nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-20722-y OPEN

Scaling of sensory information in large neural populations shows signatures of information-limiting correlations

MohammadMehdi Kafashan[1], Anna W. Jaffe[1], Selmaan N. Chettih[1], Ramon Nogueira[2], Iñigo Arandia-Romero[3,4], Christopher D. Harvey[1], Rubén Moreno-Bote[5,6] & Jan Drugowitsch[1]✉

How is information distributed across large neuronal populations within a given brain area? Information may be distributed roughly evenly across neuronal populations, so that total information scales linearly with the number of recorded neurons. Alternatively, the neural code might be highly redundant, meaning that total information saturates. Here we investigate how sensory information about the direction of a moving visual stimulus is distributed across hundreds of simultaneously recorded neurons in mouse primary visual cortex. We show that information scales sublinearly due to correlated noise in these populations. We compartmentalized noise correlations into information-limiting and nonlimiting components, then extrapolate to predict how information grows with even larger neural populations. We predict that tens of thousands of neurons encode 95% of the information about visual stimulus direction, much less than the number of neurons in primary visual cortex. These findings suggest that the brain uses a widely distributed, but nonetheless redundant code that supports recovering most sensory information from smaller subpopulations.

[1] Department of Neurobiology, Harvard Medical School, Boston, MA 02115, USA. [2] Center for Theoretical Neuroscience, Mortimer B. Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY, USA. [3] ISAAC Lab, Aragón Institute of Engineering Research, University of Zaragoza, Zaragoza, Spain. [4] IAS-Research Center for Life, Mind, and Society, Department of Logic and Philosophy of Science, University of the Basque Country, UPV-EHU, Donostia-San Sebastián, Spain. [5] Center for Brain and Cognition and Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain. [6] Serra Húnter Fellow Programme and ICREA Academia, Universitat Pompeu Fabra, Barcelona, Spain. ✉email: jan_drugowitsch@hms.harvard.edu

Our brains encode information about sensory features in the activity of large neural populations. The amount of encoded information provides an upper bound on behavioral performance, and so exposes the efficiency and structure of the computations implemented by the brain. The format of this encoding reveals how downstream brain areas ought to access the encoded information for further processing. For example, the amount of information in visual cortex about the drift direction of a moving visual stimulus determines how well one could in principle discriminate different drift directions if the brain operates at maximum efficiency, and its format tells us how downstream motion-processing areas ought to "read out" this information. Therefore, knowing how the brain encodes sensory information about the world is necessary if we are to understand the computations it performs. Unfortunately, we still know little about how sensory information is distributed across neuronal populations even within a single brain area. Is information spread evenly and largely independently across neurons, or in a way that introduces significant redundancy? In the first scenario, one would need to record from the whole neuronal population to get access to all available information, whereas in the second scenario only a fraction of neurons would be needed.

The amount of information about a stimulus feature that can be extracted from neural population activity depends on how this activity changes with a change in the stimulus feature. For information that can be extracted by a linear decoder, which is the information we focus on in this work, it depends on the neurons' tuning curves, as well as how their activity varies across repetitions of the same stimulus (i.e., "noise")[1–4]. Due to the variability in neural responses to repetitions of the same stimulus, each neuron's response provides limited information about the stimulus feature[5–9]. If the noise is independent across neurons, it can be averaged out by pooling across neurons[10], and total information would on average increase by the same amount with every neuron added to this pool (Fig. 1a, red). This corresponds to the first scenario in which information is spread evenly across neurons. If, however, the trial-to-trial variations in spiking are shared across neurons—what are referred to as "noise correlations"—the situation is different. In general, depending on their structure, noise correlations can either improve or limit the amount of information (Fig. 1b), such that the presence of correlated noise alone does not predict its impact. In a theoretical population with translation-invariant tuning curves (i.e., the individual neurons' tuning curves are shifted copies of each other) and noise correlations that are larger for neurons with similar tuning, information might quickly saturate with population size[10,11], corresponding to the second scenario (Fig. 1a, black). Even though such correlation structures, which are traditionally studied in sensory areas, have been observed across multiple brain areas[10,12–15], neural tuning is commonly more heterogeneous than assumed by Zohary et al.[10]. A consequence of this heterogeneity is that sensory information might grow without bound even with noise correlations of the aforementioned structure[16]. Overall, it remains an open question if sensory information saturates in large neural populations of human and animal brains[1].

If information saturates in such populations, then, by the theory of information-limiting correlations (TILC)[17], information in large populations is limited exclusively by one specific component of the noise correlations. This component introduces noise in the direction of the change of the mean population activity with stimulus value (e.g., drift direction; black arrow in Fig. 1b, bottom), thus limiting information about this value. Measuring this noise correlation component directly in neural population recordings is difficult, as noise correlations are, in general, difficult to estimate well[18], and the information-limiting component is usually swamped by other types of correlations that do not limit information[17,19]. Fortunately, however, TILC also predicts *how* information scales with population size if information-limiting correlations are present. We thus exploited this theory to detect the presence of information-limited correlations indirectly by examining how information scales with population size.

In this work, we search for the presence of information-limiting correlations, by simultaneously recording the activity of hundreds of neurons in V1 of awake mice in response to drifting gratings, with hundreds of repeats of each stimulus. We asked how these neurons encoded information about the direction of the moving visual stimulus. We found that noise correlations reduce information even within the limited neural populations we

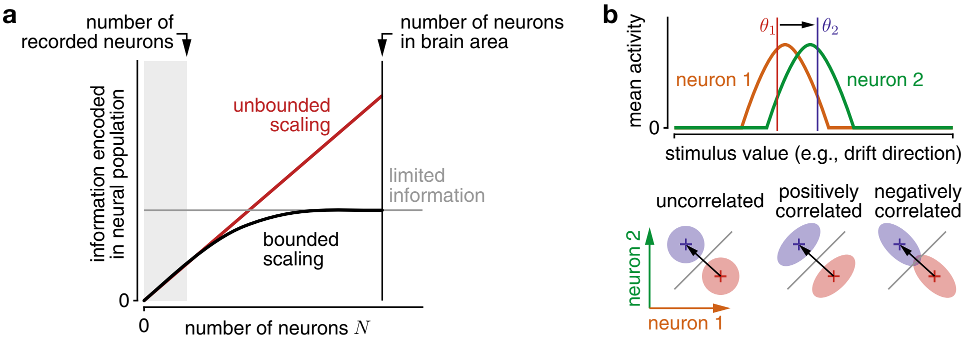


Fig. 1 Information scaling in large neural populations, and the impact of noise correlations on information. a The information that a population of neurons can encode about some stimulus value is always a non-decreasing function of the population size. Information might on average increase with every added neuron (unbounded scaling; red) if the information is evenly distributed across all neurons. In contrast, information can rapidly saturate if information is redundant, and thus it is not strictly limited by population size, but by other factors. In general, it has only been possible to record from a very small subset of neurons of a particular area (gray shaded), from which it is hard to tell the difference between the two scenarios if the sampled population size is too small. **b** The encoded information is modulated by noise correlations. This is illustrated using two neurons with different tunings to the stimulus value (top). The amount of information to discriminate between two stimulus values (θ_1/red and θ_2/blue) depends on the difference in mean population activity (crosses) between stimuli, and the noise correlations (shaded ellipsoids) for either stimulus (bottom, showing joint neural activity of both neurons). The information is largest when the noise is smallest in the direction of the mean population activity difference (black arrow), which leads to the largest separation across the optimal discrimination boundary (gray line). In this example, positive correlations boost information (middle), whereas negative correlations lower it (right), when compared to uncorrelated neurons (left). In general, the impact of noise correlations depends on how they interact with the population's tuning curves.

could record. Applying TILC to compartmentalize information-limiting correlations from nonlimiting correlations, and to extrapolate the growth of information to larger neural populations, we found that on the order of tens of thousands of neurons would be required to encode 95% of the information about the direction of the moving stimulus. Given that there are hundreds of thousands of neurons in this brain region, this means that only a small fraction of the total population is needed to encode this information. This is not because only a small fraction of neurons contains information about the stimulus; rather, we found that most neurons contain information about the stimulus, but because information is represented redundantly, only a small fraction of these neurons is actually needed. Notably, the size of the required neural population depends only weakly on stimulus contrast; thus, increasing the amount of information in this brain area does not substantially increase the number of neurons required to encode 95% of the information about the stimulus. Finally, we found that the low-dimensional neural subspace that captures a large fraction of the noise correlations does not encode a comparably large fraction of information. Overall, our results suggest that information in mouse V1 is both highly distributed and highly redundant, which is true regardless of the total amount of information encoded.

Results

Neural response to drift direction of moving visual stimuli. To measure how sensory information scales with population size, we used two-photon calcium imaging to record neural population activity from layer 2/3 of V1 in awake mice observing a low-contrast drifting grating (10% contrast). The drift direction varied across trials, with each trial drawn pseudorandomly from eight possible directions, spaced evenly around the circle (Fig. 2a). We simultaneously recorded 273–386 neurons (329 on average) across four mice and a total of 16 sessions (Fig. 2b), and analyzed temporally deconvolved calcium activity, summed up over the stimulus presentation period as a proxy for their spike counts within that period. The tuning curves of individual neurons (Fig. 2c) revealed that, on average, only a small fraction of neurons (5–45% across mice/sessions, 18% average) were tuned to the grating's drift direction, while a larger fraction of neurons (38–60% across mice/sessions, 48% average) were sensitive to the grating's orientation, but not its direction of drift. The remaining neurons had no appreciable tuning (14–52% across mice/sessions, 34% average), but were nonetheless included in the analysis, as they can contribute to the information that the population encodes through noise correlations[20,21]. See Supplementary Figs.1–3 for more examples of neural responses, tuning curves, pairwise noise correlations, and raw calcium traces. We found no significant impact of the drift direction in the previous trial on neural responses in the current trial (Supplementary Fig. 1b and Supplementary Table 1). Tuning curves were plotted for the sole purpose of characterizing individual neural responses, but our fits had no bearing on any of our further analysis.

Noise correlations limit information. To quantify stimulus information encoded in the response of neural populations, we asked how well a linear decoder of the recorded population activity (i.e., information decodable by a single neural network layer) would allow us to discriminate between a pair of drift directions (Fig. 3a). Importantly, our aim was to measure information that population activity conveyed about drift direction in general, without prioritizing specific drift directions over others. Even though subselecting a limited set of drift directions is common in animal training, we here focused on discriminating drift directions in pairs only as a tool to get at information about drift direction in general, which should be more reflective of real-world demands. We measured the decoder's performance by generalizing linear Fisher information, usually restricted to fine discriminations, to coarse discrimination (Fig. 3b). This generalization is closely related to the sensitivity index d' from signal detection theory[3,22], and has a set of appealing properties (see "Methods"). In particular, combining the activity of two uncorrelated neural populations causes their associated Fisher information to add, so that it does not trivially saturate like other measures of discrimination performance (Fig. 3c, inset).

We used generalized Fisher information to measure how information about drift direction scales with the number of neurons in the recorded population. Because this scaling depends on the order in which we add particular neurons to the population (individual neurons might contribute different amounts of additional information to a population), we measured average scaling by averaging across a large number of different random orderings (see "Methods"). Figure 3c shows this average scaling for one example session for discriminating between drift directions of 135° and 180° (arbitrary choice; as shown below, other drift direction combinations resulted in comparable information scaling). Information increases with population size, but, on average, additional neurons contribute less additional information to larger populations than to smaller ones. The resulting sublinear scaling is expected if noise correlations limit information. Indeed, trial-shuffling the data to remove pairwise correlations resulted in information that scaled linearly, with average information exceeding that of the non-shuffled data for all population sizes except, trivially, for single neurons, and a significantly higher total information within the recorded population (bootstrap, $p \approx 0.0062$). Such linear scaling was not

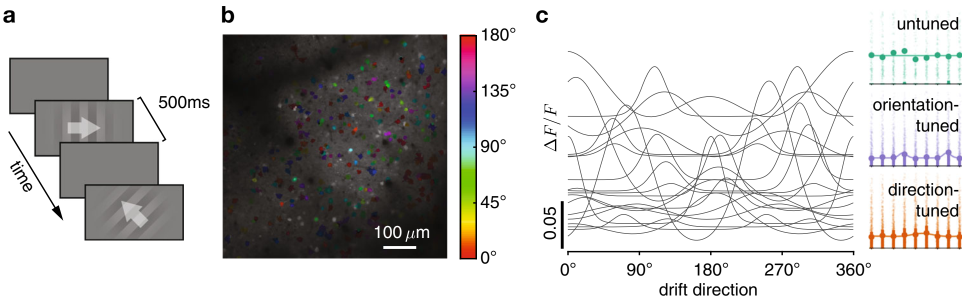


Fig. 2 Experimental design, population recordings, and neural tuning. a Mice passively observed sequences of drifting gratings (white arrows overlaid for illustration only), interleaved with blank screens. **b** Example field-of-view with significantly tuned neurons color coded by their preferred orientation tuning. **c** Left: example fitted tuning curves of 20 significantly tuned neurons. Right: example tuning curves (dots + bars: raw tuning, mean ± 25–75% percentiles; line: fitted) fitted to per-trial neural responses (dots, horizontally jittered) for an untuned (top), orientation-tuned (middle) and direction-tuned (bottom) neuron.

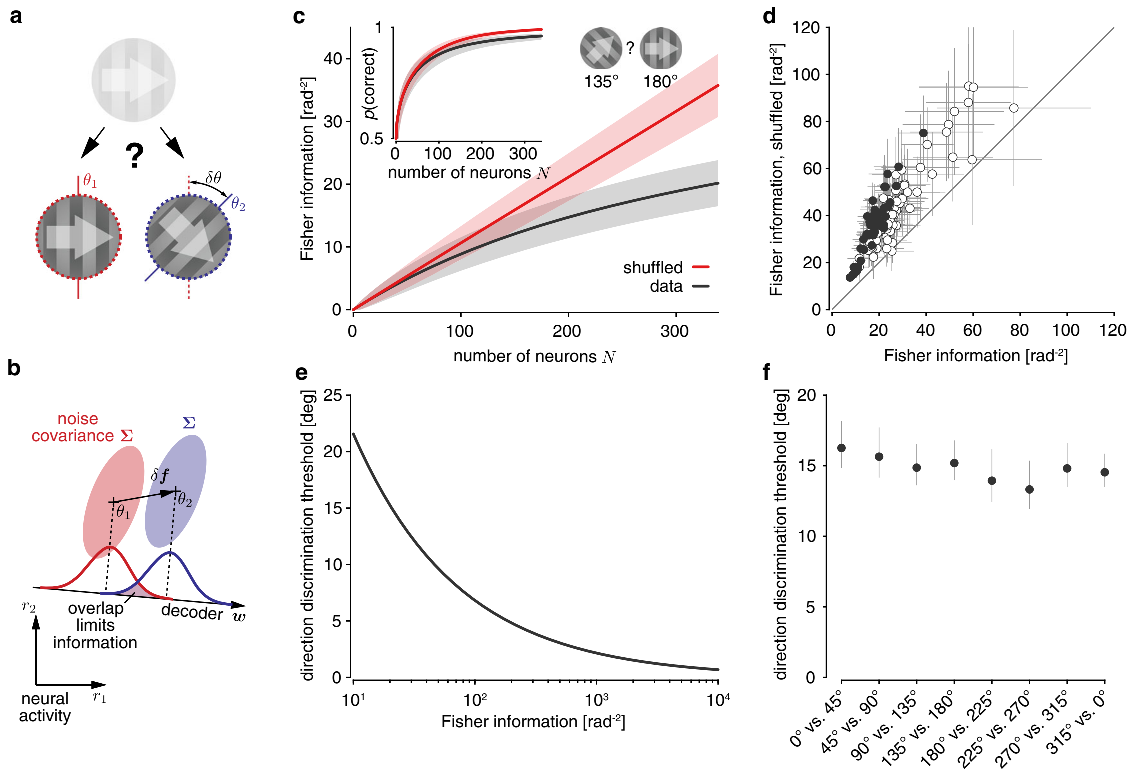


Fig. 3 Noise correlations limit information across all drift directions. a A drift direction discrimination task, in which a hypothetical observer needs to judge which of two template drift directions (θ_1 or θ_2; indicated by white arrows) an observed low-contrast drifting grating corresponds to. **b** Mean activity f (crosses) and noise covariance $\boldsymbol{\Sigma}$ (shaded area ≈ 2SDs) of a pair of neurons across repeated presentation of the same two drift directions, θ_1 (red) and θ_2 (blue). Linear information about drift direction is limited by the projection of the noise onto the optimal linear decoder w. This decoder depends on how mean activity changes with drift direction ($\delta f = f(\theta_2) - f(\theta_1)$) and the noise covariances $\boldsymbol{\Sigma}$. **c** The information associated with discriminating between drift directions 135° and 180° scales sublinearly with population size (black; mean ± 1 SD across random orderings of neurons within the population). If we remove noise correlations by shuffling trials across neurons, the information scales linearly (red). This linear growth would not be apparent from the probability of correctly identifying the stimulus' drift direction (inset), which is monotonically, but non-linearly related to Fisher information, and saturates in both cases. **d** Information in the recorded population was consistently larger for trial-shuffled data across different discriminations, sessions, and mice. Each dot (mean ± 1 SD of information estimate; filled = significant increase, bootstrap, $p < 0.05$) shows the information estimated for one discrimination with $\delta\theta = 45^{\circ}$. **e** The drift direction discrimination threshold (corresponding to 80% correct discriminations) we would expect to see in a virtual discrimination experiment drops with the amount of information that V1 encodes about drift directions. **f** The inferred drift direction discrimination threshold for the same session as in panel **c** is comparable across the different drift direction pairs with $\delta\theta = 45^{\circ}$ used to estimate Fisher information with the recorded population.

apparent if we measured discrimination performance by the fraction of correct discriminations (Fig. 3c, inset), illustrating the point that Fisher information is indeed a better measure to analyze information scaling. Removing noise correlations resulted in a significant information increase in all our datasets (Fig. 3d; paired $t_{63} = -17.93$, two-sided $p \approx 1.96 \times 10^{-26}$; statistics computed across all sessions and mice, but only across non-overlapping $\delta\theta = 45°$ discriminations to avoid duplicate use of individual drift direction trials; see Supplementary Table 2 for avg. per-neuron information for all sessions/mice), confirming that noise correlations indeed limit information in our recorded populations.

To aid interpretation of the estimated amounts of Fisher information, we translated them into quantities that are more frequently measured in experiments. Specifically, we assumed that the recorded neural population was used to discriminate between two close-by drift directions in a virtual fine discrimination task (similar to Fig. 3a). For a given estimate of Fisher information, we could then determine the expected discrimination threshold at which the ideal observer could correctly discriminate between two drift directions in 80% of the trials based solely on neuronal responses (Fig. 3e). This resulted in a discrimination threshold of ~15.2° for the Fisher information estimated from a 135° vs. 180° discrimination (Fig. 3f). Previously reported discrimination threshold of mice, as measured from behavioral performance, ranged from 6.6°[23] over 10–20°[24], to 30–40°[25]. These numbers provide an orders-of-magnitude comparison, but cannot be directly compared to our estimate, as neither study exactly matched the stimuli we used. Moreover, previous work has shown that attending to a stimulus boosts the information encoded about this stimulus[26,27]. As our animals were passive observers that were not actively engaged in any task, the estimated threshold likely underestimate discrimination capabilities. Indeed, higher running speeds, which were previously used as a proxy for increased attention[28], resulted in increased information (as shown previously by Dadarlat and Stryker[29]) and lower thresholds (Supplementary Fig. 4). In line with previous findings[29], this information boost was caused by a combination of a change in population tuning, per-neuron noise variability, and pairwise noise correlations, rather than either of these factors in isolation (Supplementary Fig. 5). Overall, the estimated thresholds provide a reasonable interpretation of the information encoded in the recorded population. Computing the discrimination threshold for all drift direction pairs with $\delta\theta = 45^{\circ}$ resulted in comparable thresholds that did not differ significantly (bootstrap, two-sided $p \approx 0.50$ for session shown in Fig. 3f, two-sided $p > 0.49$ for all sessions/mice). We found comparable information across all drift directions, confirming that we recorded from populations that were homogeneously tuned across all drift directions.

Neural signatures of limited asymptotic information. To identify neural signatures of limited encoded information, we relied on the TILC that showed that noise correlations in large populations can be compartmentalized into information-limiting and nonlimiting components[17]. The limiting component is scaled by the inverse of the asymptotic information I_∞, which is where information asymptotes in the limit of a large number of neurons[17,19]. This compartmentalization allowed us to split the information I_N in a population of N neurons into the contribution of limiting and non-limiting components (see "Methods"), resulting in

$$I_N = \frac{1}{\frac{1}{cN} + \frac{1}{I_\infty}}. \tag{1}$$

This expression assumes that the non-limiting component contributes c information per neuron on average, irrespective of the current population size. Model comparison to alternative non-limiting component scaling models confirmed that this assumption best fits our data (Supplementary Fig. 6b).

