{\mu}$ and $\boldsymbol{\varepsilon}$, respectively, along the eigenvector $\mathbf{v}^\alpha$. Then in the eigenbasis of $\boldsymbol{\Sigma}$, Eq. (6) reduces to,

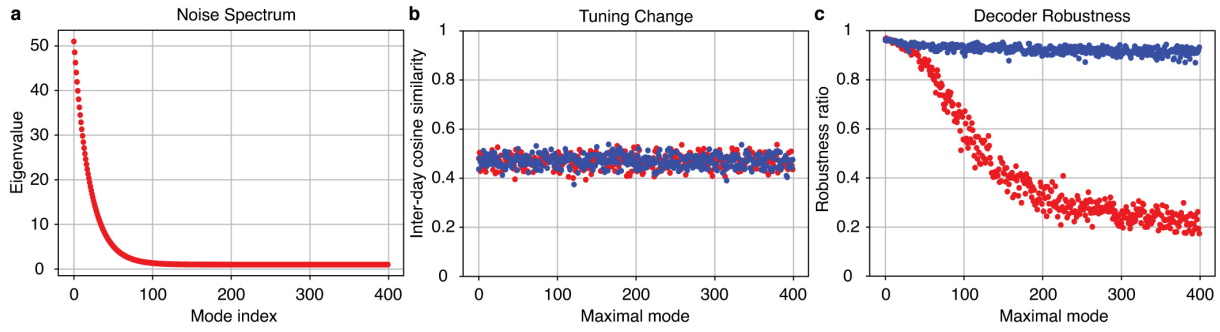
$$R = \frac{\left[\sum_\alpha \left(\frac{\Delta\mu_\alpha^2}{\lambda_\alpha} - \frac{\varepsilon_\alpha^2}{\lambda_\alpha}\right)\right]^2}{\left[\sum_\alpha \frac{(\Delta\mu_\alpha - \varepsilon_\alpha)^2}{\lambda_\alpha}\right]\left[\sum_\alpha \frac{(\Delta\mu_\alpha + \varepsilon_\alpha)^2}{\lambda_\alpha}\right]}. \tag{7}$$

Clearly when $\varepsilon_\alpha = 0$ for all α, indicating no change in neural population statistics, the decoder robustness $R = 1$. Also, inspecting Eq. (7), it is clear that a change of the stimulus tuning vector along a single noise eigenmode α with fixed magnitude ε_α leads to a smaller (larger) reduction in R when the mode α corresponds to a larger (smaller) noise eigenvalue λ_α.

This observation suggests a general principle: if the day-to-day change in the mean stimulus tuning vector, which is proportional to $\boldsymbol{\varepsilon}$, is preferentially aligned to within-day noise eigenmodes $\mathbf{v}^\alpha$ with large noise eigenvalues λ_α, then decoder robustness R may remain high even if the mean stimulus tuning vector changes substantially. We can measure the change in the mean stimulus tuning vector through the cosine similarity metric

$$C = \frac{\Delta\boldsymbol{\mu}_1^T \Delta\boldsymbol{\mu}_2}{\sqrt{\Delta\boldsymbol{\mu}_1^T \Delta\boldsymbol{\mu}_1}\sqrt{\Delta\boldsymbol{\mu}_2^T \Delta\boldsymbol{\mu}_2}}. \tag{8}$$

The key question is then, is it possible to have a mean inter-day stimulus tuning vector change $\boldsymbol{\varepsilon}$ such that there is a substantial inter-day change in stimulus tuning (*i.e.*, small C), but a large decoder robustness (*i.e.,* large R)? We show in **Appendix Fig. 1** that this is indeed possible, when the stimulus tuning vector change across days is aligned to directions of high within-day noise fluctuations, as suggested by Eq. (7). In particular, we consider a scenario with an exponentially decaying within-day noise eigenspectrum $\{\lambda_\alpha\}$ (**Appendix Fig. 1a**). We also only consider large stimulus tuning vector changes, corresponding to $\boldsymbol{\varepsilon}$ vectors with large norm, which lead to a small between-day tuning vector cosine similarity of $C \approx 0.5$ in Eq. (8) (**Appendix Fig. 1b**). We then find that the direction of between-day stimulus tuning change $\boldsymbol{\varepsilon}$ in relation to within-day noise eigenmodes can substantially impact decoder robustness. Indeed, if the magnitude of the projection of the between-day stimulus tuning change vector $\boldsymbol{\varepsilon}$ onto each within-day noise eigenmode $\mathbf{v}^\alpha$ is proportional on average to that mode's noise eigenvalue λ_α (as it is in the data in **Fig. 3f**), then the decoder robustness R in Eq. (7) can remain as high as 0.9 (blue points in **Appendix Fig. 1c**). On the other hand, for the same overall magnitude of stimulus tuning change, with a similar cosine similarity, if $\boldsymbol{\varepsilon}$ instead projects onto every noise eigenmode with similar magnitude on average (unlike the data in **Fig. 3f**), then decoder robustness can decrease substantially to as small as 0.2 (red points in **Appendix Fig. 1c**).