Increasing the population size N in Eq. (1) reveals how information ought to scale in small populations if it is limited in large populations (Fig. 1). Information would initially grow linearly, closely following cN. However, for sufficiently large N, it would start to level off and slowly approach the asymptotic information I_∞. If we were to record from a small number of neurons, we might only observe the initial linear growth and would wrongly conclude that no information limit exists (Fig. 1). Therefore, simultaneously recording from sufficiently large populations is important to identify limited asymptotic information.

To distinguish between a population in which information does not saturate from one in which it does, we fitted two models to the measured information scaling. The first assumed that, within the recorded population, information scales linearly and without bound. We might observe this information scaling if, on average, each neuron contributes the same amount of information. The second model corresponds to Eq. (1), and assumes that information asymptotes at I_∞. Our fits relied on a large number of repetitions (at least as many as the number of recorded neurons) of the same drift direction within each experimental session to ensure reliable, bias-corrected information estimates[30]. These estimates are correlated across different population sizes, as estimates for larger populations share data with estimates for smaller populations. Unlike previous work that estimated how information scales with population size[31–33], we accounted for these correlations by fitting how information increases with each additional neuron, rather than fitting the total information for each population size. This information increase turns out to be statistically independent across population sizes (see "Methods"), making the fits statistically sound and side-stepping the problem of fitting correlated data.

Figure 4a illustrates the fit of the limited-information model to the data of a single session. We fitted the average information increase with each added neuron (Fig. 4a, top), and from this predicted the total information for each population size (Fig. 4a, bottom). Bayesian model comparison to a model that assumed unbounded information scaling confirmed that a model with limited asymptotic information was better able to explain the measured information scaling (Watanabe–Akaike Information Criterion $\text{WAIC}_{\text{unlim}} = -529.25$ vs. $\text{WAIC}_{\text{lim}} = -531.59$; smaller is better). This was the case for almost all discriminations with $\delta\theta = 45^\circ$ across sessions and mice (Supplementary Fig. 6a). Furthermore, the same procedure applied to the shuffled data resulted in better model fits for the unbounded information model, confirming that our model comparison was not a priori biased towards the limited-information model (Supplementary Fig. 6a). Two sets of simulations with idealized and realistic neural models further confirmed that this model comparison was able to recover the correct underlying information scaling (Supplementary Fig. 7). Therefore, information about drift direction is limited in the neural population responses within our dataset.

This result of limited drift direction information was corroborated by a second analysis. We start by observing that Eq. (1) can be rewritten as $1/I_N = a(1/N) + 1/I_\infty$, which is linear in the inverse population size $1/N$ with slope $a = 1/c$. Increasing the population size, $N \to \infty$, causes the inverse information to approach the asymptotic information, $1/I_N \to 1/I_\infty$. Therefore, we can distinguish between limited asymptotic information and unbounded information scaling (i.e., $I_\infty \to \infty$) by plotting $1/I_N$ against $1/N$, and estimating its intercept at $1/N \to 0$. A non-zero intercept confirms limited asymptotic information, whereas a zero intercept would suggest information to scale without apparent bounds. When we analyzed the previous single-session data, we found that the inverse information indeed tightly scales linearly with the information population size (linear regression, adjusted $R^2 \approx 1$), as predicted by the model (Fig. 4b). Furthermore, the intercept at $1/N \to 0$ was significantly above zero (linear regression, $\beta_0 \approx 0.023$, two-sided $p < 10^{-6}$), suggesting that information saturates with N. We found comparably good linear fits for all sessions/mice across all $\delta\theta = 45^\circ$ discriminations (average adjusted $R^2 \approx 0.999$; Supplementary Fig. 8a), and intercepts that were all significantly above zero ($\beta_0 \approx 0.023$, $t_{63} = 17.95$, two-sided $p < 10^{-10}$ across non-overlapping discriminations; Supplementary Fig. 8b), confirming the results of our model comparison.

In addition to supporting the distinction between information-limited and unbounded information scaling, TILC also allowed us to estimate the magnitude at which information would asymptote if we increased the population size beyond that of our recorded population. This is a theoretical measure that would be reached only for infinitely large virtual populations that have the same statistical structure as the recorded neurons. Despite this limitation, it gives insight into the order of magnitude of the information that we could expect to be encoded in the large populations of neurons present in mammalian cortices. To quantify the uncertainty associated with extrapolations beyond observed population sizes, we relied on Bayesian model fits that provide posterior distributions over our estimates of I_∞, as illustrated in Fig. 4c. These posteriors were comparable across the discrimination of different drift direction pairs (Fig. 4d). Comparable information estimates across different drift direction pairs were essential to make these estimates meaningful, as different estimates would have implied that these estimates are driven by neural subsets within a heterogeneous population rather than being a statistical property of the whole population, as desired. Furthermore, it allowed us to reduce our uncertainty in the I_∞ estimates by pooling the fits across different, non-overlapping drift direction pairs (Fig. 4d; gray). Indeed, Bayesian model comparison that accounts for the larger number of parameters of multiple individual per-discrimination fits confirmed that those were outperformed by pooled fits for all but two experimental sessions across all tested drift direction differences (Supplementary Fig. 9). This provided further evidence that, for a fixed drift direction difference, the measured information scaling was statistically indistinguishable across different discriminations within each session.

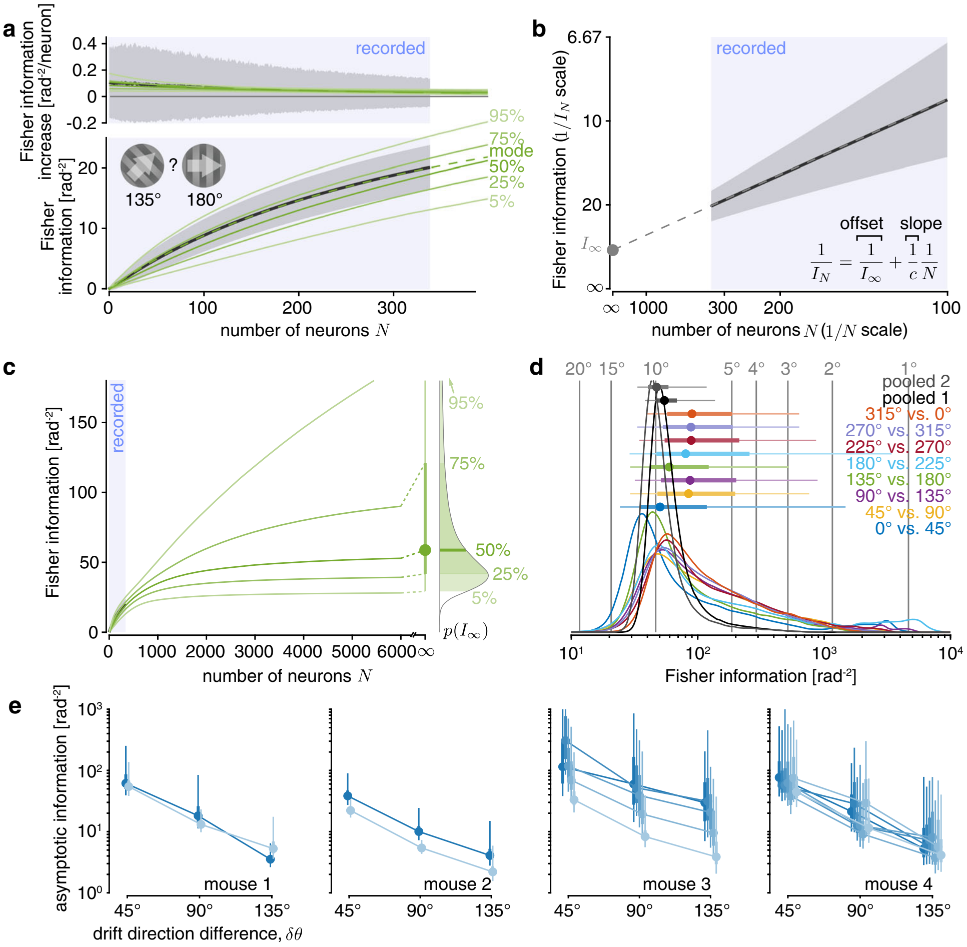


Fig. 4 Information about drift direction is estimated to asymptote in large neural populations. a Example information scaling fit, showing data (black; mean estimate ± 1 SD; computed from 135° vs. 180° drift direction trials, as in Fig. 3c) and posterior predictive density for Bayesian fit (green; solid = percentiles, dashed = mode) for the Fisher information increase (top) and Fisher information (bottom) across different population sizes N. The model is fitted to the Fisher information increase estimates (top), as these are statistically independent across different population sizes. **b** Plotting the inverse Fisher information $1/I_N$ over the inverse population size $1/N$ (mean estimate ± 1 SD; same data as in **a**) shows an almost perfect linear scaling, as predicted by our theory. Fitting a linear model (gray dashed line) reveals a non-zero asymptotic information I_∞ (gray dot) with $N \to \infty$ **c**. The fitted model supports extrapolating the posterior predictive density beyond recorded population sizes (blue shaded area in **a**–**c**) up to $N \to \infty$. This results in a Bayesian posterior estimate over the asymptotic information I_∞ (right), which we summarize by its median (dot), and its 50% (thick line) and 90% (thin line; truncated at top) credible intervals. **d** Estimates of asymptotic information resulting from different drift direction pairs (colors; $\delta\theta = 45^\circ$ for all pairs) results in comparable posterior densities (colored lines; associated density summaries above densities as in **c**) across different pairs. Therefore, we pooled the data across all non-overlapping pairs with the same $\delta\theta$ to achieve a more precise estimate. The pooled estimates were comparable across two different sets of non-overlapping pairs (gray). The vertical gray lines and numbers indicate the drift direction discrimination thresholds corresponding to different Fisher information estimates. **e** The asymptotic Fisher information estimate (density summaries as in **c**; lines connect posterior medians) is comparable across sessions (different colors; horizontally shifted to ease comparison) and mice.

Comparing these pooled estimates across sessions and mice revealed these estimates to be similar (Fig. 4e). These estimates dropped with an increase in the angular difference $\delta\theta$ in the compared drift directions, as is to be expected from a linear decoder used to discriminate between circular quantities (Supplementary Fig. 10). Together, these observations strongly suggest that the recorded populations were part of a larger population that encoded limited information about the drift direction of the presented stimuli.

No optimal neural subpopulation across all drift directions. The recorded population might contain neurons that are not only untuned to drift direction but also do not contribute information through being correlated with other neurons in the population[20,21]. As our information scaling measures are averaged across different orderings of how neurons are added to the population, uninformative neurons would contribute at different population sizes across different orderings. As a result, they make information scaling curves appear shallower than for populations that exclude uninformative neurons. These shallower scaling curves could in turn impact our estimates of asymptotic information (Fig. 4).

To ensure that uninformative neurons did not significantly affect our estimates, we asked if we could identify neural subpopulations within the set of recorded neurons that encode most of the information. Previous work identified such subpopulations in auditory cortex[34] and lateral prefrontal cortex[20] of monkeys, but we are not aware of any work that has shown this for V1. To identify highly informative subpopulations, we ordered the neurons within the recorded population by incrementally adding the neuron that resulted in

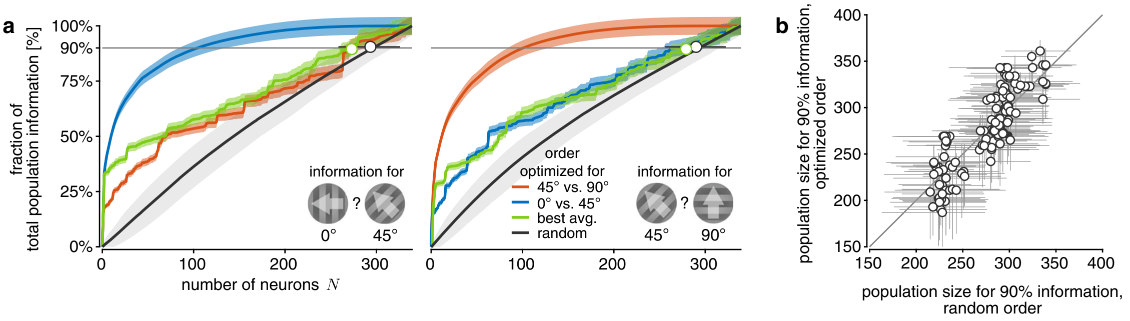


Fig. 5 No single neural subpopulation appears to encode a disproportionate amount of information across all stimulus drift directions. a Both panels show that the information increase in the recorded population depends on the order with which neurons are added to the population (colors). The panels differ in the considered drift direction discrimination (left: 0° vs. 45°; right: 45° vs. 90°). The neuron order was optimized by incrementally adding the neuron that resulted in the largest information increase for a 0° vs. 45° (blue) or 45° vs. 90° (orange) drift direction discrimination, or largest average increase across all discriminations with $\delta\theta = 45^{\circ}$ (green). The optimal ordering for the 0° vs. 45° was also applied to the 45° vs. 90° discrimination (blue line in right panel) and vice versa (orange line in left panel). The average information increase across random orders (black) is shown as baseline reference. Shaded error regions illustrate the uncertainty (mean ± 1 SD) due to limited numbers of trials (all curves), and variability across random orderings (black only). The black and green open circle (bootstrapped median ± 95% CI) show the population sizes required to capture 90% of the information in the recorded population for the associated orderings. **b** Plotting population sizes required to capture 90% of the information in the recorded population (bootstrapped median ± 95% CI) for random ordering vs. orderings optimized to maximize average information across all discriminations revealed no significant difference between the two orderings. Each dot reflects one discrimination for one session.

the largest overall information increase[20,34]. With this ordering, 90% of the information in the recorded population for a particular discrimination could be recovered from only about 30% of the recorded neurons (Fig. 5a). However, natural behavior usually requires information about a wide range of different drift directions rather than the ability to discriminate a specific drift direction pair. To identify how much information the discovered subpopulation contains about other drift directions, we asked how well its population activity supports discriminating another, close-by drift direction pair (Fig. 5a; left vs. right). We found that the same subset of neurons was only able to recover about 55% of the information about this new discrimination. Even a population ordering that boosted the average information across all drift direction pairs did not reveal a highly informative subpopulation within the recorded set of neurons (Fig. 5a; green). To determine whether there is any advantage to a particular ordering, we estimated the population size required to capture 90% of information of the recorded population if we ordered the neurons according to this objective. Across sessions/mice and discriminations, the required population size turns out to not differ significantly compared with a random ordering of the population (Fig. 5b; $t_{63} = -0.215$, two-sided $p \approx 0.83$; across non-overlapping $\delta\theta = 45^{\circ}$ discriminations). Noise correlations contribute to the observed lack of difference, as this difference becomes significant for trial-shuffled data (Supplementary Fig. 11). If a significant fraction of neurons is uninformative across all drift direction pairs, we would expect these population sizes to differ. Therefore, it is unlikely that our asymptotic information estimates were significantly influenced by the presence of uninformative neurons in the recorded populations.

Finite-population information impacts asymptotic information. If estimated asymptotic information mirrors the total information encoded by the animals' brains, it should increase if we increase the amount of information provided by the stimulus in retinal photoreceptor activity. As has been shown previously, higher contrast stimuli result in higher decoding performance from recorded population responses (e.g., see ref. [35]). However, we might observe an information increase in recorded populations even when the asymptotic information remains unchanged (Fig. 6c, right). To determine if increasing the stimulus contrast results in an increase of asymptotic information, we performed a separate set of experiments in which two mice observed the same drift directions as before, but with a grating contrast of either 10% or 25% that was pseudo-randomly chosen across trials. We hypothesized that the 25% contrast stimuli provide more information about the drift direction, and expected a corresponding increase in asymptotic information.

For most neurons, a contrast increase from 10 to 25% led to a change in baseline activity and re-scaling of their tuning curves, but no appreciable change in pairwise noise correlations (Supplementary Fig. 12). As in correlated populations we cannot predict changes in information solely from changes in tunings, we again moved to measuring information by our generalized Fisher information measure. This revealed that information encoded in the recorded populations significantly increased for higher stimulus contrasts (Fig. 6a for single discrimination and session; Fig. 6b for all sessions/mice, non-overlapping discriminations with $\delta\theta = 45^{\circ}$: paired $t_{27} = 2.78$, two-sided $p \approx 0.0098$). We in turn applied the same procedure as before (see Fig. 4e) to estimate asymptotic information, but did so separately for the two contrasts (Fig. 6d). We then compared these estimates for $\delta\theta = 45^{\circ}$ within each session between low- and high-contrast trials (Fig. 6d). In principle, increasing contrast could increase asymptotic information, or it could leave asymptotic information unchanged (Fig. 5c). For three out of the four sessions in which information in the recorded population increased with contrasts for a majority of discriminations (as shown in Fig. 6b), we also observed an increase in asymptotic information with contrast (Fig. 6e, filled dots). This suggests that a more informative stimulus not only increased information in the recorded neural populations but also in the larger (unrecorded) neural population.

Tens of thousands of neurons decode most of information. Information in the brain must saturate, as noisy sensors fundamentally limit the sensory information it receives. However, it remains unclear whether information saturates within the population size of V1 (Fig. 1). In our information scaling model, Eq. (1), saturation by definition only occurs in the limit of infinite neurons. We can nonetheless use the model to estimate saturating population sizes by asking how large these populations need to be to encode a large fraction of the asymptotic information (Fig. 7a).

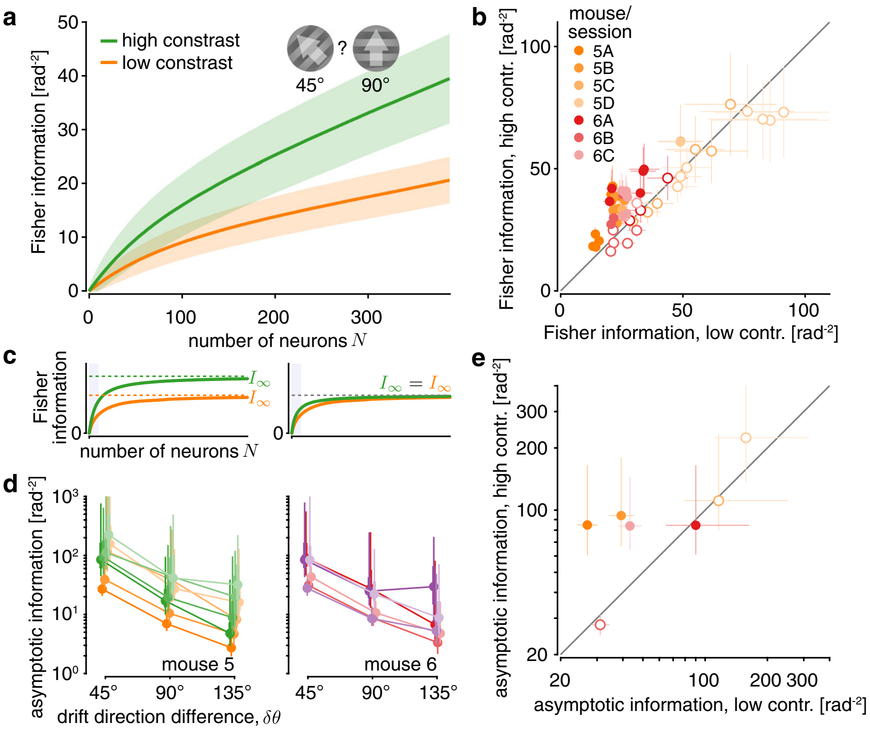


Fig. 6 Increasing stimulus contrast boosts asymptotic information in V1. a The information increases more rapidly with population size for high-contrast stimuli (green; mean ± 1 SD) than for low-contrast stimuli (orange; mean ± 1 SD), here shown for the discrimination between 45° vs. 90° drift direction trials of one session of mouse 5. An increase in stimulus contrast significantly increases the total information in the recorded population (bootstrap, two-sided $p < 10^{-5}$). **b** The information in the recorded population was larger for high than low stimulus contrast for most $\delta\theta = 45^\circ$ discriminations across mice and sessions (colors, XY = mouse X, session Y). Each dot shows the information for one discrimination between different drift directions (8 dots per session; error bars = ± 1 SD of the information estimation uncertainty). Filled dots indicate a significant information increase (bootstrap, two-sided $p \geq 0.05$). **c** Observing an information increase in the recorded population (blue shaded area) does not necessarily imply an increase in asymptotic information (left vs. right). **d** The estimated asymptotic information was generally higher for high-contrast stimuli (green/magenta; shades = sessions) than low-contrast stimuli (orange/red; shades = session; colors as in panel **b**), across different drift direction differences $\delta\theta$ (pooled estimates across different drift direction pairs; posterior density summaries as in Fig. 4c). **e** To compare the pooled asymptotic information estimates for $\delta\theta = 45^\circ$ for low-contrast trials to those for high-contrast trials, we plot them against each other (one dot per session, colors as in panel **b**, error bar centers = posterior medians, error bars = 50% posterior credible intervals). The four filled dots indicate sessions for which the information in the recorded population (panel **b**) is significantly larger for higher contrast stimuli for the majority of discriminations.

We will here focus on population sizes N_{95} that achieve 95% of asymptotic information, which can be found by setting $I_N = 0.95 I_\infty$ in Eq. (1) and solving for N. The required population sizes for other fractions of asymptotic information are easily found by a rescaling of N_{95} (Supplementary Fig. 13).

To estimate N_{95}, we again relied on the information scaling fits pooled across non-overlapping pairs of drift directions. The recovered population sizes were all on the order of tens of thousands of neurons (Fig. 7b). Our previous analysis (Fig. 5) makes it unlikely that uninformative neurons within the recorded population strongly impact our estimated population sizes. Interestingly, increasing the drift direction difference $\delta\theta$ did not strongly affect these estimates (mice 1–4 in Fig. 7b), even though it modulated asymptotic information (Fig. 4d). Increasing stimulus contrast appeared to increase the estimated population sizes (mice 5–6 in Fig. 7b, orange vs. green), but not consistently so. Thus, it was unclear if a change in information resulted in a global re-scaling of the information scaling curve without changing its shape (Fig. 7c, top), or in the need for more neurons to encode this information (Fig. 7c, bottom).

To clarify the relationship between the asymptotic information I_∞ and required population size N_{95}, we did not directly relate these two quantities, as N_{95} is derived from the estimate of I_∞. Instead, we relied on the property that N_{95} is proportional to I_∞/c, where c is the scaling factor associated with the non-limiting covariance component (see Eq. (1); Methods). Therefore, if N_{95} remains constant across different estimates of I_∞ and c, these two quantities need to vary in proportion to each other. In a log–log plot, this implies that the slope describing their relationship would be one. However, we found a slope of $\beta_1 \approx 0.72$, which is slightly, but significantly below one (Fig. 7d; F-test, $F_1 = 21.49$, $p \approx 1.2 \times 10^{-5}$). Substituting the measured relationship between c and I_∞ into the expression for N_{95} results in $N_{95} \approx 4523.8 I_\infty^{0.28}$. This implies that the population size required to encode 95% of the asymptotic information increases with I_∞, but does so only weakly. To illustrate this weak increase, let us consider sessions in which the estimated asymptotic information increased threefold with an increase in stimulus contrast (Fig. 6e). In this case, a population of the size required to capture 95% of the asymptotic information for low-contrast trials could capture 93% of the asymptotic information for high-contrast trials (see “Methods”).

Information is not aligned with principal noise dimensions. Previous work has observed that most neural population activity fluctuations are constrained to a low-dimensional linear subspace that is embedded in the high-dimensional space of neural activity[36–38]. This might suggest that focusing on such a low-dimensional subspace is sufficient to understand brain function[38]. Thus, we asked if we can recover most of the information about

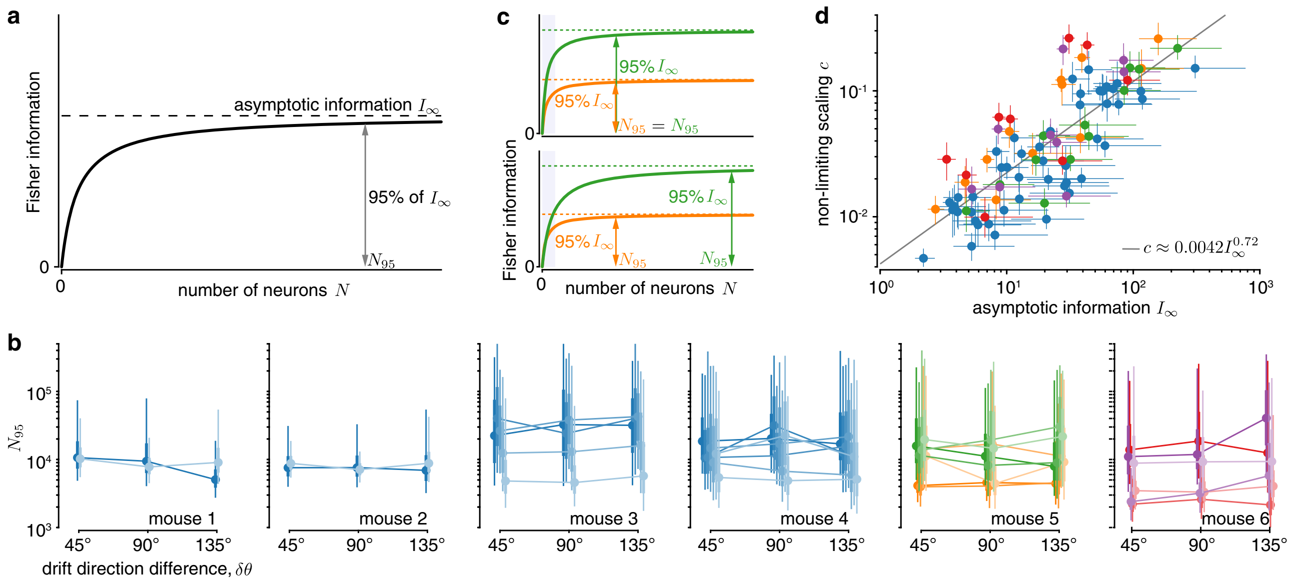


Fig. 7 Tens of thousands of neurons are required to capture most of the information about stimulus drift direction. a The TILC predicts how information grows with population size and allows us to estimate the population size N_{95} required to capture 95% of the asymptotic information I_∞. **b** Applied to our data, we find N_{95} in the order of tens of thousands of neurons, consistently across mice (panels) and sessions (colors; blue = uniform contrasts; orange/red = low contrast; green/magenta = high contrast; lines connect individual sessions; horizontally shifted to ease comparison; posterior densities as is Fig. 4c). **c** An increase in information (orange to green) could be achieved by increasing the average information per neuron (top) or by leaving the average information per neuron roughly unchanged while recruiting more neurons (bottom). We would expect N_{95} to grow in the second, but not the first case. These two cases are hard to distinguish from the observed information scaling in smaller populations (shaded blue). **d** Plotting estimated non-limiting scaling c over asymptotic information I_∞ (dot = median, lines = 50% credible interval; colors as in panel **b**) for all animals, sessions, drift direction differences, and contrasts from **b** reveals that c grows sub-linearly with I_∞ (gray line = linear regression of median estimates in log-log plot), which indicates that the estimated population size N_{95} increases weakly with the asymptotic information.

visual drift direction from such subspaces, defined by the dimensions where population activity is most variable. The information encoded in each dimension grows with how well the signal, $\mathbf{f}'$, is aligned with this dimension, but shrinks with the magnitude of noise in this dimension (Fig. 8a; see refs. [17,33]). This tradeoff makes it unclear whether the subspace where population activity is the most variable is indeed the subspace that encodes the most information.

We found the principal dimensions of the noise covariance matrix and asked how much information a subset of the most variable dimensions is able to encode. In our data, 90% of the total variance was captured by approximately 37.6% ± 12.4pp (mean % ± 1 SD percentage points across all sessions/mice, $\delta\theta =$ 45° discriminations) of all available dimensions (Fig. 8b/e), confirming previous reports that relatively few dimensions are required to capture most noise variance. Furthermore, $\mathbf{f}'$ was most strongly aligned to the first few of these principal dimensions[33] (Fig. 8c). Using cosine similarity to measure this alignment, we found that 90% of the cumulative alignment was reached by approximately 7.4% ± 9.1pp of all available dimensions (Fig. 8c/e). Finally, we asked how many dimensions were required to capture 90% of the information encoded in the recorded population. Even though later dimensions were not well-aligned with $\mathbf{f}'$ (see the shallow cumulative alignment increase in Fig. 8c), they were also less noisy (Fig. 8b) and so could contribute significantly to the encoded information. As evident by the continual information growth in Fig. 8d, this resulted in information which was fairly evenly spread across all dimensions, such that, on average, approximately 86.7% ± 2.2pp of all principal noise dimensions were required to encode 90% of all of the recorded information. This is significantly higher than the fraction required to capture 90% of all variance (difference = 48.7 ± 1.5pp, mean ± 1 SEM, paired $t_{63} = 32.53$, two-sided $p < 10^{-6}$ across non-overlapping discriminations). In fact, if we restricted ourselves to the subspace that captures 90% of all noise variance, we could only decode 58.9% ± 5.6pp of information. Therefore, in our data, relying only on information encoded in the subspace of most variable principal dimensions would result in significant information loss.

Discussion

We asked how information about the drift direction of a visual stimulus is distributed in large neural populations, and addressed this question by analyzing how information scales with population size. We observed that, in recorded populations, information scaled sublinearly with population size, indicating that noise correlations limited this information. The information scaled in line with TILC if information is indeed limited in larger populations. Based on this theory, we found that we require on the order of tens of thousands of neurons to encode 95% of the asymptotic information. When varying input information by changing stimulus contrast, the required population size appeared to change. Indeed, we found that more information required larger populations, but this relationship was extremely weak. Overall, these findings suggest the presence of information-limiting correlations that cause sensory information in mouse V1 to saturate with population size, indicating the use of a highly redundant, distributed neural code within mouse V1.

Previous attempts at measuring how sensory information scales with population size have frequently found noise correlations to either be beneficial[39] or to not affect information scaling[32,33]. These studies focused on smaller populations (<200 neurons in ref. [39]; <100 neurons in ref. [33]) in which sublinear scaling might be hard to identify (Fig. 1), and in part included spike timing information[39] in addition to the spike counts used

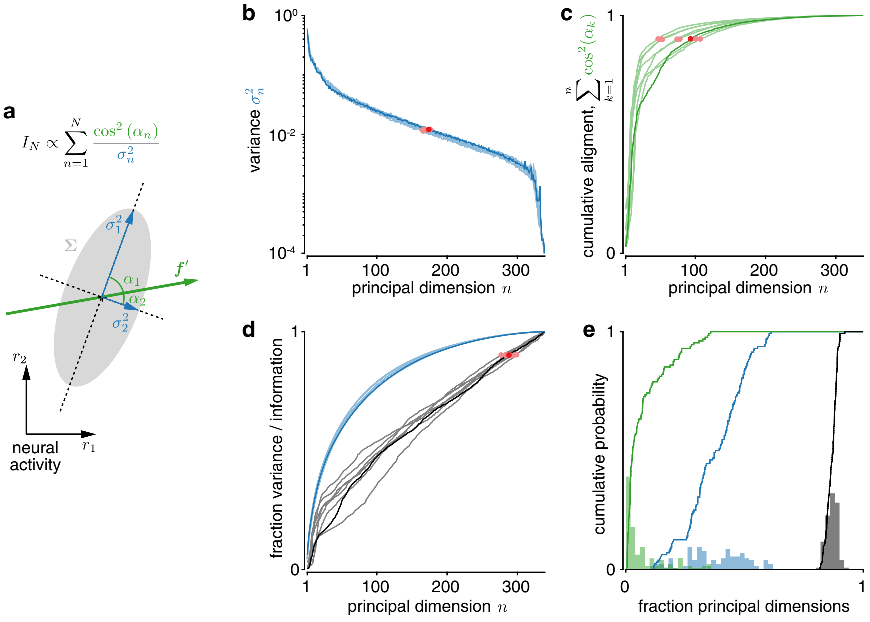


Fig. 8 Information is not well-aligned with principal noise dimensions. a The total information in a recorded population, I_N, can be decomposed into how well the change in population tuning, f', is aligned to the different principal dimensions of the noise covariance (given by $\cos^2(\alpha_n)$), as well as the noise variances, σ_n^2, in these dimensions. Low-variance dimensions that are strongly aligned to f' contribute more information. **b** The variance along principal dimensions drops rapidly with dimension (red dots: >90% of total variance). **c** The cumulative alignment of f' to the principal dimensions of the noise covariance rises rapidly with principal dimension (red dots: >90% of total alignment), indicating that f' is most strongly aligned to the first few principal dimensions. **d** The fraction of total information (black; red dots: >90% information) rises more slowly with additional principal dimensions than the fraction of total noise variance (blue). The principal dimensions in panels **b**–**d** are ordered in decreasing order of variance, and show data for the same session as in Fig. 3c (dark = same discrimination as in Fig. 3c; light = other $\delta\theta = 45^\circ$ discriminations for that session). **e** The histograms (bar plots) and cumulative probabilities (lines) of the fractions of the total number of principal dimensions at which the cumulative f' alignment (green), cumulative variance (blue), and cumulative information (black) exceed 90% of their respective totals. These fractions are shown for all $\delta\theta = 45$ discriminations, sessions, and mice. All estimates are cross-validated, and averaged across ten train/test splits (see "Methods").

here. Recent recordings from ~20,000 neurons in mouse V1 suggest information about visual stimuli does saturate[40], but it appears to do so above the population sizes we estimated. These recordings used a slower image scan rate (3 Hz vs. the 30 Hz used for this study), which introduces additional recording noise. This additional noise makes information saturate more slowly with population size (see SI, Sec. 2.3), potentially explaining the larger required population sizes. Recordings from hundreds of neurons in monkey prefrontal cortex revealed sublinear scaling of motor information, compatible with the presence of information-limiting correlations, and resulted in required population size estimates comparable to ours[31]. In contrast to our study, this work measured information about saccade direction rather than about sensory stimulus features. Furthermore, it relied on data from two saccade directions only, and so could not assess if a smaller, selected subpopulation could be used to decode a significant fraction of the total information across a wide range of saccade directions, as we do for drift directions.

Even though information is highly distributed across neurons in a population, most variability is captured by a low-dimensional subspace, leading to suggestions that we might only need to consider the information encoded in this subspace[38]. As we have shown, this argument does not consider that information does not only depend on variability, but also on how the signal aligns with this variability (Fig. 8a). Once both are taken into account, the dimensions of largest variability become a poor proxy for the most informative dimensions (Fig. 8d). This is in line with recent work showing that the most variable subspace in macaque V1 is different from the one that most co-varies between V1 and V2 (ref. [37]), which presumably transmits information between these areas. Our work explicitly shows such misalignment, and does so in larger populations.

To compare our required population size estimates to the total number of neurons in mouse V1, we conservatively estimated the need for about 48,000 neurons (see "Methods") to achieve drift direction discrimination performance that most likely exceeds that of the animals[23–25]. Our use of time-deconvolved calcium activity as a noisy proxy for spike counts[41,42] makes these estimates upper bounds on required population sizes (see SI). Nonetheless, they compare favorably to the number of neurons in mouse V1, whose estimates range from 283,000 to 655,500 (refs. [43,44]). If we instead compare to the number of neurons in V1 that correspond to the retinotopic area of the visual stimulus, using the entire stimulus or only the full-contrast portion as best and conservative worst-case scenarios, we estimate that the lower and upper bounds on the responsive number of neurons are the same to 10 times higher than our required population size estimates (see "Methods"). This confirms that mouse V1 has more neurons than required to encode most of the estimated asymptotic information about the direction of a moving visual stimulus. Would fewer neurons be required to encode information about natural scenes, which tend to evoke sparser population responses than drifting gratings[45–47]? We do not expect this to be the case, as the fraction of neurons that respond to individual natural stimuli are in fact lower than for drifting gratings, but overall more neurons are required to represent a broad set of natural stimuli[45,47]. This implies that, as for drifting gratings (Fig. 5), we cannot focus on smaller subpopulations that might well discriminate specific image pairs[47], but might fail to convey information about other natural images.

If animals are required to perform tasks that rely on the encoded information we measured (e.g., to discriminate between different drift directions), each neuron in the population would ideally contribute to the animal's choices. Quantified by choice correlations[48,49], an optimal read-out requires the choice correlations of individual neurons to be the fraction of the population's discrimination threshold over that of the neuron[50]. In contrast to previous work (e.g., refs. [51,52]) that found that individual neurons' thresholds match that of the animal, the neurons' average threshold in our data (see information for $N = 1$ in Fig. 3c) is exceedingly small when compared to that of the recorded population (Fig. 3c for full population), and even smaller when compared to estimated asymptotic information (Fig. 4e). This mismatch might arise from shorter stimulus presentation, not tailoring the stimuli to match the neuron's tuning (as done in Britten et al.[51]), recording from lower-level visual areas (V1 vs. V4 or MT) with smaller receptive fields, as well as increased recording noise with calcium imaging as compared to electrophysiological recordings. These lower discrimination thresholds predict increasingly small choice correlations, in line with recent reports from area V1 of monkeys, where fewer than 7% of V1 neurons were found to feature significant choice correlations[53]. In general, the estimated asymptotic information predicted direction discrimination thresholds compatible with previous behavioral reports in mice[23–25], but the use of different stimuli in these experiments precludes a direct quantitative comparison. We furthermore cannot exclude the possibility that mice used a different read-out than the linear one we assumed, or lacked motivation to perform the task to their full potential, further impacting their behavioral performance. A more detailed analysis of the relation between neural activity and choice would require training animals to report their percepts, and then relating these reports to population activity fluctuations.

Multiple factors could have impacted our information measures, and with them our asymptotic information and discrimination threshold estimates. First, the mouse's state of arousal, commonly assessed by their pupil dilation, has been found to fluctuate during similar experiments[28], and such fluctuations could modulate information encoded in V1. Locomotion is linked to arousal[28], and has previously been shown to impact information[29]. In our data, periods of increased locomotion also result in more information in the recorded populations and increase asymptotic information estimates, but do not significantly affect the estimated population sizes required to encode 95% of this asymptotic information (Supplementary Fig. 4). Second, any eye movement within the stimulus presentation period will shift the association between the stimulus and the cells' receptive fields, and result in a relative drop in information. Our stimulus was designed to minimize the effect of eye movements occurring between consecutive stimuli (see "Methods"). Furthermore, eye movement in mice tend to be rare[54] and small[54,55] when compared to the V1 neuron receptive field sizes[56] and size of our stimulus, such that we expect them to have little effect on our estimates of information-limiting correlations. This was confirmed in simulations and theoretical analysis of a simple eye movement model, which revealed that the assumed eye movements might result in over-estimating N_{95}, but only in a minor underestimation of I_∞ (Supplementary Fig. 14). Third, we used calcium imaging to obtain dense sampling from large neural populations. Although viral expression of GCaMP6s, as we used here, has been shown to detect nearly all single spikes in some conditions[41], with our imaging conditions, it is likely that we were unable to detect some single spikes. Furthermore, saturation of GCaMP responses might have caused a non-linear mapping between spike counts and measured GCaMP responses, which would quantitatively lower the measured information, but not qualitatively impact how information scales with population size (Supplementary Fig. 15). Also, neuropil fluorescence has the potential to create shared changes in nearby neurons[57]. We expect that neuropil contamination is unlikely to have a major impact on our information scaling results because such contamination would create redundant signals across neurons and would thus have little impact on information levels that must arise from genuine, non-redundant signals in neurons. However, it is possible that neuropil contamination could have made some uninformative neurons appear informative, in which case a smaller fraction of neurons might be genuinely informative than suggested by Fig. 5. Moreover, residual neuropil fluorescence could cause the non-recorded neuron's signal to "leak out" to recorded neurons, which might result in an underestimation of N_{95}. In general, only those factors that modulate information-limiting correlations, which are a small component of the overall noise correlation matrix, impact our information estimates (illustrated in Supplementary Fig. 3). Therefore, while we cannot rule out the presence of such factors, we expect that they did not qualitatively impact our findings.

A prediction of our findings is that neural information should continue to scale according to Eq. (1) in larger populations than those recorded in our experiments. Testing these predictions involves precise estimates of noise correlations, which require about the same number of trials in which the same stimulus (e.g., drift direction) is presented as there are neurons in the population[17,19]. Therefore, even with more powerful recording techniques, information estimates might be limited by the number of trials that can be collected within individual sessions. The use of decoders to estimate information might sidestep these estimates[30,31], with the downside of potentially confounding decoder biases. A further challenge is to record from a population that homogeneously encodes the same amount of information about each stimulus. Such homogeneity ensures that the estimated asymptotic information and population sizes are not specific to particular stimulus values. The weak spatial organization of drift direction selectivity in mouse V1 (ref. [58]) supports this, but the same would be harder to achieve in monkeys due to the much stronger spatial correlations of orientation and direction selectivity in their visual cortices[59]. Finally, even if Eq. (1) is confirmed to match the information in larger populations than used here, it does not allow us to guarantee that the cortex's information is limited by sensory noise and suboptimal computations. Though unlikely, information might continue to grow linearly after an initial sublinear growth[16]. The only way to conclusively rule out this scenario is to record from all neurons in the information-encoding population, which, at least in mammals, will likely not be possible in the foreseeable future[60].

Although all information entering the brain is limited by sensory noise[6], such that it can never grow without bound, the information could be so plentiful or broadly distributed across multiple independent chunks as to not saturate within the population sizes of mammalian sensory areas. In this case, we would expect information to grow on average linearly with the recorded population size, as has been frequently observed in smaller populations. Our findings suggest this not to be the case. However, we suspect the main limiting factor not to be noisy sensors. Instead, most problems that the brain has to deal with require fundamentally intractable computations that need to be approximated, resulting in substantial information loss[61]. Indeed, suboptimal computations can dominate overall information loss, and resulting behavioral variability[62,63], such that they might be the main contributor to the information limitations we observe in our experiments.

If the brain operates in a regime in which information in sensory areas is limited, all information the brain deals with is uncertain. This idea finds support in the large body of work showing that behavior is well-described by Bayesian decision theory[64–66], which makes effective use of uncertainty. This, in turn, implies that the brain encodes this uncertainty, but its exact neural representations remain unclear[66,67]. A further consequence of limited information is that theories that operate on trial averages (e.g., refs. [68–70]) or assume essentially unlimited information (e.g., ref. [16]) only provide an incomplete picture of the brain's operation. Therefore, an important next step is to refine these theories to account for trial-by-trial variation in the encoded information to achieve a more complete picture of how the brain processes information in individual trials, rather than on average.