Appendix Fig. 1 Decoder robustness despite substantial changes in stimulus tuning,when inter-day tuning changes co-align with intra-day fluctuations.

(A) We simulate a population of *N*=400 neurons. We assume a generic within-day noise covariance $\mathbf{\Sigma}$ with a random orthonormal basis of eigenvectors $\mathbf{v}^{\alpha}$ and an exponentially decaying eigenvalue spectrum $\lambda_{\alpha} = 1 + \Gamma e^{-\frac{\alpha}{K}}$ for $\alpha = 0$ to $N-1$. This corresponds to approximately K large eigenvalues ranging from $\lambda_0 = 1 + \Gamma$ down to approximately 1. An example spectrum is shown in (A) for $\Gamma = 50$ and $\mathrm{K} = 20$. These parameters were chosen to qualitatively match the eigenvalue spectrum observed in **Fig. 3f**, but none of our conclusions depend in detail on this specific choice of Γ and K.

(B) We further model the stimulus tuning vectors on day 1 and day 2 as $\Delta\boldsymbol{\mu}_1 = \Delta\boldsymbol{\mu} - \boldsymbol{\varepsilon}$ and $\Delta\boldsymbol{\mu}_2 = \Delta\boldsymbol{\mu} + \boldsymbol{\varepsilon}$ respectively, where $\Delta\boldsymbol{\mu}$ is a random Gaussian vector whose elements are chosen i.i.d. from a zero mean Gaussian with variance 1 (which simply sets an overall scale for $(d')^2$). We then model the day-to-day change $\boldsymbol{\varepsilon}$ in stimulus tuning using two methods. In the first method, we simply choose the components ε_{α} (in the eigenbasis $\mathbf{v}^{\alpha}$) to be i.i.d. random variables drawn from a zero mean unit variance Gaussian distribution for all $\alpha \leq \alpha_{max}$ and 0 for all $\alpha > \alpha_{max}$. Thus in this method, the between-day stimulus tuning change $\boldsymbol{\varepsilon}$ has uniform power on the largest within-day noise modes up to a maximal mode index α_{max} and zero power on the smaller noise modes. In the second method, we make the same random choice for $\boldsymbol{\varepsilon}$ up to some maximal mode index α_{max}, but we additionally scale up each component ε_{α} by the noise eigenvalue λ_{α}. This yields on average a larger magnitude for projections of the between-day stimulus

tuning change $\boldsymbol{\varepsilon}$ onto eigenvectors $\mathbf{v}^{\alpha}$ of the within-day noise covariance $\boldsymbol{\Sigma}$ with larger eigenvalues λ_{α}, similar to what we observe in the data in **Fig. 3f.** For both methods we multiply $\boldsymbol{\varepsilon}$ by an overall scale factor so that its norm is 60% of the expected norm of $\Delta\boldsymbol{\mu}$. We chose this large percentage to force a large fractional change in the stimulus tuning vector from day to day that is approximately the same for every maximal mode index α_{max}. To confirm this, for each value of α_{max}, we plot the cosine similarity C in Eq. (8) for both methods, with method 1 given by the red points and method 2 given by the blue points. The low value of C in (B) indicates a substantial change in stimulus tuning from day to day for all maximal mode indices α_{max}.