Methods

All experimental procedures were approved by the Harvard Medical School Institutional Animal Care and Use Committee (IACUC).

Animals and surgery. Male C57BL/6J mice were obtained from The Jackson Laboratory and housed at 65–75 °F with 35–65% humidity and on a 12-h reverse light/dark cycle. Mice were used for imaging experiments between 4 and 7 months of age. Prior to imaging, mice underwent surgery to implant a chronic cranial window and headplate. Mice were injected intraperitoneally with dexamethasone (3 μg per g body weight) 3–6 h before surgery to reduce brain swelling. During surgery, mice were stably anesthetized with isoflurane (1–2% in air). A titanium headplate was attached to the skull using dental cement (C&B Metabond, Parkell). A ~3.5-mm diameter craniotomy was made over left V1 (stereotaxic coordinates: 2.5 mm lateral, 3.4 mm posterior to bregma). AAV2/1-syn-GCaMP6s (Penn Vector Core) was diluted into phosphate-buffered saline at a final titer of ~2.5E12 gc/ml and mixed 10:1 with 0.5% Fast Green FCF dye (Sigma-Aldrich) for visualization. Virus was injected in a 3 × 3 grid with 350 μm spacing near the center of the craniotomy at 250 μm below the dura, with ~75 nl at each site. Injections were made slowly (over 2–5 min) and continuously using beveled glass pipettes and a custom air pressure injection system. The pipette was left in place for an additional 2–5 min after each injection. Following injections, the dura was removed. A glass plug consisting of two 3.5-mm coverslips and one 4.5-mm coverslip (#1 thickness, Warner Instruments) glued together with UV-curable transparent optical adhesive (Norland Optics, NOA 65) was inserted into the craniotomy and cemented in place with cyanoacrylate (Insta-Cure, Bob Smith Industries) and metabond mixed with carbon powder (Sigma-Aldrich) to prevent light contamination from the visual stimulus. An aluminum ring was then cemented on top of the headplate, which interfaced with the objective lens of the microscope through black rubber light shielding to provide additional light-proofing. Data from mouse 1 and 2 were collected as part of a previously published study[71], following a similar surgical protocol. Imaging datasets were collected at least 2 weeks post-surgery, and data collection was discontinued once baseline GCaMP levels and expression in nuclei appeared to be high.

Visual stimuli. Visual stimuli were displayed on a gamma-corrected 27-inch IPS LCD gaming monitor (ASUS MG279Q). The monitor was positioned at an angle of 30° relative to the animal and such that the closest point to the mouse's right eye was ~24 cm away, with visual field coverage ~103° in width and ~71° in height. Visual stimuli were generated using PsychoPy[72] or Psychtoolbox (for mice 1 and 2 only) and consisted of square-wave gratings presented on a gray background to match average luminance across stimuli. Gratings were windowed outside of a central circle of radius 20° with a Gaussian of 19° standard deviation, or windowed with a Gaussian central aperture mask of 44° standard deviation (for mice 1 and 2 only) to prevent monitor edge artifacts. Grating drift directions were pseudorandomly sampled from 45° to 360° in 45° increments at 10 or 25% contrast, spatial frequency of 0.035 cycles per degree, and temporal frequency of 2 Hz. Stimuli were presented for 500 ms, followed by a 500 ms gray stimulus during the inter-stimulus interval (1 Hz presentation). Digital triggers from the computer controlling visual stimuli were recorded simultaneously with the output of the ScanImage frame clock for offline alignment. The visual stimulus was designed to be minimally sensitive to the small eye movements typical of mice[54,55]. In addition to using a full field grating, the stimulus presentation of 500 ms and temporal frequency of 2 Hz was chosen so that each trial consisted of exactly one complete cycle. The effect of fixational eye movements was thus mostly a small shift in phase of the perceived stimulus, which should have little impact on spike counts summed over the full stimulus presentation.

Microscope design. Data were collected using a custom-built two-photon microscope. A Ti:Sapphire laser (Coherent Chameleon Vision II) was used to deliver 950 nm excitation light for calcium imaging through a Nikon 16 × 0.8 NA water immersion objective, with an average power of ~60–70 mW at the sample. The scan head consisted of a resonant-galvonometric scanning mirror pair separated by a scan lens-based relay. Collection optics were housed in a light-tight aluminum box to prevent contamination from visual stimuli. Emitted light was filtered (525/50, Semrock) and collected by a GaAsP photomultiplier tube (Hamamatsu). Microscope hardware was controlled by ScanImage 2018 (Vidrio Technologies). Rotation of the spherical treadmill along three axes was monitored by a pair of optical sensors (ADNS-9800) embedded into the treadmill support communicating with a microcontroller (Teensy, 3.1). The treadmill was mounted on an XYZ translation stage (Dover Motion) to position the mouse under the objective.

Experimental protocol. Before data acquisition, mice were habituated to handling, head-fixation on a spherical treadmill[73], and visual stimuli for 2–4 days. For each experiment, a field-of-view (FOV) was selected. Multiple experiments conducted in each animal were performed at different locations within V1 or different depths within layer 2/3 (120–180 μm below the brain surface). Before each experiment, the monitor position was adjusted such that a movable flashing stimulus or drifting grating in the center of the screen drove the strongest responses in the imaged FOV, as determined by online observation of neural activity. A single experiment consisted of three blocks of ~45 min each. Once a FOV was chosen, a baseline image (~680 × 680 μm) was stored and used throughout the entire experiment to compare with a live image of the current FOV and manually correct for axial and lateral drift (typically <3 μm between blocks and <10 μm over the full experiment) by adjusting the stage. Drift and image quality stability were verified post hoc by examining 1000 × sped-up movies of the entire experiment after motion correction and temporal downsampling, and experiments that were unstable were discarded without further analysis. Data from mouse 1 and 2 were from previously published experiments[71], where a small fraction of neurons were photostimulated simultaneous to drifting gratings presentation. All photostimulated neurons were excluded from analysis for this paper.

Data processing. Imaging frames were first motion-corrected using custom MATLAB code (https://github.com/HarveyLab/Acquisition2P_class) on subframe, full-frame, and long (minutes to hours) timescales. Batches of 1000 frames were corrected for rigid translation using subpixel image registration, after which frames were corrected for non-rigid warping on sub-frame timescales using a Lucas-Kanade method. Non-rigid deformation on long timescales was corrected by selecting a global alignment reference image (average of a 1000-frame batch) and aligning other batches by fitting a rigid 2D translation, followed by an affine transform and then nonlinear warping. After motion correction, due to large dataset size (~130 GB), imaging frames were temporally downsampled by a factor of 25 from 30 to 1.2 Hz. Downsampled data were used to find spatial footprints, using a modified version of the constrained nonnegative matrix factorization (CNMF) framework[74] (https://github.com/Selmaan/NMF-Source-Extraction). Three unregularized background components (instead of the default number, one) were used to model spatially and temporally varying neuropil fluorescence, as we observed that the spatial footprints of neuropil activity were distinct from the GCaMP baseline fluorescence background component. We modified the procedure used by CNMF to initialize sources, and instead used an approach to identify sources independently of their spatial profile by using a procedure to cluster pixels based on temporal activity correlations[71]. These sources were then used as initializations for subsequent iterations of the original CNMF algorithm. The resulting spatial footprints from CNMF were used to extract full temporal-resolution fluorescence traces for each source. Traces were deconvolved using the constrained AR-1 OASIS method[75] and individually optimized decay constants. To obtain dF/F, CNMF traces were divided by the average pixel intensity in the absence of neural activity (i.e., the sum of background components and inferred baseline fluorescence from deconvolution of the source's CNMF trace). Because our modified version of CNMF returned sources with both cell-shaped and irregular spatial profiles, we used a convolutional neural network trained on manually annotated labels to classify sources as cell bodies, axial processes (bright spots), horizontal processes, or unclassified. Only data from cell bodies were used in this paper.

To assess neural variability in our recordings, we computed the coefficient of variation (CV; i.e., relative standard deviation) for orientation- and direction-tuned neurons. We found this CV to be roughly one on average, which compares favorably to previously reported mouse V1 data. Bennett et al.[76], for example, found in whole-cell patch clamp recordings a CV of between ~1 (moving) to 2 (stationary) in response to drifting sinusoidal gratings. De Vries et al.[45] found a higher CV of ~2.5 from two-photon calcium imaging data in response to drifting gratings. As fluorescence responses are scaled by some unknown, arbitrary factor relative to spiking activity, we could not compute the neurons' Fano factors. This scaling did not impact our linear Fisher information estimates, as these estimates are invariant to (invertible) linear transformations of neural activity.

Tuning curve fits. We used three nested models to fit tuning curves for each neuron. In the direction-tuned model, the average neural response of each neuron was fitted by a mixture of two Von Mises function given by

$$f_1(\theta) = a + b_1 \exp\left(c\cos\left(\theta - \theta_{\text{preferred}}\right)\right) + b_2 \exp\left(-c\cos\left(\theta - \theta_{\text{preferred}}\right)\right), \tag{2}$$

where a, b_1, b_2, c, and θ_{prefered} are model parameters, and θ is the stimulus' drift direction. In the orientation-tuned model, the average neural response of each neuron was fitted using a single Von Mises function given by

$$f_2(\theta) = a + b\exp\left(c\cos\left(2\left(\theta - \theta_{\text{preferred}}\right)\right)\right), \tag{3}$$

with parameters a, b, c, and $\theta_{\text{preferred}}$. The third and last model is a null model that assumes neurons are not significantly tuned to drift direction, and fits a constant value to neural responses, that is $f_3(\theta) = a$. We fitted all three models to the response of neuron across all trials by minimizing the sum of squared residuals between observed neural response and the tuning function across different stimulus drift direction (see Supplementary Fig. 1 for the R^2's associated with these fits). We then compared the nested models by an F-test (with Bonferroni correction for multiple comparisons) to test whether neurons are direction-tuned, orientation-tuned or untuned.

Generalized Fisher information. Linear Fisher information[17,77,78], which is the Fisher information that can be recovered by a linear decoder, can for stimulus θ_0 be computed by $I(\theta_0) = \mathbf{f}'(\theta_0)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\theta_0)\mathbf{f}'(\theta_0)$. Here, $\mathbf{f}'(\theta_0)$ is the vector of derivatives of each neuron's average response with respect to θ, with the ith element given by $\partial f_i(\theta_0)/\partial\theta = \partial <r_i|\theta_0>/\partial\theta$, and $\boldsymbol{\Sigma}(\theta_0) = \mathrm{cov}(\mathbf{r}|\theta_0)$ is the noise covariance of the population activity vector $\mathbf{r}$. Therefore, linear Fisher information is fully determined by the first two moments of the population activity, irrespective of the presence of higher-order moments. Furthermore, if $\widehat{\theta} = \mathbf{w}^{\mathrm{T}}(\mathbf{r} - \mathbf{f}(\theta_0)) + \theta_0$ is the unbiased minimum-variance locally linear estimate of θ, its variance is given by $\mathrm{var}\left(\widehat{\theta}|\theta_0\right) = 1/I(\theta_0)$[79]. In practice, $\mathbf{f}'(\theta_0)$ and $\boldsymbol{\Sigma}(\theta_0)$ are approximated by their empirical estimates, $\mathbf{f}'(\theta_0) \approx \left(\widehat{\mathbf{f}}(\theta_2) - \widehat{\mathbf{f}}(\theta_1)\right)/\delta\theta$, and $\boldsymbol{\Sigma}(\theta_0) \approx (\mathrm{cov}(\mathbf{r}|\theta_1) + \mathrm{cov}(\mathbf{r}|\theta_2))$, where $\theta_{1,2} = \theta_0 \mp \delta\theta/2$. This naïve estimate is biased but a bias-corrected estimate can be used[30].

By definition, Fisher information is a measure of fine discrimination performance around a specific reference θ_0, requiring small $\delta\theta$. As we show in the SI, the same measure with $\mathbf{f}'(\theta_0)$ and $\boldsymbol{\Sigma}(\theta_0)$ replaced by their empirical estimate can be used for coarse discrimination for which $\delta\theta$ is larger. Furthermore, this generalization corresponds to $(d'/\delta\theta)^2$, where d' is the sensitivity index used in signal detection theory[22], becomes equivalent to Fisher information in the $\delta\theta \to 0$ limit, and shares many properties with the original Fisher information estimate. In particular, the same bias correction leads to unbiased estimates. Kanitscheider et al.[30] lack an estimate of the variance of the bias-corrected Fisher information estimate that can be computed from data, so we provide a derivation thereof in the SI.

To relate (generalized) Fisher information to discrimination thresholds, we observe that the variance of the stimulus estimate $\widehat{\theta}$ is $1/I(\theta_0)$. Assuming this estimate to be Gaussian across trials, the difference in estimates across two stimuli which differ by $\Delta\theta$ is distributed as $\mathrm{N}(\Delta\theta, 2/I(\theta_0))$. Therefore, the probability of correctly discriminating these stimuli is $\Phi\left(\Delta\theta\sqrt{I(\theta_0)/2}\right)$[3,80,81], where $\Phi(\cdot)$ is the cumulative function of a standard Gaussian. Setting the desired probability correct to 80% and solving for $\Delta\theta$ results in the drift direction discrimination threshold $\Delta\theta = \Phi^{-1}(0.8)\sqrt{2/I(\theta_0)}$.

Estimating Fisher information from neural data. Our Fisher information estimates have two sources of uncertainty. First, they rely on empirical estimates of $\mathbf{f}'(\theta_0)$ and $\boldsymbol{\Sigma}(\theta_0)$ from a limited number of trials that are thus noisy. Second, we assume that recorded neurons to be a small, random subsample of the full population. As we want to estimate the average Fisher information across such subsamples across different population sizes, observing only a single subsample introduces additional uncertainty.

We will first focus on the uncertainty due to a limited number of trials. We can find an unbiased estimate of I_N for a population of N neurons by a biased-corrected estimate $\widehat{I}_N$. Our aim is to fit models to how $\widehat{I}_N$ changes with N. We can estimate this change by computing $\widehat{I}_1$ for a single neuron, and then successively add neurons to the population to find $\widehat{I}_2, \widehat{I}_3, \ldots$ However, this procedure causes $\widehat{I}_N$ and $\widehat{I}_{N+1}$ to be correlated, as their estimates share the data of the previous N neurons. Therefore, although previous work did not correct for these correlations when fitting the information scaling curves[31–33], it is important to account for them when fitting the information estimates across multiple N. Fortunately, the change in information across successive N, $\Delta\widehat{I}_N = \widehat{I}_N - \widehat{I}_{N-1}$ is uncorrelated, that is $\mathrm{cov}\left(\Delta\widehat{I}_N, \Delta\widehat{I}_{N+1}\right) = 0$ (see SI). The intuition underlying this independence is that the response of each neuron can be decomposed into a component that is collinear to the remaining population and one that is independent of it. Only the independent component contributes additional information, making the information increase due to adding this neuron independent of the information encoded in the remaining population. Overall, rather than fitting the information estimates, we will instead fit the information increases across different N.

To handle the uncertainty associated with subsampling larger populations, we assumed that the small recorded population is statistically representative of the full population. Then, our aim is to simulate random draws of the size of the recorded population from the full, much larger population. We achieved this simulation by randomly drawing neurons from the recorded population, without replacement, up to the full recorded population size, effectively resulting in a random order of adding recorded neurons to the population. For each such ordering, we estimated the information increase with each additional neuron. As the information in the total recorded population is the same, irrespective of this ordering, the information increases ΔI_N and ΔI_M for $N \neq M$ will on average be negatively correlated across different orderings. This is an artifact of re-using the same data to simulate samples from a larger population. As long as the full population is significantly larger than the one we recorded from, the probability of re-sampling the same pair of neurons from the full population is exceedingly small, such that we can ignore these correlations (see SI). Any negative correlations between information increases, however small, will reduce the variance of our Fisher information estimates. Therefore, by ignoring these correlations, we will estimate an upper bound of this variance, and thus overestimate the uncertainty. In summary, we estimated the uncertainty associated with subsampling larger populations by estimating the moments of the Fisher information increase by bootstrap estimates across different orderings with which neurons are added to the population. As shown in Supplementary Fig. 16a, this procedure also captures the uncertainty associated with a limited number of trials, such that no extra correction is needed to account for this second source of uncertainty.

Overall, we estimated the moments of the Fisher information increase $\Delta\widehat{I}_N$ for the discrimination of θ_1 and θ_2 as follows. First, we estimated the empirical moments $\widehat{\mathbf{f}}'$ and $\widehat{\Sigma}$ using the same number of trials for θ_1 and θ_2. Second, we chose a particular random order with which to add neurons to the population. Third, we used this order to estimate $\Delta\widehat{I}_1, \Delta\widehat{I}_2, \ldots$ by use of the biased-corrected Fisher information estimate applied to $\widehat{\mathbf{f}}'$ and $\widehat{\boldsymbol{\Sigma}}$. Fourth, we repeated this estimate across 10^4 different neural ordering to get 10^4 bootstrap estimates of the Fisher information increase sequence. Fifth, we used the bootstrap estimate to compute the moments $\mu_N = <\Delta\widehat{I}_N>$ and $\sigma_N^2 = \mathrm{var}\left(\Delta\widehat{I}_N\right)$ for each N, which we in turn use to fit the information scaling curves (see below). As the individual increases are independent across N, we used its moments to additionally estimate the moments of $\widehat{I}_N = \sum_{n=1}^N \Delta\widehat{I}_n$, which are given by $<\widehat{I}_N> = \sum_{n=1}^N \mu_n$ and $\mathrm{var}\left(\widehat{I}_N\right) = \sum_{n=1}^N \sigma_n^2$. We used these moments to plot the Fisher information estimates in Figs. 3a, 4b/d and 5a.

Fisher information scaling with limited information. Moreno-Bote et al.[17] have shown that for large populations encoding limited asymptotic information I_∞, the noise covariance can be decomposed into $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_0 + I_\infty^{-1}\mathbf{f}'\mathbf{f}'^{\mathrm{T}}$, where only the $\mathbf{f}'\mathbf{f}'^{\mathrm{T}}$ component, called *differential correlations*, limits information. Assuming a population size of N neurons, we can apply the Sherman–Morrison formula to the above noise covariance decomposition[17,50] to find $I_N^{-1} = I_{0,N}^{-1} + I_\infty^{-1}$, where $I_N = \mathbf{f}'^{\mathrm{T}}\boldsymbol{\Sigma}_N^{-1}\mathbf{f}'$ is the Fisher information in this population, and $I_{0,N} = \mathbf{f}'^{\mathrm{T}}\boldsymbol{\Sigma}_0^{-1}\mathbf{f}'$ is the Fisher information associated with the non-limiting noise covariance component $\boldsymbol{\Sigma}_0$. Furthermore, assuming that this non-limiting component contributes average information c per neuron, that is $I_{0,N} = cN$, results in Eq. (1) in the main text. While similar expressions have been suggested before[10,11], they were derived from models that made significantly more restrictive assumptions about neural tuning and shared variability. We also tested a model in which $I_{0,N}$ initially scaled supralinearly in N. We found this model by integrating $c(1 - e^{-N/\tau})$ from zero to N, resulting in $I_{0,N} = c(N + \tau(e^{-N/\tau} - 1))$ with parameter τ that controls the extent of the initial supralinearity. The two models become equivalent with $\tau \to 0$. The above derivation relies on the traditional Fisher information definition for fine discrimination. The results remain unchanged when moving to Fisher information generalized to coarse discrimination.

Fitting information scaling models. We compared three models for how Fisher information I_N scales with population size N. The first *unlim* model assumes linear scaling, $I_N = cN$, and has one parameter, $\phi_1 = \{c\}$. The second *lim* model, given by Eq. (1) in the main text, assumes asymptotic information I_∞, and that the Fisher information associated with the non-limiting covariance component increased linearly, $I_{0,N} = cN$. This model thus has two parameters, $\phi_2 = \{c, I_\infty\}$. The third *lim-exp* model assumes an initial supralinear scaling of $I_{0,N}$, as described above, and has three parameters, $\phi_3 = \{c, I_\infty, \tau\}$. The lim-exp model fits the data consistently worse than the *lim* model (Supplementary Fig. 6b), such we did not consider it in the main text.

As the Fisher information estimates in data are correlated across different population sizes, we did not directly fit these estimates. Instead, we fitted how they changed when adding additional neurons, as the estimated Fisher information increase is uncorrelated across different population sizes. That is, we used the likelihood function $p(X|\phi) = \prod_{n=1}^N \mathrm{N}\left(\mu_n(X)\middle|\Delta I_{n,\phi}, \sigma_{\mathrm{n}}^2(X)\right)$, where X is the recorded data (that is, the recorded population activity in all trials with the drift directions that are being discriminated, yielding the desired moments $\mu_1,\ldots,\mu_N$ and $\sigma_1^2, \ldots, \sigma_{\mathrm{N}}^2$), ϕ are the model parameters, $\Delta I_{n,\phi} = I_{n,\phi} - I_{n-1,\phi}$ is the information increase predicted by that model, and μ_n and σ_n^2 are the mean and variance of the estimated information increase in data X for a particular discrimination when moving from population size $n-1$ to n (see further above).