(C) We next plot the day-to-day decoder robustness R defined in Eq. (7) for both method 1 (weak alignment of stimulus tuning change with noise, red points) and method 2 (stronger alignment of stimulus tuning change with noise, blue points). The red points with α_{max} near $N-1$ correspond to the stimulus tuning vector change $\boldsymbol{\varepsilon}$ being uniformly spread out across almost all noise modes, resulting in a very low day-to-day decoder robustness of $R \approx 0.2$. The blue points with α_{max} near $N-1$ correspond to the stimulus tuning vector change $\boldsymbol{\varepsilon}$ being spread out across almost all noise modes, but with a strength proportional to each mode's within-day noise eigenvalue λ_{α}, yielding a substantially high decoder robustness of $R \approx 0.9$.

.

In summary, we have shown that if between-day changes in the mean stimulus tuning vector are preferentially aligned to directions of large within-day noise fluctuations, then a

decoder trained to be optimal on day 1 will also tend to do well on day 2, relative to a decoder trained specifically to be optimal on day 2. A key intuition for why this is true can be gleaned from the structure of the optimal decoder on day 1, given by $\mathbf{w}_{\mathrm{opt},1} = \boldsymbol{\Sigma}^{-1}\Delta\boldsymbol{\mu}_1$. This decoder will tend to avoid directions in which the within-day noise is large, in order to be robust to this noise. Such a decoder will then also be robust to day-to-day changes in the tuning vector, provided such day-to-day changes in the tuning vector preferentially lie along directions in which the within-day noise is large. More generally, when this connection between inter-day mean changes and within-day noise fluctuations persists across multiple days, it should be straightforward to find a single decoder that achieves high performance across multiple days despite subtantial changes in single neuron coding properties.

§4. A simple neural network model to account for key observations

The previous section elucidated the theoretical significance of **Fig. 3f**, namely the observation that between day changes in the mean stimulus tuning vector are preferentially aligned to directions of large within-day noise fluctuations. In particular we showed that this observation makes it easier to construct a single decoder that performs well across many days. However, the previous section did not address any particular neural mechanism that could naturally achieve a tight relation connecting substantial between-day changes in mean stimulus responses to the eigenstructure strength of within-day noise fluctuations. Here, we provide an exceedingly simple mechanistic neural model that can generate such a relation in a biologically plausible manner without any fine tuning. We make no claim that this is the only way such a relation can arise mechanistically; our model is intended to just serve as a proof of principle that this relation is realizable in one simple setting.

We consider a two-layer circuit with N_s sensory neurons in its first layer and N_c visual cortical neurons in the second layer. We respectively denote the activity patterns of these cells as $\mathbf{s}_j$, for $j = 1, \ldots, N_s$ and $\mathbf{r}_i$, for $i = 1, \ldots, N_c$. We take a simple linear input-output relationship given by

$$\boldsymbol{r} = \mathbf{G}^d\mathbf{W}\left(\mathbf{s} + \boldsymbol{\xi}^{\text{in}}\right) + \boldsymbol{\xi}^{\text{out}}. \tag{9}$$

Here $\boldsymbol{\xi}^{\text{in}}$ and $\boldsymbol{\xi}^{\text{out}}$ are, respectively, vectors, or patterns, of sensory input noise and cortical output noise that fluctuate from trial to trial. We model them as Gaussian random vectors with zero means and covariance matrices $\boldsymbol{\Sigma}^{\text{in,d}}$ and $\boldsymbol{\Sigma}^{\text{out,d}}$. Here $d = 1{,}2$ is a day index indicating that the covariance of both the input noise and the output noise can vary from day to day. $\mathbf{G}^d$ in Eq (9) is a diagonal gain matrix that can also vary from day to day. Each diagonal element $\boldsymbol{G}_{ii}^d$ reflects a gain, or single neuron excitabilty level, of a cortical neuron i on day d. The N_c by N_c matrix $\mathbf{W}$ reflects a constant synaptic connectivity matrix from the sensory layer to the cortical layer, and we assume it does not vary from day to day.

Now the conditional distribution of cortical activity, $\mathbf{r}^A$, that results from a specific pattern of input activity, $\mathbf{s}^A$ is Gaussian with a mean on day d given by

$$\boldsymbol{\mu}_d^A = \mathbf{G}^d\mathbf{W}\mathbf{s}^A. \tag{10}$$

For two stimuli $\mathbf{s}^A$ and $\mathbf{s}^B$, this implies a stimulus tuning vector on day d given by

$$\Delta\boldsymbol{\mu}_d = \boldsymbol{\mu}_d^A - \boldsymbol{\mu}_d^B = \mathbf{G}^d\mathbf{W}\Delta\mathbf{s}, \tag{11}$$

where $\Delta\mathbf{s} = \mathbf{s}^A - \mathbf{s}^B$ is the stimulus tuning vector in the first layer. We are assuming that the mean stimulus representations $\mathbf{s}^A$ and $\mathbf{s}^B$ in the first layer do not change from day to day, although we did also explore that possibility and obtained qualitatively similar results. Here, we

focused on day-to-day changes in the statistics of the input noise (through $\mathbf{\Sigma}^{\mathrm{in,d}}$), the output noise (through $\mathbf{\Sigma}^{\mathrm{out,d}}$), and the excitability of each cortical neuron (through $\mathbf{G}^{d}$). This last day-to-day variability can cause a day-to-day change in the cortical stimulus tuning vector $\Delta\boldsymbol{\mu}_d$ even though the mean sensory tuning vector $\Delta\mathbf{s}$ does not change from day to day. Next, the trial-to-trial noise covariance of cortical activity on day d is given by

$$\mathbf{\Sigma}_d = \mathbf{G}^{d}\mathbf{W}\,\mathbf{\Sigma}^{\mathrm{in,d}}\mathbf{W}^{T}\mathbf{G}^{\mathrm{d}} + \mathbf{\Sigma}^{\mathrm{out,d}}. \tag{12}$$

This within-day noise covariance can vary from day to day via combined day-to-day changes in cortical excitability $\mathbf{G}^{d}$, input noise statistics $\mathbf{\Sigma}^{\mathrm{in,d}}$, and output noise statistics $\mathbf{\Sigma}^{\mathrm{out,d}}$. Thus overall this model gives us a stimulus tuning vector $\Delta\boldsymbol{\mu}_d$ in Eq. (10) and noise covariance $\mathbf{\Sigma}_d$ in Eq. (12) which can be inserted into Eq. (6) to determine the day-to-day decoder robustness R. Moreover, this model yields day-to-day changes in single neuron $(d')^2$ values. Indeed the $(d')^2$ value for cortical neuron i on day d is given by

$$\left(d'_{i,d}\right)^2 = \frac{\left(\Delta\boldsymbol{\mu}_{d,i}\right)^2}{\mathbf{\Sigma}_{d,ii}}. \tag{13}$$

We next describe the constant properties $\mathbf{W}$ and $\Delta\mathbf{s}$, as well as the nature of the day-to-day changes in $\mathbf{G}^{d}$, $\mathbf{\Sigma}^{\mathrm{in,d}}$ and $\mathbf{\Sigma}^{\mathrm{out,d}}$. We model $\mathbf{W}$ as a random synaptic connectivity matrix with a prescribed singular value distribution through the singular value decomposition $\mathbf{W} = \mathbf{U}\mathbf{D}\mathbf{V}^{T}$. We assume $\mathbf{U}$ is an $N_c \times N_s$ matrix with random orthonormal columns that are the output singular vectors of $\mathbf{W}$, $\mathbf{V}$ is an $N_s \times N_s$ random orthonormal matrix, and $\mathbf{D}$ is an $N_s \times N_s$ diagonal matrix of singular values, where $\mathbf{D}_{jj} = d_j$. We choose the singular value spectrum d_j to be exponentially decaying of the form $d_j = \Gamma e^{-\frac{j}{K}}$, corresponding to roughly K nontrivially large singular values. Note that the rows of $\mathbf{W}$ can be thought of in this model as the set of

receptive fields of cortical neurons, and if this set of receptive fields approximately spans a low-dimensional space, then we naturally expect the singular value spectrum to decay with a small number of large singular values. The detailed structure of the singular vectors in $\mathbf{U}$ and $\mathbf{V}$ do not impact our final conclusions, hence we simply chose them to be random. For the stimulus tuning vector $\Delta\mathbf{s}$ in the sensory layer, we simply choose its components i.i.d. from a zero mean unit variance Gaussian. The unit variance of this Gaussian simply sets an overall scale by which all other variances are measured. We now turn to the properties in $\mathbf{G}^d$, $\mathbf{\Sigma}^{\mathrm{in,d}}$ and $\mathbf{\Sigma}^{\mathrm{out,d}}$ which each change day by day.