We regularized the fits by weakly informative parameter priors. For c we used $p(c) \propto \mathrm{St}_1(<\mu_n>, 100(<\mu_n>+0.5)^2)$, which is a Student's t distribution with mean $<\mu_n>$, variance $100(<\mu_n>+0.5)^2$ and one degree of freedom, and where $<\mu_n>$ is the average estimated information increase in the recorded population. Thus, the prior is centered on the empirical estimate for c for the linear scaling model, but has a wide variance around this estimate. We furthermore limited c to the range $c \in [0,\infty]$. For I_∞ we used $p(I_\infty) \propto \mathrm{St}_1\left(<\widehat{I}_N>, 100 \max\left\{1, <\widehat{I}_N>\right\}^2\right)$ over $I_\infty \in [0,\infty]$, which is a weak prior centered on the empirical information estimate $<\widehat{I}_N> = \sum_{n=1}^{N} \mu_n$ for the recorded population. For τ we used $p(\tau) \propto \mathrm{St}_1(0, N^2)$ over $\tau \in [0,\infty]$. Technically, the data should not inform the priors, as it does here. However, this is not a concern for the extremely weak and uninformative priors used here.

We fitted the different models to data X of individual sessions/mice and discriminations by sampling the associated parameter posteriors, $p(\phi|X) \propto p(X|\phi)p(\phi)$, by slice sampling[82]. The slice sampling interval widths were set to $(<\mu_n>+0.5)/2$ for c, to $\max\left\{1, <\widehat{I}_N>\right\}/5$ for I_∞, and to 10 for τ. The samplers were initiated by parameter values found by maximum-likelihood fits for the respective model. For each fit, we sampled four chains with 10^5 posterior samples each, after discarding 100 burn-in samples, and keeping only each 10th sample. We used the Gelman-Rubin potential scale reduction factor[83] to assess MCMC convergence. To fit the same model to multiple discriminations simultaneously (i.e., our *pooled* fits), we sampled from the pooled posterior $p(\phi|X_{1:K}) \propto p(\phi)\prod_{k=1}^{K} p(X_k|\phi)$, where X_k is the data associated with the kth discrimination.

We compared the fit quality of different models by the Watanabe-Akaike information criterion (WAIC; see ref. [84]). This criterion supports comparing models with different numbers of parameters, as it takes the associated change in model complexity into account. It is preferable to the Akaike information criterion or Bayesian information criterion, as it provides a better approximation to the cross-validated predictive density than other methods[85].

We found posterior predictive densities by empirically marginalizing over the posterior parameter samples, $\phi^{(1)},\ldots,\phi^{(J)}$, pooled across all four chains. That is, we approximated the density of any function $f(\phi)$ of these parameters by $p(f|X) \approx J^{-1}\sum_{j=1}^{J} \delta\left(f - f\left(\phi^{(j)}\right)\right)$, where $\delta(\cdot)$ is the Dirac delta function. This approach was used to find the predictive density of the fitted information increase in Fig. 4a (top), as well as the information in Fig. 4a (bottom) and Fig. 4c. We also used it to estimate the posterior distribution of the required population size N_{95} to capture 95% of the asymptotic information.

Additional data analysis and statistical tests. Except for Figs. 6 and 7, all statistical tests across sessions/mice were restricted to mice 1–4.

Figure 3. We removed noise correlations in the recorded data by, for each neuron, randomly permuting the trial order across all trials in which the same drift direction was presented. We then compared the total information in the recorded population with (I_N^{Shuffled}) and without (I_N) trial-shuffling by a bootstrap test (Fig. 3d). To do so, we estimated mean and variance of that total recorded information as described above, and then computed the probability of the null hypotheses ($I_N^{\mathrm{Shuffled}} \leq I_N$) by $p = \mathrm{pr}\left(I_N^{\mathrm{Shuffled}} - I_N < 0\right)$, where we assumed Gaussian information estimates. We compared I_N^{Shuffled} to I_N across sessions/mice by a paired t-test across all non-overlapping discriminations with $\delta\theta = 45^\circ$ (Fig. 3d). We focused exclusively on discriminations that did not share any drift directions, to avoid comparing estimates that rely on the same underlying set of trials. Unless otherwise noted, all non-overlapping discriminations with $\delta\theta = 45^\circ$ were performed on the 0° vs. 45°, 90° vs. 135°, 180° vs. 225°, and 270° vs. 315° discriminations. To test for significant differences in the drift direction discrimination thresholds (Fig. 3f) across multiple discriminations with the same difference in drift directions, θ, we relied on the one-to-one mapping between information and discrimination threshold, and performed the test directly on the estimated information. For K discriminations (in our case $K = 4$ for non-overlapping discriminations), let $I_{N,k}$, $k = 1,\ldots, K$ denote the information in the recorded population for discrimination k, $I_{N,k} \sim \mathrm{N}\left(\mu_{N,k}, \sigma_{N,K}^2\right)$. To test the null hypothesis that all $I_{N,k}$ share the same mean, we drew 10^5 bootstrap samples each from $TS_{H_1} = \sum_{k=1}^{K}\left(I_{N,k} - \mu_{N,k}\right)^2$ and $TS_{H_0} = \sum_{k=1}^{K}\left(I_{N,k} - \mu_N\right)^2$ with $\mu_N = K^{-1}\sum_{k=1}^{K}\mu_{N,k}$, and then computed the probability that TS_{H_0} is larger than TS_{H1} by $p = \mathrm{pr}(TS_{H_1} - TS_{H_0} < 0)$.

Figure 4. To test how $1/I_N$ scales with $1/N$ (Fig. 4b), we found the moments of $1/I_N$ by $<1/I_N> \approx 1/<I_N>$ and $\mathrm{var}(1/I_N) \approx \mathrm{var}(I_N)/I_N^4$. To fit $<1/I_N>$ over $1/N$, we performed weighted linear regression with weights $1/\mathrm{var}(1/I_N)$ for each N. The pooling across different discriminations in Fig. 4d was performed over 45° vs. 90°, 135° vs. 180°, 225° vs. 270°, and 0° vs. 315° for pooled 1, and 0° vs. 45°, 90° vs. 135°, 180° vs. 225°, and 270° vs. 315° for pooled 2. All other pooled estimates (Figs. 4e, 6d and e, and 7b) were pooled across 45° vs. 90°, 135° vs. 180°, 225° vs. 270°, and 0° vs. 315° for $\delta\theta = 45^\circ$, across 45° vs. 135°, 90° vs. 180°, 225° vs. 315°, and 0° vs. 270° for $\delta\theta = 90^\circ$, and across 45° vs. 180°, 90° vs. 315°, and 0° vs. 225° for $\delta\theta = 135^\circ$. Note that the estimate I_N's are correlated across different N's, and we did not correct for these correlations. Such a correction might lower the reported R^2 values. Therefore, the Bayesian model comparison across different information scaling models, as reported in the main text, provides a statistically sounder confirmation of limited asymptotic information.

Figure 5. The shaded error regions in Fig. 5a relied on parametric bootstrap estimates. For information scaling for a fixed ordering, we computed the estimate and variance of I_1, I_2,… by the Fisher information and the variance of this estimator (see SI), and used these estimates to compute mean and variance of the information increase associated with adding individual neurons to the population. We then re-sampled these information increases from Gaussian distributions with the found moments, and summed the individual samples to find different samples for the whole information scaling curve. These samples were in turn used to estimate mean and variance of the information scaling for a fixed order with which neurons were added to the population. This procedure was chosen, as the increase in Fisher information is independent across added neurons, whereas the total Fisher information is not. A similar procedure was used to find the estimates for random orderings, for which we additionally shuffled the order of neurons across different samples of the information scaling curve. The above procedures yielded 10^3 bootstrap samples for each information scaling curve, which we in turn used to find samples for the population sizes required to capture 90% of the total information (Fig. 5a, b). In neither case did we apply bias correction of the Fisher information estimate. This bias correction would have been stronger for larger population sizes, which would have led to a seeming (but not real) drop of information with population size, resulting from a lower number of trials per neuron in the population, and an associated stronger bias correction.

Figure 6. To identify for individual discriminations if increasing the stimulus contrast increased information in the recorded population (Fig. 6a, b), we estimated information in the recorded population by the bias-corrected Fisher information estimate[30], and its variance by our analytical expression for this estimate's variance (see SI). We assumed the estimate for low and high contrast, I_N^{LO} and I_N^{HI}, to be Gaussian, and found the probability of no information increase by $\mathrm{pr}\left(\mathrm{I_N^{HI}} \leq \mathrm{I_N^{LO}}\right)$, using the aforementioned moments. The paired t-test across sessions/mice (Fig. 6b) did not take into account the information estimates' variance. For Fig. 6e, higher contrast was considered to significantly increase the information in the recorded population (filled dots in Fig. 6e), if it did so for at least five out of eight possible discriminations with $\delta\theta = 45^\circ$.

Figure 7. To test the relationship between c and I_∞ in Fig. 7d, we performed the linear regression $log_{10}(c) = \beta_0 + \beta_1 log_{10}(I_\infty)$. The relationship between N_{95} and I_∞ was found by substituting $c = 10^{\beta_0} I_\infty^{\beta_1}$ into the expression for N_{95}, resulting in $N_{95} = 0.95 I_\infty^{1-\beta_1}/\left(0.05 \times 10^{\beta_0}\right)$. To find the information loss for using a smaller population size than required, we assumed $I_\infty^{\mathrm{hi}} = \alpha I_\infty^{\mathrm{lo}}$ and computed the fraction $I_N^{\mathrm{hi}}/I_\infty^{\mathrm{hi}}$ at $N = N_{95}^{\mathrm{lo}}$, which is the population size that captures 95% of I_∞^{lo}. Substituting the found relationships between I_∞, c, and N_{95} results in this fraction to be given by $0.95/(0.95 + 0.05\alpha^{1-\beta_1})$, which, for $\alpha = 3$, equals 0.93. Interestingly, this fraction depends only the relationship between I_∞^{lo} and I_∞^{hi}, as quantified by α, but not on their individual values.

Figure 8. All estimates in Fig. 8 are averages across 10 random splits of the recorded data. For each split, half of the trials were used to compute the principal dimensions, $\mathbf{Q}_{\mathrm{train}}$, using the spectral decomposition $\mathbf{\Sigma}_{\mathrm{train}} = \mathbf{Q}_{\mathrm{train}}\mathbf{D}_{\mathrm{train}}\mathbf{Q}_{\mathrm{train}}^{\mathrm{T}}$, where $\mathbf{D}_{\mathrm{train}}$ is diagonal, $\mathbf{Q}_{\mathrm{train}}$ is the matrix of unit eigenvectors, and we denote the nth column vector of $\mathbf{Q}_{\mathrm{train}}$ by $\mathbf{q}_{n,\mathrm{train}}$. The second half of trials was used to find $\mathbf{f}'_{\mathrm{test}}$ and $\mathbf{\Sigma}_{\mathrm{test}}$, from which we computed the shown estimates as follows. The noise variance associated with the nth principal dimension was found by $\mathbf{q}_{n,\mathrm{train}}^{\mathrm{T}}\mathbf{\Sigma}_{\mathrm{test}}\mathbf{q}_{n,\mathrm{train}}$. The $\mathbf{f}'$ alignment to the nth principal dimension was found by $\cos^2(\alpha_n) = \left(\mathbf{q}_{n,\mathrm{train}}^{\mathrm{T}}\mathbf{f}'_{\mathrm{test}}\right)^2/\mathbf{f}'^{\,\mathrm{T}}_{\mathrm{test}}\mathbf{f}'_{\mathrm{test}}$. The information encoded in the first n principal dimensions was found by $I_n = \mathbf{f}'^{\,\mathrm{T}}_{\mathrm{test}}\mathbf{Q}_{1:n,\mathrm{train}}\left(\mathbf{Q}_{1:n,\mathrm{train}}^{\mathrm{T}}\mathbf{\Sigma}_{\mathrm{test}}\mathbf{Q}_{1:n,\mathrm{train}}\right)^{-1}\mathbf{Q}_{1:n,\mathrm{train}}^{\mathrm{T}}\mathbf{f}'_{\mathrm{test}}$, where $\mathbf{Q}_{1:n,\mathrm{train}}$ is the matrix formed by the first n columns of $\mathbf{Q}_{\mathrm{train}}$.

Additional analyses in discussion. To compare the estimated population sizes to the number of neurons in V1, we asked for the number of neurons required to encode 95% of the asymptotic information associated with a direction discrimination threshold of 1°. This threshold most likely exceeds the behavioral performance that mice can reach even for high contrast stimuli[23,25] and thus provides an upper bound on the required population size. Achieving such a low threshold requires an asymptotic information of 4651 rad^{-2} (Fig. 3e), and approximately 48,000 neurons are necessary to encode 95% of this information (Fig. 7d). Current estimates of the neural density of mouse V1 range from 92,400 to 214,000 neurons per mm^3 (refs. [43,44]). For area V1 with an approximate size of 3.063 mm^3 (ref. [43]), this amounts to 283,000 to 655,500 neurons[44]. Therefore, our estimated population sizes are well within those available in V1 of mice. In addition to comparing our estimates to the total number of neurons in V1, we also considered best and worst-case scenarios for the number of neurons in V1 that correspond to the retinotopic area of the visual stimulus (103° azimuth, 71° elevation). To convert between degrees of visual space and mm of cortical space, we used the conversion factors 63°/mm in azimuth and 40°/mm in elevation[86]. In the best-case scenario, the entire visual stimulus corresponds to ~1.65 × 1.78 mm, or 2.95 mm^2 in the cortex. Relative to the total area of V1, estimated as ~3.25–4 mm^2 (refs. [87,88]), 75–90% of V1 neurons would be activated by the stimulus. Using the range above for total neurons in V1, this is on the order of ~10× our estimates for the number of neurons encoding 95% of asymptotic information. For a conservative worst-case scenario, we consider only the full-contrast portion of the stimulus (circle with radius 20°), for which the retinotopic area covered is ~0.5 mm^2, or ~12.5–15% of V1 neurons. This conservative estimate of a lower bound on the number of responsive neurons is ~1× our required population size estimates. Thus, mouse V1 has more neurons than required to encode most of the estimated asymptotic information about the direction of a moving visual stimulus.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The datasets generated and analyzed during this study are available in the Figshare repository, https://doi.org/10.6084/m9.figshare.13274951. Source data are provided with this paper.

Code availability

MATLAB code performing the described analyzes and generating the resulting figures is available at https://doi.org/10.5281/zenodo.4291863.

Received: 10 February 2020; Accepted: 16 December 2020;
Published online: 20 January 2021

References

1. Kohn, A., Coen-cagli, R., Kanitscheider, I. & Pouget, A. Correlations and neuronal population information. *Annu. Rev. Neurosci.* **39**, 237–256 (2016).
2. Averbeck, B. B., Latham, P. E. & Pouget, A. Neural correlations, population coding and computation. *Nat. Rev. Neurosci.* **7**, 358–366 (2006).
3. Nogueira, R. et al. The effects of population tuning and trial-by-trial variability on information encoding and behavior. *J. Neurosci.* **40**, 1066–1083 (2020).
4. Shamir, M. Emerging principles of population coding: in search for the neural code. *Curr. Opin. Neurobiol.* **25**, 140–148 (2014).
5. Carandini, M. Amplification of trial-to-trial response variability by neurons in visual cortex. *PLoS Biol.* **2**, e264 (2004).
6. Faisal, A. A., Selen, L. P. J. & Wolpert, D. M. Noise in the nervous system. *Nat. Rev. Neurosci.* **9**, 292–303 (2008).
7. Shadlen, M. N. & Newsome, W. T. The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. *J. Neurosci.* **18**, 3870–3896 (1998).
8. Softky, W. & Koch, C. The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. *J. Neurosci.* **13**, 334–350 (1993).
9. Tolhurst, D. J., Movshon, J. A. & Dean, A. F. The statistical reliability of signals in single neurons in cat and monkey visual cortex. *Vis. Res.* **23**, 775–785 (1983).
10. Zohary, E., Shadlen, M. N. & Newsome, W. T. Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature* **370**, 140–143 (1994).
11. Abbott, L. F. & Dayan, P. The effect of correlated variability on the accuracy of a population code. *Neural Comput.* **11**, 91–101 (1999).
12. Adibi, M., McDonald, J. S., Clifford, C. W. G. & Arabzadeh, E. Adaptation improves neural coding efficiency despite increasing correlations in variability. *J. Neurosci.* **33**, 2108–2120 (2013).
13. Gu, Y. et al. Perceptual learning reduces interneuronal correlations in macaque visual cortex. *Neuron* **71**, 750–761 (2011).
14. Maynard, E. M. et al. Neuronal interactions improve cortical population coding of movement direction. *J. Neurosci.* **19**, 8083–8093 (1999).
15. Averbeck, B. B. & Lee, D. Neural noise and movement-related codes in the macaque supplementary motor area. *J. Neurosci.* **23**, 7630–7641 (2003).
16. Ecker, A. S., Berens, P., Tolias, A. S. & Bethge, M. The effect of noise correlations in populations of diversely tuned neurons. *J. Neurosci.* **31**, 14272–14283 (2011).
17. Moreno-Bote, R. et al. Information-limiting correlations. *Nat. Neurosci.* https://doi.org/10.1038/nn.3807 (2014).
18. Cohen, M. R. & Kohn, A. Measuring and interpreting neuronal correlations. *Nat. Neurosci.* **14**, 811–819 (2011).
19. Kanitscheider, I., Coen-Cagli, R., & Pouget, A. Origin of information-limiting noise correlations. *Proc. Natl Acad. Sci. USA* **112**, E6973-82 (2015).
20. Leavitt, M. L., Pieper, F., Sachs, A. J., & Martinez-Trujillo, J. C. Correlated variability modifies working memory fidelity in primate prefrontal neuronal ensembles. *Proc. Natl Acad. Sci. USA* **114**, E2494–E2503 (2017).
21. Pruszynski, J. A. & Zylberberg, J. The language of the brain: real-world neural population codes. *Curr. Opin. Neurobiol.* **58**, 30–36 (2019).
22. Green, D. M. & Swets, J. A. *Signal Detection Theory and Psychophysics* (Wiley, New York, 1966).
23. Glickfeld, L. L., Histed, M. H. & Maunsell, J. H. R. Mouse primary visual cortex is used to detect both orientation and contrast changes. *J. Neurosci.* **33**, 19416–19422 (2013).
24. Andermann, M. L. Chronic cellular imaging of mouse visual cortex during operant behavior and passive viewing. *Front. Cell. Neurosci.* **4**, 1–16 (2010).
25. Abdolrahmani, M., Lyamzin, D. R., Aoki, R. & Benucci, A. Cognitive modulation of interacting corollary discharges in the visual cortex. Preprint at https://www.biorxiv.org/content/10.1101/615229v1 (2019).
26. Ni, A. M., Ruff, D. A., Alberts, J. J., Symmonds, J. & Cohen, M. R. Learning and attention reveal a general relationship between population activity and behavior. *Science* **359**, 463–465 (2018).
27. Otazu, G. H., Tai, L.-H., Yang, Y. & Zador, A. M. Engaging in an auditory task suppresses responses in auditory cortex. *Nat. Neurosci.* **12**, 646–654 (2009).
28. McGinley, M. J. et al. Waking state: rapid variations modulate neural and behavioral responses. *Neuron* **87**, 1143–1161 (2015).
29. Dadarlat, M. C. & Stryker, M. P. Locomotion enhances neural encoding of visual stimuli in mouse V1. *J. Neurosci.* **37**, 3764–3775 (2017).
30. Kanitscheider, I., Coen-Cagli, R., Kohn, A. & Pouget, A. Measuring Fisher information accurately in correlated neural populations. *PLoS Comput. Biol.* **11**, 1–27 (2015).
31. Bartolo, R., Saunders, R. C., Mitz, A. R. & Averbeck, B. B. Information limiting correlations in large neural populations. *J. Neurosci.* https://doi.org/10.1523/JNEUROSCI.2072-19.2019 (2020).
32. Cotton, R. J. et al. Accuracy of sensory in*formation does not saturate for large neuronal populations*. 2018 Neuroscience Meeting Planner, 219.02/BB10 (Society for Neuroscience: San Diego, CA, 2018).
33. Mendels, O. P. & Shamir, M. Relating the structure of noise correlations in Macaque primary visual cortex to decoder performance. *Front. Comput. Neurosci.* https://doi.org/10.3389/fncom.2018.00012 (2018).
34. Ince, R. A. A., Panzeri, S. & Kayser, C. Neural codes formed by small and temporally precise populations in auditory cortex. *J. Neurosci.* **33**, 18277–18287 (2013).
35. Busse, L. et al. The detection of visual contrast in the behaving mouse. *J. Neurosci.* **31**, 11351–11361 (2011).
36. Engel, T. A. & Steinmetz, N. A. New perspectives on dimensionality and variability from large-scale cortical dynamics. *Curr. Opin. Neurobiol.* **58**, 181–190 (2019).
37. Semedo, J. D., Zandvakili, A., Machens, C. K., Yu, B. M. & Kohn, A. Cortical areas interact through a communication subspace. *Neuron* **102**, 1–11 (2019).
38. Williamson, R. C. et al. Scaling properties of dimensionality reduction for neural populations and network models. *PLoS Comput. Biol.* **12**, e1005141 (2016).
39. Denman, D. J. & Reid, R. C. Synergistic population encoding and precise coordinated variability across interlaminar ensembles in the early visual system. Preprint at https://www.biorxiv.org/content/10.1101/812859v1 (2019).
40. Stringer, C., Michaelos, M. & Pachitariu, M. High precision coding in mouse visual cortex. Preprint at https://www.biorxiv.org/content/10.1101/679324v1 (2019).
41. Chen, T.-W. et al. Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499**, 295–300 (2013).
42. Ledochowitsch, P. et al On the correspondence of electrical and optical physiology in in vivo population-scale two-photon calcium imaging. Preprint at https://www.biorxiv.org/content/10.1101/800102v1 (2019).
43. Herculano-Houzel, S., Watson, C. & Paxinos, G. Distribution of neurons in functional areas of the mouse cerebral cortex reveals quantitatively different cortical zones. *Front. Neuroanat.* **7**, 1–14 (2013).
44. Keller, D., Erö, C., & Markram, H. Cell densities in the mouse brain: a systematic review. *Front. Neuroanat.* https://doi.org/10.3389/fnana.2018.00083 (2018).
45. de Vries, S. E. J. et al. A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. *Nat. Neurosci.* **23**, 138–151 (2020).
46. Vinje, W. E. & Gallant, J. L. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* **287**, 1273–1276 (2000).
47. Yoshida, T. & Ohki, K. Natural images are reliably represented by sparse and variable populations of neurons in visual cortex. *Nat. Commun.* **11**, 872 (2020).
48. Britten, K. H., Newsome, W. T., Shadlen, M. N., Celebrini, S. & Movshon, J. A. A relationship between behavioral choice and the visual responses of neurons in macaque MT. *Vis. Neurosci.* **13**, 87–100 (1996).
49. Haefner, R. M., Gerwinn, S., Macke, J. H. & Bethge, M. Inferring decoding strategies from choice probabilities in the presence of correlated variability. *Nat. Neurosci.* **16**, 235–242 (2013).
50. Pitkow, X., Liu, S., Angelaki, D. E., DeAngelis, G. C. & Pouget, A. How can single sensory neurons predict behavior? *Neuron* **87**, 411–424 (2015).