We assume the day-to-day changes in the gain matix are given by $\mathbf{G}^d = \mathbf{I} + \mathrm{f}\,\mathrm{diag}\left(\boldsymbol{\varepsilon}^{\mathrm{G,d}}\right)$, where $\mathbf{I}$ is the identity matrix, $\boldsymbol{\varepsilon}^{\mathrm{G,d}}$ is a zero mean unit variance random Gaussian vector denoting a day specific gain perturbation, $\mathrm{diag}\left(\boldsymbol{\varepsilon}^{\mathrm{G,d}}\right)$ is a diagonal matrix with diagonal elements specified by the components of $\boldsymbol{\varepsilon}^{\mathrm{G,d}}$, and f is a positive fraction between 0 and 1. Under this scheme, every cortical neuron i has a base gain of 1, which can change on day d to the value $1 + \mathrm{f}\,\varepsilon_i^{\mathrm{G,d}}$. Similarly, we assume day-to-day changes in the input noise covariance matrix are given by $\mathbf{\Sigma}^{\mathrm{in,d}} = \sigma_{\mathrm{in}}^2\,\mathbf{I} + \mathrm{f}\,\sigma_{\mathrm{in}}^2\ \mathrm{diag}\left(\boldsymbol{\varepsilon}^{\mathrm{in,d}}\right)$, where $\boldsymbol{\varepsilon}^{\mathrm{in,d}}$ is a zero mean unit variance random Gaussian vector. Under this scheme, every sensory neuron j has a base input noise variance σ_{in}^2 which can change on day d to the value $\sigma_{\mathrm{in}}^2\left(1 + \mathrm{f}\,\varepsilon_j^{\mathrm{in,d}}\right)$. Also, we assume day-to-day changes in the output noise covariance matrix are given by $\mathbf{\Sigma}^{\mathrm{out,d}} = \sigma_{\mathrm{out}}^2\,\mathbf{I} + \mathrm{f}\,\sigma_{\mathrm{out}}^2\ \mathrm{diag}\left(\boldsymbol{\varepsilon}^{\mathrm{out,d}}\right)$, where $\boldsymbol{\varepsilon}^{\mathrm{out,d}}$ is a zero mean unit variance random Gaussian vector. Under this scheme, every cortical neuron i has a base output noise variance σ_{out}^2 which can change on day d to the value $\sigma_{\mathrm{out}}^2\left(1 + \mathrm{f}\,\varepsilon_i^{\mathrm{out,d}}\right)$. Finally, for any sensory or cortical neuron in which

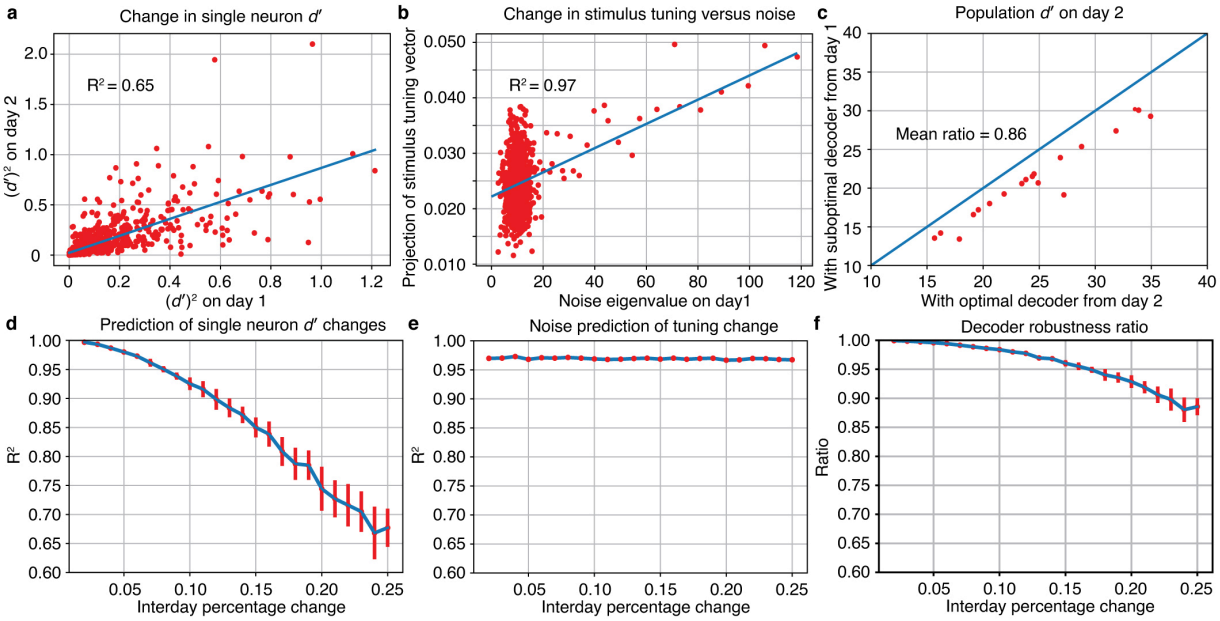
the sampled noise variance happens to be negative (an event that occurs with low probability for small f), we rectify it to a small value of 0.01.