51. Britten, K. H., Shadlen, M. N., Newsome, W. T. & Movshon, J. A. The analysis of visual motion: a comparison of neuronal and psychophysical performance. *J. Neurosci.* **12**, 4745–4765 (1992).
52. Nienborg, H. & Cumming, B. G. Macaque V2 neurons, but not V1 neurons, show choice-related activity. *J. Neurosci.* **26**, 9567–9578 (2006).
53. Jasper, A. I., Tanabe, S. & Kohn, A. Predicting perceptual decisions using visual cortical population responses and choice history. *J. Neurosci.* **39**, 6714–6727 (2019).
54. Keller, G. B., Bonhoeffer, T. & Hübener, M. Sensorimotor mismatch signals in primary visual cortex of the behaving mouse. *Neuron* **74**, 809–815 (2012).
55. Ayaz, A., Saleem, A. B., Schölvinck, M. L. & Carandini, M. Locomotion controls spatial integration in mouse visual cortex. *Curr. Biol.* **23**, 890–894 (2013).
56. Niell, C. M. & Stryker, M. P. Highly selective receptive fields in mouse visual cortex. *J. Neurosci.* **28**, 7520–7536 (2008).
57. Lee, S., Meyer, J. F., Park, J. & Smirnakis, S. M. Visually driven neuropil activity and information encoding in mouse primary visual cortex. *Front. Neural Circuits* **11**, 1–18 (2017).
58. Ringach, D. L. et al. Spatial clustering of tuning in mouse primary visual cortex. *Nat. Commun.* **7**, 12270 (2016).
59. Dow, B. M. Orientation and color columns in monkey visual cortex. *Cereb. Cortex* **12**, 1005–1015 (2002).
60. Mott, M. C., Gordon, J. A. & Koroshetz, W. J. The NIH BRAIN Initiative: advancing neurotechnologies, integrating disciplines. *PLoS Biol.* **16**, e3000066 (2018).
61. Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E. & Pouget, A. Not noisy, just wrong: the role of suboptimal inference in behavioral variability. *Neuron* **74**, 30–39 (2012).
62. Drugowitsch, J., Wyart, V., Devauchelle, A.-D. & Koechlin, E. Computational precision of mental inference as critical source of human choice suboptimality. *Neuron* **92**, 1–14 (2016).
63. Acerbi, L., Vijayakumar, S. & Wolpert, D. M. On the origins of suboptimality in human probabilistic inference. *PLoS Comput. Biol.* **10**, e1003661 (2014).
64. Doya, K., Ishii, S., Pouget, A., & Rao, R. P. N. *Bayesian Brain: Probabilistic Approaches to Neural Coding* (MIT Press, 2006).
65. Moreno-Bote, R., Knill, D. C. & Pouget, A. Bayesian sampling in visual perception. *Proc. Natl Acad. Sci. USA* **108**, 12491–12496 (2011).
66. Pouget, A., Beck, J. M., Ma, W. J. & Latham, P. E. Probabilistic brains: knowns and unknowns. *Nat. Neurosci.* **16**, 1170–1178 (2013).
67. Fiser, J., Berkes, P., Orbán, G. & Lengyel, M. Statistically optimal perception and learning: from behavior to neural representations. *Trends Cogn. Sci.* **14**, 119–130 (2010).
68. Gao, P. & Ganguli, S. On simplicity and complexity in the brave new world of large-scale neuroscience. *Curr. Opin. Neurobiol.* **32**, 148–155 (2015).
69. Gao, P. et al. A theory of multineuronal dimensionality, dynamics and measurement. Preprint at https://www.biorxiv.org/content/10.1101/214262v2 (2017).
70. Kobak, D. et al. Demixed principal component analysis of neural population data. *ELife* **5**, 1–36 (2016).
71. Chettih, S. N. & Harvey, C. D. Single-neuron perturbations reveal feature-specific competition in V1. *Nature* **567**, 334–340 (2019).
72. Peirce, J. W. PsychoPy—Psychophysics software in Python. *J. Neurosci. Methods* **162**, 8–13 (2007).
73. Harvey, C. D., Coen, P. & Tank, D. W. Choice-specific sequences in parietal cortex during a virtual-navigation decision task. *Nature* **484**, 62–68 (2012).
74. Pnevmatikakis, E. A. et al. Simultaneous denoising, deconvolution, and demixing of calcium imaging data. *Neuron* **89**, 285–299 (2016).
75. Friedrich, J., Zhou, P. & Paninski, L. Fast online deconvolution of calcium imaging data. *PLoS Comput. Biol.* **13**, e1005423 (2017).
76. Bennett, C., Arroyo, S. & Hestrin, S. Subthreshold mechanisms underlying state-dependent modulation of visual responses. *Neuron* **80**, 350–357 (2013).
77. Ganguli, D. & Simoncelli, E. P. Efficient sensory encoding and bayesian inference with heterogeneous neural populations. *Neural Comput.* **26**, 2103–2134 (2014).
78. Seriès, P., Latham, P. E. & Pouget, A. Tuning curve sharpening for orientation selectivity: coding efficiency and the impact of correlations. *Nat. Neurosci.* **7**, 1129–1135 (2004).
79. Cover, T. M. & Thomas, J. A. *Elements of Information Theory* 2nd edn. (Wiley, 2006).
80. Chen, Y., Geisler, W. S. & Seidemann, E. Optimal decoding of correlated neural population responses in the primate visual cortex. *Nat. Neurosci.* **9**, 1412–1420 (2006).
81. Averbeck, B. B. & Lee, D. Effects of noise correlations on information encoding and decoding. *J. Neurophysiol.* **95**, 3633–3644 (2006).
82. Neal, R. M. Slice sampling. *Annals of Statistics* **31**, 705–767 (2003)
83. Gelman, A. & Rubin, D. B. Inference from iterative simulation using multiple sequences. *Stat. Sci.* **7**, 457–472 (1992).
84. Watanabe, S. A widely applicable Bayesian information criterion. *J. Mach. Learn. Res.* **14**, 867–897 (2013).
85. Gelman, A., Hwang, J. & Vehtari, A. Understanding predictive information criteria for Bayesian models. *Stat. Comput.* **24**, 997–1016 (2014).
86. Kalatsky, V. A. & Stryker, M. P. New paradigm for optical imaging: temporally encoded maps of intrinsic signal. *Neuron* **38**, 529–545 (2003).
87. Garrett, M. E., Nauhaus, I., Marshel, J. H. & Callaway, E. M. Topography and areal organization of mouse visual cortex. *J. Neurosci.* **34**, 12587–12600 (2014).
88. Waters, J. et al. Biological variation in the sizes, shapes and locations of visual cortical areas in the mouse. *PLoS ONE* **14**, e0213924 (2019).

Acknowledgements

We would like to thank Alexandre Pouget, Peter Latham, and members of the HMS Neurobiology Department for useful discussions and feedback on the work, and Rachel Wilson and Richard Born for comments on early versions of the manuscript. The work was supported by a scholar award from the James S. McDonnell Foundation (grant# 220020462 to J.D.), grants from the NIH (R01MH115554 to J.D.; R01MH107620 to C.D. H.; R01NS089521 to C.D.H.; R01NS108410 to C.D.H.; F31EY031562 to A.W.J.), the NSF's NeuroNex program (DBI-1707398. to R.N.), MINECO (Spain; BFU2017-85936-P to R.M.-B.), the Howard Hughes Medical Institute (HHMI, ref 55008742 to R.M.-B.), the ICREA Academia (2016 to R.M.-B.), the Government of Aragon (Spain; ISAAC lab, cod T33 17D to I.A.-R.), the Spanish Ministry of Economy and Competitiveness (TIN2016-80347-R to I.A.-R.), the Gatsby Charitable Foundation (to R.N.), and an NSF Graduate Research Fellowship (to A.W.J.).

Author contributions

All authors designed the research and wrote the paper; A.W.J. and S.N.C. performed the experiments; M.K., R.N., I.A.-R., R.M.-B., and J.D. developed the theory; and M.K., A.W.J., S.N.C., and J.D. analyzed the data.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-20722-y.

Correspondence and requests for materials should be addressed to J.D.

Peer review information *Nature Communications* thanks Joel Zylberberg and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer review reports are available.0202F

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Scaling of sensory information in large neural populations shows signatures of information-limiting correlations Supplementary Information

MohammadMehdi Kafashan, Anna W. Jaffe, Selmaan N. Chettih, Ramon Nogueira,
Iñigo Arandia-Romero, Christopher D. Harvey, Rubén Moreno-Bote, and Jan Drugowitsch

Contents

Supplementary Note 1. Generalized Fisher information

Fisher information quantifies how much neural activity r tells us about a stimulus θ around a particular reference θ_0. As such, it is a measure of fine discrimination performance. Here, we show how *linear* Fisher information relates to Fisher information in general, show how it can be generalized beyond fine discrimination, and describe some properties of this generalization.

1.1 Definition and properties of linear Fisher information

We can derive *linear* Fisher information in two ways [1, 2]. The first is to assume that $p(\mathbf{r}|\theta)$ is a member of the exponential family with linear sufficient statistics. The second is to show that it is the Fisher information that can be extracted with a minimum-variance unbiased linear decoder. We will provide both derivations in turn.

Let us first assume that neural activity r in response to a stimulus θ follows an exponential family distribution with linear sufficient statistics,

$$p(\mathbf{r}|\theta) = g(\theta)\Phi(\mathbf{r})\exp(\mathbf{h}(\theta)^T\mathbf{r}), \tag{1}$$

where

$$g(\theta) = \frac{1}{\int \Phi(\mathbf{r})\exp(\mathbf{h}(\theta)^T\mathbf{r})d\mathbf{r}}, \tag{2}$$

in which $g(\theta)$, $\Phi(\mathbf{r})$, and $\mathbf{h}(\theta)$ are known functions.

The partial derivative with respect to θ of the log-likelihood function, $\frac{\partial}{\partial\theta}\log p(\mathbf{r}|\theta)$, is called the "score" which is given by

$$\frac{\partial}{\partial\theta}\log p(\mathbf{r}|\theta) = \mathbf{h}'^T(\theta)(\mathbf{r}(\theta) - \mathbf{f}(\theta)), \tag{3}$$

where $\mathbf{f}(\theta) = \mathbb{E}\left(\mathbf{r}(\theta)\right)$ is the population activity vector. Note that the first moment of the score function is zero.

The Fisher information can be derived using the variance of the score function [3] which can be written as follows:

$$I(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log p(\mathbf{r}|\theta)\right)^2\right] = \mathbf{h}'(\theta)^T\mathbf{\Sigma}(\theta)\mathbf{h}'(\theta), \tag{4}$$

where $\mathbf{\Sigma}(\theta) = \mathbb{E}\left[\left(\mathbf{r}(\theta) - \mathbf{f}(\theta)\right)\left(\mathbf{r}(\theta) - \mathbf{f}(\theta)\right)^T\right]$ is the noise covariance matrix. To express the Fisher information in terms of $\mathbf{f}(\theta)$, we note that

$$\mathbf{f}'(\theta) = \frac{\mathrm{d}}{\mathrm{d}\theta}\int \mathbf{r}p(\mathbf{r}|\theta)d\mathbf{r} = \mathbf{\Sigma}(\theta)\mathbf{h}'(\theta). \tag{5}$$

Thus, we have $\mathbf{h}'(\theta) = \mathbf{\Sigma}^{-1}(\theta)\mathbf{f}'(\theta)$ [4]. Taking this expression to substitute both instances of $\mathbf{h}'(\theta)$ in the Fisher information results in

$$I(\theta) = \mathbf{f}'(\theta)^T\mathbf{\Sigma}^{-1}(\theta)\mathbf{f}'(\theta). \tag{6}$$

To show that linear Fisher information is the information extractable by a minimum-variance unbiased linear decoder, assume that the decoder linearly combines neural activity of neurons with a projection vector w. For fine discrimination task with two close-by stimuli $\theta_1 = \theta_0 - \delta\theta$ and $\theta_2 = \theta_0 + \delta\theta$ with small $\delta\theta$, the unbiased locally linear estimator for $\hat{\theta}$ is given by

$$\hat{\theta} - \theta_0 = \mathbf{w}^T(\mathbf{r} - \mathbf{f}(\theta_0)). \tag{7}$$

The expectation of the right-hand side around θ_0 is $\mathbf{w}^T\left(\langle\mathbf{r}\rangle - \mathbf{f}(\theta_0)\right) = 0$, demonstrating that the estimator is unbiased. Our aim is to find a w that yields a locally unbiased estimate, that is

$$\frac{\mathrm{d}\mathbb{E}_\theta(\hat{\theta})}{\mathrm{d}\theta} = 1, \tag{8}$$

imposing the constraint

$$\mathbf{w}^T\mathbf{f}'(\theta) = 1. \tag{9}$$

To find the minimum variance estimator satisfying this constraint, note that its variance is given by var $\left(\hat{\theta}\right) = \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}$, where $\boldsymbol{\Sigma}$ is the noise covariance matrix around θ_0. Therefore, we aim to find

$$\min_{\mathbf{w}} \mathbf{w}^T\boldsymbol{\Sigma}\mathbf{w}, \qquad \text{s.t. } \mathbf{w}^T\mathbf{f}'(\theta) = 1. \tag{10}$$

Using a Lagrange multiplier to solve the constraint optimization for w results in

$$\mathbf{w}^* = \frac{\boldsymbol{\Sigma}^{-1}\mathbf{f}'}{\mathbf{f}'^T\boldsymbol{\Sigma}^{-1}\mathbf{f}'}, \tag{11}$$

with associated estimator variance

$$\text{var}\left(\hat{\theta}\right) = \frac{1}{\mathbf{f}'^T\boldsymbol{\Sigma}^{-1}\mathbf{f}'}. \tag{12}$$

By the Cramér-Rao bound [3], the Fisher information is the inverse of this variance, resulting in

$$I(\theta) = \frac{1}{\text{var}\left(\hat{\theta}\right)} = \mathbf{f}'^T\boldsymbol{\Sigma}^{-1}\mathbf{f}', \tag{13}$$

which matches the previously derived expression for the linear Fisher information. This demonstrates that linear Fisher information can be interpreted in multiple ways: it is either the Fisher information when restricting the distribution of neural activity to the exponential family with linear sufficient statistics (which contains independent-Poisson populations with dense tuning curves, as well as other distributions [4], or the Fisher information that can be extracted with a linear decoder.

1.2 Generalizing Fisher information beyond fine discrimination

Let us generalize the above to coarse discrimination. To do so, assume two classes, C_1 and C_2, which represent a pair of stimulus orientations at θ_1 and θ_2 in the experiment. As before, we will derive generalized Fisher information in two ways. First, we will derive it by making particular distributional assumptions on $p(\mathbf{r}|\theta_1)$ and $p(\mathbf{r}|\theta_2)$. Then, we will derive it from the perspective of optimal linear discrimination.

For the first approach, assume that $p(\mathbf{r}|\theta_j)$ for both $j \in \{1, 2\}$ follows a Gaussian distribution,

$$\begin{aligned} C_1 &: \; p(\mathbf{r}|\theta_1) = \mathcal{N}(\mathbf{r}|\mathbf{f}_1, \boldsymbol{\Sigma}) \\ C_2 &: \; p(\mathbf{r}|\theta_2) = \mathcal{N}(\mathbf{r}|\mathbf{f}_2, \boldsymbol{\Sigma}), \end{aligned} \tag{14}$$

which have different means, but the same covariance matrix. Under the assumption that θ is a random variable (which takes two values, $\theta \in \{\theta_1, \theta_2\}$, in coarse discrimination tasks), it is easy to find a decision rule that minimize the expected Bayes risk [5]. We will denote L_{ij} as the loss of choosing C_j when C_i is correct. Furthermore, we assume a symmetric decision problem with symmetric loss, that is $L_{12} = L_{21}$ and $L_{11} = L_{22}$, a uniform prior $p(C_1) = p(C_2) = 1/2$, and a preference for making correct choices, that is $L_{11} < L_{12}$. In this case, the expected Bayesian risk, $\sum_{i\in\{1,2\}} L_{i\mathcal{D}(\mathbf{r})} p(C_i|\mathbf{r})$, associated with decision rule $\mathcal{D}(\mathbf{r}) \in \{1, 2\}$ is minimized by

$$\mathcal{D}\left(\mathbf{r}\right) = \begin{cases} 2 & \text{if } \Lambda(\mathbf{r}) = \log \frac{p(\mathbf{r}|\theta_2)}{p(\mathbf{r}|\theta_1)} > 0, \\ 1 & \text{otherwise ,} \end{cases} \tag{15}$$

where $\Lambda(\mathbf{r})$ is the log-likelihood ratio. For the assumed Gaussian likelihoods, this log-likelihood ratio is given by

$$\Lambda(\mathbf{r}) = (\mathbf{f}_2 - \mathbf{f}_1)^T\boldsymbol{\Sigma}^{-1}(\mathbf{r} - \mathbf{f}_0), \tag{16}$$

where $\mathbf{f}_0 = \frac{1}{2}(\mathbf{f}_1 + \mathbf{f}_2)$, and $\mathbf{f}_j = \mathbb{E}_{\mathbf{r}|\theta_j}(\mathbf{r})$ for $j \in \{1, 2\}$. Letting $\mathbf{w} = \boldsymbol{\Sigma}^{-1}\delta\mathbf{f}$ with $\delta\mathbf{f} = \mathbf{f}_2 - \mathbf{f}_1$, we can rewrite $\Lambda(\mathbf{r})$ as

$$\Lambda(\mathbf{r}) = \mathbf{w}^T(\mathbf{r} - \mathbf{f}_0). \tag{17}$$

In order to identify how likely this decision rule makes the correct choice, observe that $\Lambda(\mathbf{r})$ follows the following distributions under C_1 and C_2,

$$\Lambda(\mathbf{r})|C_1 \sim \mathcal{N}\left(-\frac{1}{2}\mathbf{w}^T\mathbf{\Sigma}\mathbf{w}, \mathbf{w}^T\mathbf{\Sigma}\mathbf{w}\right), \qquad \Lambda(\mathbf{r})|C_2 \sim \mathcal{N}\left(\frac{1}{2}\mathbf{w}^T\mathbf{\Sigma}\mathbf{w}, \mathbf{w}^T\mathbf{\Sigma}\mathbf{w}\right). \tag{18}$$

Therefore, we can find the probability of making a correct choice under $\mathcal{D}(\mathbf{r})$ by

$$p(\text{correct}) = \frac{1}{2}p\left(\Lambda(\mathbf{r}) \leq 0|C_1\right) + \frac{1}{2}p\left(\Lambda(\mathbf{r}) > 0|C_2\right) = \Phi\left(\frac{1}{2}\sqrt{\mathbf{w}^T\mathbf{\Sigma}\mathbf{w}}\right), \tag{19}$$

where $\Phi(\cdot)$ is the cumulative function of the standard normal distribution. After replacing both instances of w by its definition, $\mathbf{w} = \mathbf{\Sigma}^{-1}\delta\mathbf{f}$, p(correct) becomes

$$p(\text{correct}) = \Phi\left(\frac{1}{2}\sqrt{\delta\mathbf{f}^T\mathbf{\Sigma}^{-1}\delta\mathbf{f}}\right). \tag{20}$$

Comparing this expression to Eq. (6) reveals a close similarity which we can utilize to define the generalized linear Fisher information for coarse discrimination tasks by

$$I_g(\theta) = \frac{\delta\mathbf{f}^T\mathbf{\Sigma}^{-1}\delta\mathbf{f}}{\delta\theta^2}, \tag{21}$$

where $\delta\theta = \theta_2 - \theta_1$ is the stimulus difference. It is easy to see that, for small $\delta\theta$, generalized linear Fisher information converges to linear Fisher information,

$$\lim_{\delta\theta\to 0} I_g(\theta) = \lim_{\delta\theta\to 0} \frac{\delta\mathbf{f}^T\mathbf{\Sigma}^{-1}\delta\mathbf{f}}{\delta\theta^2} = \mathbf{f}'^T\mathbf{\Sigma}^{-1}\mathbf{f}' = I(\theta) \tag{22}$$

As the sensitivity index d' [6] in our case is given by $d' = \sqrt{\delta\mathbf{f}^T\mathbf{\Sigma}^{-1}\delta\mathbf{f}}$ [7, 8, 9], the generalized linear Fisher information can be re-expressed in terms of d' by

$$I_g(\theta) = \frac{d'^2}{\delta\theta^2}. \tag{23}$$

This relationship furthermore results in

$$p(\text{correct}) = \Phi\left(\frac{\delta\theta}{2}\sqrt{I_g(\theta)}\right) = \Phi\left(\frac{d'}{2}\right) \tag{24}$$

illustrating the close relationship between p(correct), d', and $I_g(\theta)$.