Overall, a key parameter in this framework is f, which denotes the common fractional change on average in single neuron gain, input variance, and output variance across days. For small (large) f there are negligible (substantial) day-to-day changes in single neuron gain, and input and output noise variance. The key question is then: is there an intermediate range of f such that there are substantial changes in single neuron coding fidelity, as measured for example by substantial day-to-day changes in single neuron $(d')^2$ values in Eq. (13), yet at a population level, the overall day-to-day decoder robustness R in Eq. (6) remains high?

We show in **Appendix Fig. 2** that this dichotomy between substantial day-to-day changes in single neuron coding fidelity, yet population level decoder robustness, is indeed possible in this model. We modelled $N_s = 200$ sensory neurons and $N_c = 1000$ cortical neurons. We chose $\Gamma = 10$ and $K = 20$ so that the singular spectrum of $\mathbf{W}$, given by $d_j = \Gamma e^{-\frac{j}{K}}$, corresponds to the amplification of a small number of modes, with an order of magnitude range of amplification from largest to smallest. This ampification of a small number of modes from sensory inputs to cortical outputs is a key property required to achieve our results, and is a property shared by prior models of V1[3]. We chose our baseline output noise variance $\sigma^2_{\text{out}} = 10$ to compensate for this amplfication in $\mathbf{W}$ so that single neuron $(d')^2$ values in the model would remain primarily less than 1 as they do in the data (**Extended Data Fig. 2jkl**). We chose our baseline input noise variance $\sigma^2_{\text{out}} = 1$ so that the overall optimal population decoder $(d')^2$ values would remain in the range of 10 to 40 as they do in the data (**Fig. 3c**). With these baseline parameters chosen to qualitatively match the characteristics of the data, we examined a relatively large fractional

day-to-day change of $f = 0.25$ of single neuron cortical gains, cortical output noise and sensory input noise. We found for this value of f a substantial variation in single neuron $(d')^2$ values from day to day, where the R^2, or fraction of variance explained in the $(d')^2$ value of neurons on day 2, given their $(d')^2$ values on day 1, was as low as 0.65 (**Appendix Fig. 2a**). This day-to-day variation in single neuron $(d')^2$ values is qualitatively similar to that seen in the data (**Extended Data Fig. 2k**). We then confirmed that despite this variation at a single neuron level, between day changes in the stimulus tuning vector (i.e. $\Delta\boldsymbol{\mu}_2 - \Delta\boldsymbol{\mu}_1$ obtained from Eq. (11)) had a magnitude of projection onto a noise eigenmode $\mathbf{v}^\alpha$ of $\boldsymbol{\Sigma}_1$ in Eq. (12), which was tightly related to the corresponding noise eigenvalue λ_α, especially for large noise eigenvalues (**Appendix Fig. 2b**).

This relation connecting between day stimulus tuning vector changes to within-day noise eigenmodes is qualitatively similar to what is observed in the data (**Fig. 3f**). Finally, for multiple pairs of days, we plot the $(d')^2$ value on day 2 obtained by a decoder that is optimal for day 1, against the necessarily higher $(d')^2$ value for a decoder that is optimal for day 2 (**Appendix Fig. 2c**). We find that these points lie near the unity line with a mean decoder robustness of $R = 0.86$. We also compute how these properties vary as a function of the day-to-day fractional change f, finding that day-to-day changes in single neuron $(d')^2$ values increase rapidly with f (**Appendix Fig. 2d**), while the correlation between day-to-day stimulus tuning changes and within-day noise eigenmodes remains high (**Appendix Fig. 2e**), as does the decoder robustness R (**Appendix Fig. 2f**) (See next page).