An alternative derivation for generalized linear Fisher information is through an optimal linear discriminator with less stringent assumptions on the class-conditional distribution. In this second approach, we assume a linear decoder projecting the neural activity to a one-dimensional readout using

$$\hat{\theta} = \mathbf{w}^T\mathbf{r}. \tag{25}$$

To assign an observed neural activity to a class, we just need to place a threshold on the readout $\hat{\theta}$. To do so, we optimize w to maximize the class separation following Fisher's linear discriminant analysis [10], which minimizes the within-class variance while maximizing the between-class variance of r. As before, let $\mathbf{f}_j$ and $\mathbf{\Sigma}_j$ be mean and noise covariance of neural activity in class C_j, but without making any further assumptions about the class-conditional densities $p(\mathbf{r}|C_j)$. We aim to find the w that maximizes the ratio of the between-class variance to the within-class variance, which is formulated as

$$\max_{\mathbf{w}} \frac{\mathbf{w}^T\delta\mathbf{f}\delta\mathbf{f}^T\mathbf{w}}{\mathbf{w}^T\mathbf{\Sigma}\mathbf{w}}, \qquad \text{s.t. } \|\mathbf{w}\|^2 = 1, \tag{26}$$

where $\delta\mathbf{f}\delta\mathbf{f}^T$ is the between-class covariance matrix and $\mathbf{\Sigma}$ is the average within-class covariance matrix given by

$$\mathbf{\Sigma} = \frac{\mathbf{\Sigma}_1 + \mathbf{\Sigma}_2}{2}. \tag{27}$$

Here, we fix $\|\mathrm{w}\|^2 = 1$, as we are interested in the direction of w but not its length. Using a Lagrange multiplier to solve the constraint optimization for w results in

$$\mathbf{w} = \frac{\mathbf{\Sigma}^{-1}\delta\mathbf{f}}{\delta\mathbf{f}^T\mathbf{\Sigma}^{-1}\delta\mathbf{f}}. \tag{28}$$

This yields the direction, w, to best project the neural activity into one dimension.

To find the associated $p(\text{correct})$, note that $\hat{\theta}$ is the sum of a (potentially large) set of random variables. These random variables are correlated, such that the central limit theorem does not directly apply. Nonetheless, we assume this sum to be approximately Gaussian for both $\hat{\theta}|C_1$ and $\hat{\theta}|C_2$, and given by

$$\hat{\theta}|C_1 \sim \mathcal{N}\left(\mathbf{w}^T\mathbf{f_1}, \mathbf{w}^T\mathbf{\Sigma}_1\mathbf{w}\right), \qquad \hat{\theta}|C_2 \sim \mathcal{N}\left(\mathbf{w}^T\mathbf{f_2}, \mathbf{w}^T\mathbf{\Sigma}_2\mathbf{w}\right). \tag{29}$$

This results in the sensitivity index, d', to be given by

$$d' = \frac{\mathbf{w}^T\mathbf{f_2} - \mathbf{w}^T\mathbf{f_1}}{\sqrt{\frac{1}{2}(\mathbf{w}^T\mathbf{\Sigma}_1\mathbf{w} + \mathbf{w}^T\mathbf{\Sigma}_2\mathbf{w})}} = \frac{\mathbf{w}^T\delta\mathbf{f}}{\sqrt{\mathbf{w}^T\mathbf{\Sigma}\mathbf{w}}} = \sqrt{\delta\mathbf{f}^T\mathbf{\Sigma}^{-1}\delta\mathbf{f}}, \tag{30}$$

yielding the same expression as before. This makes it straightforward to derive the generalized Fisher information as before.

1.3 Bias-corrected generalized Fisher information

Evaluating the generalized Fisher information, Eq. (21), by replacing $\delta\mathbf{f}$ and $\mathbf{\Sigma}$ by its empirical moments estimated from neural data with a limited number of trials leads to biased estimates [11]. In [11], they provide a bias correction for standard Fisher information, but it is unclear if this bias correction also applies to our generalization of Fisher information. In this section, we will derive such a bias correction for our generalization. This correction turns out to be the same as that provided by [11]. This is unsurprising in hindsight, as [11] do not restrict the size of $\delta\theta$ in their derivation, such that it applies to both fine and coarse discrimination.

We assume neural activity $\mathbf{r}_j^t,\ j = 1, 2$ in response to stimulus θ_j in trials $t = 1, ..., T$ to follow a multivariate Gaussian distribution given by

$$\boldsymbol{r}_j^t \sim \mathcal{N}\left(\mathbf{f}_j, \mathbf{\Sigma}\right), \qquad j = 1, 2, \tag{31}$$

where we assume the same covariance matrix for neural activity in response to θ_1 and θ_2. This is not a restriction, as our above derivation from the perspective of a linear discriminator has shown that, if these covariances differ, we can replace them by their average (which is what we do in practice, see below). Under this assumption, the empirical mean and covariance over T trials for each stimulus is distributed as [12]

$$\mu_j = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{r}_j^t \sim \mathcal{N}\left(\mathbf{f}_j, \frac{\mathbf{\Sigma}}{T}\right), \qquad \mathbf{S}_j = \frac{1}{T-1}\sum_{t=1}^{T}(\boldsymbol{r}_j^t - \mu_j)(\boldsymbol{r}_j^t - \mu_j)^T \sim \mathcal{W}\left(\frac{\mathbf{\Sigma}}{T-1}, T-1\right), \tag{32}$$

where $\mathcal{W}(V_{p\times p}, n)$ is the p-dimensional Wishart distribution with n degrees of freedom.

The naïve estimation of generalized Fisher information, Eq. (21), is obtained by replacing $\delta\mathbf{f}$ and $\mathbf{\Sigma}$ with their unbiased estimates, $\delta\mu$ and $\mathbf{S}$, given by

$$\delta\mu = \mu_1 - \mu_2 \sim \mathcal{N}\left(\delta\mathbf{f}, \frac{2\mathbf{\Sigma}}{T}\right), \qquad \mathbf{S} = \frac{1}{2}(\mathbf{S}_1 + \mathbf{S}_2) \sim \mathcal{W}\left(\frac{\mathbf{\Sigma}}{2(T-1)}, 2(T-1)\right), \tag{33}$$

where $\mathbb{E}(\delta\mu) = \delta\mathbf{f}$ and $\mathbb{E}(\mathbf{S}) = \mathbf{\Sigma}$. Furthermore, the inverse of sample covariance, $\mathbf{S}^{-1}$, follows an inverse Wishart distribution [12] given by

$$\mathbf{S}^{-1} \sim \mathcal{W}^{-1}\left(2(T-1)\mathbf{\Sigma}^{-1}, 2(T-1)\right), \tag{34}$$

which has mean

$$\mathbb{E}(\mathbf{S}^{-1}) = \frac{2(T-1)}{2T-N-3}\mathbf{\Sigma}^{-1} \tag{35}$$

Replacing δf and $\mathbf{\Sigma}$ with $\delta\mu$ and S in Eq. (21) results in the following naive estimator of the generalized Fisher information to be given by

$$\hat{I}_{g,nv}(\theta) = \frac{\delta\mu^T\mathbf{S}^{-1}\delta\mu}{\delta\theta^2}. \tag{36}$$

To evaluate the bias of $I_{g,nv}$, we utilize the fact that the sample mean and sample covariance of Gaussian distributions are independent [12], such that we can express the first moment of $I_{g,nv}$ by

$$\mathbb{E}\left(\hat{I}_{g,nv}\right) = \frac{\mathbb{E}_{\delta\mu,\mathbf{S}}\left(\delta\mu^T\mathbf{S}^{-1}\delta\mu\right)}{\delta\theta^2}, \tag{37}$$

where

$$\begin{aligned}
\mathbb{E}_{\delta\mu,\mathbf{S}}\left(\delta\mu^T\mathbf{S}^{-1}\delta\mu\right) &= \mathbb{E}_{\delta\mu,\mathbf{S}}\left(\mathrm{Tr}\left(\delta\mu\delta\mu^T\mathbf{S}^{-1}\right)\right) \\
&= \mathrm{Tr}\left(\mathbb{E}_{\delta\mu,\mathbf{S}}\left(\delta\mu\delta\mu^T\mathbf{S}^{-1}\right)\right) \\
&= \mathrm{Tr}\left(\mathbb{E}_{\delta\mu}\left(\delta\mu\delta\mu^T\right)\mathbb{E}_{\mathbf{S}}\left(\mathbf{S}^{-1}\right)\right) \\
&= \mathrm{Tr}\left(\left(\delta\mathbf{f}\delta\mathbf{f}^T + \frac{2\mathbf{\Sigma}}{T}\right)\left(\frac{2(T-1)}{2T-N-3}\mathbf{\Sigma}^{-1}\right)\right) \\
&= \frac{2(T-1)}{2T-N-3}\left(\mathrm{Tr}\left(\delta\mathbf{f}\delta\mathbf{f}^T\mathbf{\Sigma}^{-1}\right) + \frac{2N}{T}\right) \\
&= \frac{2(T-1)}{2T-N-3}\left(\delta\mathbf{f}^T\mathbf{\Sigma}^{-1}\delta\mathbf{f} + \frac{2N}{T}\right) \\
&= \frac{2(T-1)}{2T-N-3}\left(I_g\delta\theta^2 + \frac{2N}{T}\right).
\end{aligned} \tag{38}$$

Having the first moment of $\hat{I}_{g,nv}$, we can obtain the expression for the bias-corrected generalized Fisher information, $\hat{I}_{g,bc}$, given by

$$\hat{I}_{g,bc} = \frac{2T-N-3}{2(T-1)}\frac{\delta\mu^T\mathbf{S}^{-1}\delta\mu}{\delta\theta^2} - \frac{2N}{T\delta\theta^2}. \tag{39}$$

This estimate is the same as provided by [11], and will, in expectation, equal the true Fisher information, that is, $\mathbb{E}\left(\hat{I}_{g,bc}\right) = I_g$.

1.4 Variance of bias-corrected generalized Fisher information

Let us now consider the variance of the bias-corrected generalized Fisher information across different draws of T trial/samples from the same neural population. This variance has already been computed by [11], but only as a function of the true information, I_g, which is an unknown quantity. Here, we re-derive this expression for completeness, and additionally derive an unbiased estimated thereof as a function of $\hat{I}_{g,bc}$, which can be computed from experimental data.

The variance of $\hat{I}_{g,bc}$ is given by

$$\mathrm{var}\left(\hat{I}_{g,bc}\right) = \frac{(2T-N-3)^2}{4(T-1)^2\delta\theta^4}\mathrm{var}(\delta\mu^T\mathbf{S}^{-1}\delta\mu), \tag{40}$$

where $\mathrm{var}\left(\delta\mu^T\mathbf{S}^{-1}\delta\mu\right)$ can be decomposed into

$$\mathrm{var}(\delta\mu^T\mathbf{S}^{-1}\delta\mu) = \mathbb{E}\left((\delta\mu^T\mathbf{S}^{-1}\delta\mu)^2\right) - \mathbb{E}\left(\delta\mu^T\mathbf{S}^{-1}\delta\mu\right)^2. \tag{41}$$

The first term in Eq. (41) can be expressed as

$$\begin{aligned}\mathbb{E}\left((\delta\mu^T\mathbf{S}^{-1}\delta\mu)^2\right) &= \mathbb{E}\left(\delta\mu^T\mathbf{S}^{-1}\delta\mu\delta\mu^T\mathbf{S}^{-1}\delta\mu\right)\\ &= \mathbb{E}_{\delta\mu}\left(\delta\mu^T\mathbb{E}_{\mathbf{S}}\left(\mathbf{S}^{-1}\delta\mu\delta\mu^T\mathbf{S}^{-1}\right)\delta\mu\right)\\ &= \frac{4(T-1)^2}{(2T-N-3)(2T-N-5)}\mathbb{E}_{\delta\mu}\left(\delta\mu^T\mathbf{\Sigma}^{-1}\delta\mu\delta\mu^T\mathbf{\Sigma}^{-1}\delta\mu\right)\\ &= \frac{4(T-1)^2}{(2T-N-3)(2T-N-5)}\mathbb{E}_{\delta\mu}\left((\delta\mu^T\mathbf{\Sigma}^{-1}\delta\mu)^2\right)\\ &= \frac{4(T-1)^2}{(2T-N-3)(2T-N-5)}\left(\mathrm{var}(\delta\mu^T\mathbf{\Sigma}^{-1}\delta\mu) + \mathbb{E}_{\delta\mu}\left(\delta\mu^T\mathbf{\Sigma}^{-1}\delta\mu\right)^2\right).\end{aligned} \tag{42}$$

The second term in Eq. (41) can be expressed as

$$\mathbb{E}\left(\delta\mu^T\mathbf{S}^{-1}\delta\mu\right)^2 = \frac{4(T-1)^2}{(2T-N-3)^2}\mathbb{E}_{\delta\mu}\left(\delta\mu^T\mathbf{\Sigma}^{-1}\delta\mu\right)^2. \tag{43}$$

Together, this results in Eq. (41) to be given by

$$\mathrm{var}(\delta\mu^T\mathbf{S}^{-1}\delta\mu) = \frac{4(T-1)^2}{(2T-N-3)(2T-N-5)}\left(\mathrm{var}(\delta\mu^T\mathbf{\Sigma}^{-1}\delta\mu) + \frac{2}{2T-N-3}\mathbb{E}_{\delta\mu}\left(\delta\mu^T\mathbf{\Sigma}^{-1}\delta\mu\right)^2\right). \tag{44}$$

Therefore, $\mathrm{var}(I_{g,bc})$ can be simplified to

$$\mathrm{var}\left(\hat{I}_{g,bc}\right) = \frac{2}{2T-N-5}\left(\frac{2T-N-3}{2\delta\theta^4}\mathrm{var}(\delta\mu^T\mathbf{\Sigma}^{-1}\delta\mu) + \frac{1}{\delta\theta^4}\mathbb{E}_{\delta\mu}\left(\delta\mu^T\mathbf{\Sigma}^{-1}\delta\mu\right)^2\right). \tag{45}$$

To simplify this expression, note that if $\epsilon \sim \mathcal{N}(\mu, \mathbf{\Sigma})$, then, for a constant matrix $\mathbf{\Lambda}$, we have

$$\mathbb{E}(\epsilon^T\mathbf{\Lambda}\epsilon) = \mathrm{Tr}(\mathbf{\Lambda\Sigma}) + \mu^T\mathbf{\Lambda}\mu. \tag{46}$$

Additionally, for a symmetric matrix $\mathbf{\Lambda}$, the variance of the quadratic form is expressed as

$$\mathrm{var}(\epsilon^T\mathbf{\Lambda}\epsilon) = 2\,\mathrm{Tr}\left(\mathbf{\Lambda\Sigma\Lambda\Sigma}\right) + 4\mu^T\mathbf{\Lambda\Sigma\Lambda}\mu. \tag{47}$$

Applying Eqs. (46) and (47) yields

$$\mathrm{var}(\delta\mu^T\mathbf{\Sigma}^{-1}\delta\mu) = \frac{8N}{T^2} + \frac{8}{T}\delta\theta^2 I_g, \qquad \mathbb{E}_{\delta\mu}\left(\delta\mu^T\mathbf{\Sigma}^{-1}\delta\mu\right)^2 = \frac{4N^2}{T^2} + \frac{4N}{T}\delta\theta^2 I_g + \delta\theta^4 I_g^2. \tag{48}$$

Using these expressions results in the final variance

$$\mathrm{var}\left(\hat{I}_{g,bc}\right) = \frac{2}{2T-N-5}\left(I_g^2 + \frac{4(2T-3)}{T\delta\theta^2}I_g + \frac{4N(2T-3)}{T^2\delta\theta^4}\right) \tag{49}$$

This is the expression provided by [11]. Unfortunately, it is a function of the true information I_g, which is unknown, such that the variance cannot be evaluated from data.

To find an unbiased estimate of this variance, note that the true information, I_g, shows up as I_g and I_g^2. We already have an unbiased estimate of I_g, and will now derive such an unbiased estimate for I_g^2. Let us denote this estimate by $\left(\hat{I_g^2}\right)_{bc}$ (in contrast to the squared $\hat{I}_{g,bc}$, which is $\hat{I}_{g,bc}^2$). We find it by

$$\begin{aligned}\mathbb{E}\left((\hat{I}_{g,bc})^2\right) &= \mathrm{var}\left(\hat{I}_{g,bc}\right) + \mathbb{E}\left(\hat{I}_{g,bc}\right)^2\\ &= \frac{2}{2T-N-5}\left(I_g^2 + \frac{4(2T-3)}{T\delta\theta^2}I_g + \frac{4N(2T-3)}{T^2\delta\theta^4}\right) + I_g^2\\ &= \frac{1}{2T-N-5}\left((2T-N-3)I_g^2 + \frac{8(2T-3)}{T\delta\theta^2}I_g + \frac{8N(2T-3)}{T^2\delta\theta^4}\right).\end{aligned} \tag{50}$$

Solving for I_g^2 and substituting I_g by its bias-corrected estimate $\hat{I}_{g,bc}$ reveals the bias-corrected estimate

$$\left(\hat{I_g^2}\right)_{bc} = \frac{2T-N-5}{2T-N-3}\hat{I}^2_{g,bc} - \frac{1}{2T-N-3}\frac{8(2T-3)}{T\delta\theta^2}\hat{I}_{g,bc} - \frac{1}{2T-N-3}\frac{8N(2T-3)}{T^2\delta\theta^4}, \tag{51}$$

which satisfies $\mathbb{E}\left(\left(\hat{I_g^2}\right)_{bc}\right) = I_g^2$. Substituting the bias corrected estimates of I_g and I_g^2 into Eq. (49) results after some algebra in the unbiased variance estimate

$$\mathrm{var}\left(\hat{I}_{g,bc}\right) = \frac{2}{2T-N-3}\left(\hat{I}^2_{g,bc} + \frac{4(2T-3)}{T\delta\theta^2}\hat{I}_{g,bc} + \frac{4N(2T-3)}{T^2\delta\theta^4}\right), \tag{52}$$

which can be computed from data.

1.5 Covariance of bias-corrected generalized Fisher information

As we are interested in how information scales with population size, we also need to know how information estimates for different subpopulations relate to each other. Knowing this relationship is essential to our model fits, as fitting the information scaling models to information estimates that are correlated across different population sizes could results in significant mis-estimates if these correlations are ignored. In fact, we will use the results from this section to show in Sec. 3.5 that the increase in information due to adding one more neuron to a population is uncorrelated across different subpopulations. Based on this insight, we thus fitted these information increases rather than absolute informations, as illustrated in Supplementary Figure 4 in the main text.

To identify the relation between the information estimates for different subpopulations, we will focus on two subpopulations with N_x and N_y neurons ($N_y \leq N_x$) where the latter consists of a subset of neurons of the former. That is, the subpopulation with N_x neurons contains all of the N_y neurons in the (possibly) smaller subpopulation. We are interested in how their information estimates co-vary if we estimate both information measures from the same set of T trials.