Appendix Fig. 2 | Coexistence of single neuron level variability and population level decoder robustness in a simple model.

(A) For the model parameters described in the Appendix text, and for an interday fractional change f=0.25, for a single pair of days we plot single neuron $(d')^2$ values on day 2 against the same neuron's $(d')^2$ value on day 1. The small R^2 value, or fraction of variance explained from day 1 to day 2, indicates that the model parameters chosen allow substantial variaiblity in coding fidelity from day to day at the single neuron level.

(B) For twenty pairs of days (corresponding to the analog of averaging over 5 mice with 5 sessions each, qualitatively similar to what is done in **Fig. 3f**), we compute the noise eigenvalues λ_α of $\mathbf{\Sigma}_1$ in Eq. (12), and the magnitude of the projection of the stimulus change vector $\Delta\boldsymbol{\mu}_2 - \Delta\boldsymbol{\mu}_1$ obtained from Eq. (11) onto the corresponding noise mode $\mathbf{v}^\alpha$ of $\mathbf{\Sigma}_1$. We average these two quantities over the 20 pairs of days (sorting eigenvalues from largest to smallest on each day to identify

modes across days) and plot them against each other, revealing that the average eigenvalue magnitude can predict the average projection of the stimulus change vector well with a high R^2 value.

(C) For the same twenty pairs of days in (B) we plot the $(d')^2$ value on day 2 obtained by a decoder that is optimal for day 1, against the $(d')^2$ value for a decoder that specifically is optimal for day 2. The points lie near the unity line with a mean decoder robustness across 20 pairs of days of decoder robustness $R = 0.86$.

(D) The mean (blue line) and standard deviation (red bars), across 20 pairs of days, of the fraction of variance explained in single neuron $(d')^2$ values on day 2 given the same neuron's $(d')^2$ value on day 1, for the same model parameters in the main text, but for a range of day-to-day fractional change values of f. We see that single neuron $(d')^2$ values change rapidly from day to day as f increases.

(E) The fraction of variance explained in projections of the stimulus change vector onto a noise eigenmode, given the noise eigenvalue, with both averaged over 20 pairs of days, identical to the R^2 value computed in (B), but now plotted for a range of f. The alignment of between-day stimulus tuning changes to within-day noise remains a robust property of this model over this range of f.

(F) The mean (blue line) and standard deviation (red bars), across 20 pairs of days of the decoder robustness R in Eq. (6), plotted for a range of f. This population level decoder robustness degrades gracefully with the fractional day-to-day change f even though the day-to-day stability in single neuron $(d')^2$ values degrades rapidly with f in (D).

Overall this model serves as a simple proof of principle that a network can tolerate large degrees of changes in single neuron $(d')^2$ values from day to day through completely independent and relatively large fractional changes in every single neuron gain, and every single neuron input noise variance, both in an earlier sensory layer and in the recorded cortical layer, without necessarily destroying a structured relationship connecting between-day stimulus tuning changes to within-day noise fluctuations, and without precluding the existence of a robust decoder that performs well across multiple days. Of course we can make no claim that this is the only model that accomplishes this, nor that the changes in biophysical properties posited in the model from day to day are analogous to the actual day-to-day changes in biophysical properties in the brain. However, we believe this model provides the beginnings of a simple conceptual framework to explain how in principle the seemingly contradictory properties of day-to-day single neuron variability and population level decoder robustness can simultaneously coexist.

§5. References

1 Cover, T. M. & Thomas, J. A. *Elements of information theory*. (John Wiley & Sons, 2012).

2 Fisher, R. A. Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society* **22**, 700-725 (1925).

3 Rumyantsev, Oleg I., Jérôme A. Lecoq, Oscar Hernandez, Yanping Zhang, Joan Savall, Radosław Chrapkiewicz, Jane Li, Hongkui Zeng, Surya Ganguli, and Mark J. Schnitzer. 2020. "Fundamental Bounds on the Fidelity of Sensory Cortical Coding." *Nature* 580 (7801): 100–105.