To find this covariance, let us decompose the true (i.e., non-empirical) moments of the larger subpopulation into

$$\delta\mathbf{f}_x = \begin{pmatrix}\delta\mathbf{f}_y \\ \delta\mathbf{f}_z\end{pmatrix}, \qquad \boldsymbol{\Sigma}_x = \begin{pmatrix}\boldsymbol{\Sigma}_y & \boldsymbol{\Sigma}_u \\ \boldsymbol{\Sigma}_u^T & \boldsymbol{\Sigma}_z\end{pmatrix}. \tag{53}$$

Here, $\delta\mathbf{f}_x$ and $\delta\mathbf{f}_y$ are the population tuning differences of the larger and smaller subpopulation, respectively, and we have ordered the neurons in the larger subpopulation such that it contains all shared neurons first, followed by all non-shared neurons. This re-ordering is possible, as the information estimates are independent or how neurons are ordered within a population. Furthermore, $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_y$ are the noise covariance matrices of the larger and smaller subpopulation, and $\boldsymbol{\Sigma}_u$ is the the covariance of shared with non-shared neurons.

Experimentally, we cannot directly observe these moments, but instead estimate them through the empirical moments,

$$\delta\mu_x = \begin{pmatrix}\delta\mu_y \\ \delta\mu_{z.}\end{pmatrix}, \qquad \mathbf{S}_x = \begin{pmatrix}\mathbf{S}_y & \mathbf{S}_u \\ \mathbf{S}_u^T & \mathbf{S}_{z.}\end{pmatrix} \tag{54}$$

Using the same properties as in the previous section, these empirical moments relate to the true moments by

$$\delta\mu_x \sim \mathcal{N}\left(\delta\mathbf{f}_x, \frac{2}{T}\boldsymbol{\Sigma}_x\right), \qquad \mathbf{S}_x^{-1} \sim \mathcal{W}^{-1}\left(2(T-1)\boldsymbol{\Sigma}_x^{-1}, 2(T-1)\right), \tag{55}$$

$$\delta\mu_y \sim \mathcal{N}\left(\delta\mathbf{f}_y, \frac{2}{T}\boldsymbol{\Sigma}_y\right), \qquad \mathbf{S}_y^{-1} \sim \mathcal{W}^{-1}\left(2(T-1)\boldsymbol{\Sigma}_y^{-1}, 2(T-1)\right), \tag{56}$$

The empirical covariances additionally have the properties [13]

$$(\mathbf{S}_z - \mathbf{S}_u\mathbf{S}_y^{-1}\mathbf{S}_u^T)^{-1} \sim \mathcal{W}^{-1}\left(2(T-1)(\boldsymbol{\Sigma}_z - \boldsymbol{\Sigma}_u\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_u^T)^{-1}, 2(T-1)\right), \tag{57}$$

$$\mathbf{S}_u\mathbf{S}_y^{-1}|\mathbf{S}_y^{-1} \sim \mathcal{MN}_{N_y\times(N_x-N_y)}\left(\boldsymbol{\Sigma}_u\boldsymbol{\Sigma}_y^{-1}, \frac{1}{2(T-1)}(\boldsymbol{\Sigma}_x - \boldsymbol{\Sigma}_u\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_u^T), \mathbf{S}_y^{-1}\right), \tag{58}$$

where $\mathcal{MN}$ is the matrix-normal distribution.

From Eq. (39), the bias-corrected generalized information for two subpopulations denoted as $I^x_{g,bc}$ and $I^y_{g,bc}$ can be written as

$$\hat{I}^x_{g,bc} = \frac{2T - N_x - 3}{2T - 2}\frac{\delta\mu_x^T \mathbf{S}_x^{-1}\delta\mu_x}{\delta\theta^2} - \frac{2N_x}{T\delta\theta^2}, \tag{59}$$

$$\hat{I}^y_{g,bc} = \frac{2T - N_y - 3}{2T - 2}\frac{\delta\mu_y^T \mathbf{S}_y^{-1}\delta\mu_y}{\delta\theta^2} - \frac{2N_y}{T\delta\theta^2}. \tag{60}$$

We can decompose $\hat{I}^x_{g,bc}$ into two terms. The first term is the shared information which is common between subpopulations x and y as both of them contains all of neurons in subpopulation y. The second term is the information gain that is gained by adding the non-shared neurons. This decomposition can be expressed as

$$\hat{I}^x_{g,bc} = \hat{I}^y_{g,bc} + \delta\hat{I}^{x-y}_{g,bc}, \tag{61}$$

where $\delta\hat{I}^{x-y}_{g,bc}$ is the information gain due to the non-shared components between subpopulations x and y. The covariance of $\hat{I}^x_{g,bc}$ and $\hat{I}^y_{g,bc}$ is given by

$$\text{cov}\left(\hat{I}^x_{g,bc}, \hat{I}^y_{g,bc}\right) = \text{var}\left(\hat{I}^y_{g,bc}\right) + \text{cov}\left(\hat{I}^y_{g,bc}, \delta\hat{I}^{x-y}_{g,bc}\right), \tag{62}$$

where we already have expression for the variance on the right-hand side (i.e., Eq. (52)), and only need to find an expression for the covariance.

To calculate $\text{cov}\left(\hat{I}^y_{g,bc}, \delta\hat{I}^{x-y}_{g,bc}\right)$, let us first find an expression for $\delta\hat{I}^{x-y}_{g,bc}$. To find this expression, note that, by the decomposition of $\delta\mu_x$ and $\mathbf{S}_x$, and using block matrix inversion,

$$\delta\mu_x^T \mathbf{S}_x^{-1}\delta\mu_x = \delta\mu_y^T \mathbf{S}_y^{-1}\delta\mu_y + \left(\delta\mu_z - \mathbf{S}_u\mathbf{S}_y^{-1}\delta\mu_y\right)^T \left(\mathbf{S}_z - \mathbf{S}_u\mathbf{S}_y^{-1}\mathbf{S}_u^T\right)^{-1}\left(\delta\mu_z - \mathbf{S}_u\mathbf{S}_y^{-1}\delta\mu_y\right), \tag{63}$$

Substituting this relationship into Eqs. (59) and (60) results in the bias-corrected information gain

$$\begin{aligned}\delta\hat{I}^{x-y}_{g,bc} &= \hat{I}^x_{g,bc} - \hat{I}^y_{g,bc} \\ &= \frac{N_y - N_x}{2T - N_y - 3}\hat{I}^y_{g,bc} + \frac{2T - N_x - 3}{2T - 2}\frac{\left(\delta\mu_z - \mathbf{S}_u\mathbf{S}_y^{-1}\delta\mu_y\right)^T \left(\mathbf{S}_z - \mathbf{S}_u\mathbf{S}_y^{-1}\mathbf{S}_u^T\right)^{-1}\left(\delta\mu_z - \mathbf{S}_u\mathbf{S}_y^{-1}\delta\mu_y\right)}{\delta\theta^2} + \text{const},\end{aligned} \tag{64}$$

where "const" captures all non-stochastic terms that do not contribute to the covariance. Overall, this results in

$$\begin{aligned}\text{cov}\left(\hat{I}^y_{g,bc}, \delta\hat{I}^{x-y}_{g,bc}\right) = \frac{N_y - N_x}{2T - N_y - 3}\text{var}\left(\hat{I}^y_{g,bc}\right) + \frac{(2T - N_x - 3)(2T - N_y - 3)}{(2T-2)^2\delta\theta^4} \\ \times\, \text{cov}\left(\delta\mu_y^T \mathbf{S}_y^{-1}\delta\mu_y, \left(\delta\mu_z - \mathbf{S}_u\mathbf{S}_y^{-1}\delta\mu_y\right)^T \left(\mathbf{S}_z - \mathbf{S}_u\mathbf{S}_y^{-1}\mathbf{S}_u^T\right)^{-1}\left(\delta\mu_z - \mathbf{S}_u\mathbf{S}_y^{-1}\delta\mu_y\right)\right),\end{aligned} \tag{65}$$

where we have substituted Eq. (60) for $I^y_{g,bc}$ to find the second term on the right-hand side. The first term of Eq. (65) is known from Eq. (52). The covariance expression in the second term can be expressed as

$$\begin{aligned}&\text{cov}\left(\delta\mu_y^T \mathbf{S}_y^{-1}\delta\mu_y, \left(\delta\mu_z - \mathbf{S}_u\mathbf{S}_y^{-1}\delta\mu_y\right)^T \left(\mathbf{S}_z - \mathbf{S}_u\mathbf{S}_y^{-1}\mathbf{S}_u^T\right)^{-1}\left(\delta\mu_z - \mathbf{S}_u\mathbf{S}_y^{-1}\delta\mu_y\right)\right) \\ &\quad = \mathbb{E}\left(\delta\mu_y^T \mathbf{S}_y^{-1}\delta\mu_y \left(\delta\mu_z - \mathbf{S}_u\mathbf{S}_y^{-1}\delta\mu_y\right)^T \left(\mathbf{S}_z - \mathbf{S}_u\mathbf{S}_y^{-1}\mathbf{S}_u^T\right)^{-1}\left(\delta\mu_z - \mathbf{S}_u\mathbf{S}_y^{-1}\delta\mu_y\right)\right) \\ &\quad - \mathbb{E}\left(\delta\mu_y^T \mathbf{S}_y^{-1}\delta\mu_y\right)\mathbb{E}\left(\left(\delta\mu_z - \mathbf{S}_u\mathbf{S}_y^{-1}\delta\mu_y\right)^T \left(\mathbf{S}_z - \mathbf{S}_u\mathbf{S}_y^{-1}\mathbf{S}_u^T\right)^{-1}\left(\delta\mu_z - \mathbf{S}_u\mathbf{S}_y^{-1}\delta\mu_y\right)\right)\end{aligned} \tag{66}$$

First we evaluate the last expectation in Eq. (66) which is

$$\mathbb{E}\left(\left(\delta\mu_z - \mathbf{S}_u\mathbf{S}_y^{-1}\delta\mu_y\right)^T \left(\mathbf{S}_z - \mathbf{S}_u\mathbf{S}_y^{-1}\mathbf{S}_u^T\right)^{-1}\left(\delta\mu_z - \mathbf{S}_u\mathbf{S}_y^{-1}\delta\mu_y\right)\right) \tag{67}$$

Conditioned on $\mathbf{S}_y^{-1}$, we observe that $(\mathbf{S}_z - \mathbf{S}_u\mathbf{S}_y^{-1}\mathbf{S}_u^T)^{-1}$ is independent of $\mathbf{S}_u\mathbf{S}_y^{-1}$ [13]. Thus we can first take the expectation of $(\mathbf{S}_z - \mathbf{S}_u\mathbf{S}_y^{-1}\mathbf{S}_u^T)^{-1}$ to get

$$\frac{2T-2}{2T-N_x-3}\left(\delta\mu_z - \mathbf{S}_u\mathbf{S}_y^{-1}\delta\mu_y\right)^T\left(\mathbf{\Sigma}_z - \mathbf{\Sigma}_u\mathbf{\Sigma}_y^{-1}\mathbf{\Sigma}_u^T\right)^{-1}\left(\delta\mu_z - \mathbf{S}_u\mathbf{S}_y^{-1}\delta\mu_y\right). \tag{68}$$

Next, we observe that $\mathbf{S}_u\mathbf{S}_y^{-1}|\mathbf{S}_y^{-1}$ is matrix normal, which has a simple expression for the expectation of its quadratic form. Using this expression yields

$$\frac{2T-2}{2T-N_x-3}\left(\delta\mu_z - \mathbf{\Sigma}_u\mathbf{\Sigma}_y^{-1}\delta\mu_y\right)^T\left(\mathbf{\Sigma}_z - \mathbf{\Sigma}_u\mathbf{\Sigma}_y^{-1}\mathbf{\Sigma}_u^T\right)^{-1}\left(\delta\mu_z - \mathbf{\Sigma}_u\mathbf{\Sigma}_y^{-1}\delta\mu_y\right) + \frac{N_x-N_y}{2T-N_x-3}\delta\mu_y^T\mathbf{S}_y^{-1}\delta\mu_y \tag{69}$$

We do not need to complete the expectation over the remaining random variables because most involved terms cancel out each other later on.

Utilizing the same strategy we evaluate the expectation of the first term in Eq. (66) which is given by

$$\mathbb{E}\left(\delta\mu_y^T\mathbf{S}_y^{-1}\delta\mu_y\left(\delta\mu_z - \mathbf{S}_u\mathbf{S}_y^{-1}\delta\mu_y\right)^T\left(\mathbf{S}_z - \mathbf{S}_u\mathbf{S}_y^{-1}\mathbf{S}_u^T\right)^{-1}\left(\delta\mu_z - \mathbf{S}_u\mathbf{S}_y^{-1}\delta\mu_y\right)\right). \tag{70}$$

Its expectation with respect to $\left(\mathbf{S}_z - \mathbf{S}_u\mathbf{S}_y^{-1}\mathbf{S}_u^T\right)^{-1}$ is

$$\frac{2T-2}{2T-N-3}\delta\mu_y^T\mathbf{S}_y^{-1}\delta\mu_y\left(\delta\mu_z - \mathbf{S}_u\mathbf{S}_y^{-1}\delta\mu_y\right)^T\left(\mathbf{\Sigma}_z - \mathbf{\Sigma}_u\mathbf{\Sigma}_y^{-1}\mathbf{\Sigma}_u^T\right)^{-1}\left(\delta\mu_z - \mathbf{S}_u\mathbf{S}_y^{-1}\delta\mu_y\right). \tag{71}$$

The expectation with respect to $\mathbf{S}_u\mathbf{S}_y^{-1}|\mathbf{S}_y^{-1}$ is given by

$$\frac{2T-2}{2T-N_x-3}\delta\mu_y^T\mathbf{S}_y^{-1}\delta\mu_y\left(\delta\mu_z - \mathbf{\Sigma}_u\mathbf{\Sigma}_y^{-1}\delta\mu_y\right)^T\left(\mathbf{\Sigma}_z - \mathbf{\Sigma}_u\mathbf{\Sigma}_y^{-1}\mathbf{\Sigma}_u^T\right)^{-1}\left(\delta\mu_z - \mathbf{\Sigma}_u\mathbf{\Sigma}_y^{-1}\delta\mu_y\right) + \frac{N_x-N_y}{2T-N_x-3}\left(\delta\mu_y^T\mathbf{S}_y^{-1}\delta\mu_y\right)^2 \tag{72}$$

Utilizing the fact that $\delta\mu_y$ and $\delta\mu_z - \mathbf{\Sigma}_u\mathbf{\Sigma}_y^{-1}\delta\mu_y$ are jointly Gaussian and uncorrelated, which means they are independent, we can combine Eqs. (72) and (69) to simplify the expression in Eq. (66) to

$$\begin{aligned}
&\mathrm{cov}\left(\delta\mu_y^T\mathbf{S}_y^{-1}\delta\mu_y, \left(\delta\mu_z - \mathbf{S}_u\mathbf{S}_y^{-1}\delta\mu_y\right)^T\left(\mathbf{S}_z - \mathbf{S}_u\mathbf{S}_y^{-1}\mathbf{S}_u^T\right)^{-1}\left(\delta\mu_z - \mathbf{S}_u\mathbf{S}_y^{-1}\delta\mu_y\right)\right)\\
&\quad= \frac{N_x-N_y}{2T-N_x-3}\left(\mathbb{E}\left(\left(\delta\mu_y^T\mathbf{S}_y^{-1}\delta\mu_y\right)^2\right) - \mathbb{E}\left(\delta\mu_y^T\mathbf{S}_y^{-1}\delta\mu_y\right)^2\right)\\
&\quad= \frac{N_x-N_y}{2T-N_x-3}\mathrm{var}\left(\delta\mu_y^T\mathbf{S}_y^{-1}\delta\mu_y\right)\\
&\quad= \frac{(N_x-N_y)(2T-2)^2\delta\theta^4}{(2T-N_x-3)(2T-N_y-3)^2}\mathrm{var}\left(\hat{I}_{g,bc}^y\right).
\end{aligned} \tag{73}$$

Substituting Eq. (73) into Eq. (65) results in

$$\mathrm{cov}\left(\hat{I}_{g,bc}^y, \delta\hat{I}_{g,bc}^{x-y}\right) = \frac{N_y-N_x}{2T-N_y-3}\mathrm{var}\left(\hat{I}_{g,bc}^y\right) + \frac{N_x-N_y}{2T-N_y-3}\mathrm{var}\left(\hat{I}_{g,bc}^y\right) = 0, \tag{74}$$

which means that information in the smaller population is uncorrelated to the information gain obtained from non-shared neurons. As a consequence,

$$\mathrm{cov}\left(\hat{I}_{g,bc}^x, \hat{I}_{g,bc}^y\right) = \mathrm{var}\left(\hat{I}_{g,bc}^y\right). \tag{75}$$

Note the this only holds for the bias-corrected information estimates. For the naïve estimates, a similar derivation shows that $\hat{I}_{g,nv}^y$ and $\delta\hat{I}_{g,nv}^{x-y}$ are correlated.

Supplementary Note 2. Information scaling models

We assume that information in the recorded population is limited by the presence of information-limiting correlations [1]. In this case, the noise covariance matrix $\mathbf{\Sigma}_N$ for a population of N neurons decomposes into

$$\mathbf{\Sigma}_N = \mathbf{\Sigma}_{0,N} + \frac{1}{I_\infty}\mathbf{f}'_N\mathbf{f}'^T_N, \tag{76}$$

where $\mathbf{\Sigma}_{0,N}$ is the non-limiting covariance component, I_∞ is the asymptotic information, and $\mathbf{f}'_N$ is the derivative of the mean population activity. All of these quantities depend on the stimulus, θ, but we will keep this dependency implicit for notational convenience. In the $N \to \infty$ limit, only the second component limits information, while the information associated with $\mathbf{\Sigma}_{0,N}$ grows without bounds.

To see how information grows in the presence of information-limiting correlations, note that the Sherman-Morrison formula allows us to express $\mathbf{\Sigma}_N^{-1}$ by

$$\mathbf{\Sigma}_N^{-1} = \mathbf{\Sigma}_{0,N}^{-1} - \frac{\mathbf{\Sigma}_{0,N}^{-1}\mathbf{f}'_N\mathbf{f}'^T_N\mathbf{\Sigma}_{0,N}^{-1}}{I_\infty + \mathbf{f}'^T_N\mathbf{\Sigma}_{0,N}^{-1}\mathbf{f}'_N} \tag{77}$$

Let us denote the linear Fisher information associated with the non-limiting component by $I_{0,N} = \mathbf{f}'^T_N\mathbf{\Sigma}_{0,N}^{-1}\mathbf{f}'_N$. Then, after some algebra, the total Fisher information is given by

$$I_N = \mathbf{f}'^T_N\mathbf{\Sigma}_N^{-1}\mathbf{f}'_N = \mathbf{f}'^T_N\mathbf{\Sigma}_{0,N}^{-1}\mathbf{f}'_N - \frac{\mathbf{f}'^T_N\mathbf{\Sigma}_{0,N}^{-1}\mathbf{f}'_N\mathbf{f}'^T_N\mathbf{\Sigma}_{0,N}^{-1}\mathbf{f}'_N}{I_\infty + \mathbf{f}'^T_N\mathbf{\Sigma}_{0,N}^{-1}\mathbf{f}'_N} = \frac{1}{\frac{1}{I_{0,N}} + \frac{1}{I_\infty}}, \tag{78}$$

or, equally, $I_N^{-1} = I_{0,N}^{-1} + I_\infty^{-1}$. This result forms the core of our information scaling models. For the remainder of this section we will discuss how we would expect information $\mathbf{I}_{0,N}$ in the non-limiting component to scale, and the impact of measurement noise on overall information scaling.

2.1 Linear non-limiting information scaling for large N

To characterize the scaling of $I_{0,N}$ with N, let us use the spectral decomposition

$$\mathbf{\Sigma}_{0,N} = \sum_{n=1}^{N} \sigma^2_{N,n}\mathbf{z}_{N,n}\mathbf{z}^T_{N,n}, \tag{79}$$

with variances $\sigma^2_{N,1}, \dots, \sigma^2_{N,N}$ and principal directions $\mathbf{z}_{N,1}, \dots, \mathbf{z}_{N,N}$. Then, $I_{0,N}$ is given by

$$I_{0,N} = \sum_{n=1}^{N} \frac{\left(\mathbf{f}'^T_N\mathbf{z}_{N,n}\right)^2}{\sigma^2_{N,n}} = \|\mathbf{f}'_N\|^2 \sum_{n=1}^{N} \frac{\cos^2\left(\alpha_{N,n}\right)}{\sigma^2_{N,n}}, \tag{80}$$

where $\alpha_{N,n}$ is the angle between $\mathbf{f}'_N$ and $\mathbf{z}_n$.

To see how $I_{0,N}$ scales with N, let us assume that the $\alpha_{N,n}$'s are independent of the $\sigma^2_{N,n}$'s. Furthermore, $f'_{N,n}$ (that is, the nth component of $\mathbf{f}'_N$) is $\mathcal{O}(1)$, such that $\|\mathbf{f}'_N\|^2$ will be $\mathcal{O}(N)$. In addition, geometry requires that $\sum_{n=1}^{N}\cos^2\left(\alpha_{N,n}\right) = 1$, such that each $\cos^2\left(\alpha_{N_n}\right)$ is $\mathcal{O}(1/N)$ [1]. Together, this yields

$$I_{0,N} \propto N\sum_{n=1}^{N}\frac{1}{N}\frac{1}{\sigma^2_{N,n}} = \sum_{n=1}^{N}\frac{1}{\sigma^2_{N,n}}. \tag{81}$$

Therefore, under these assumptions, the scaling of $I_{0,N}$ only depends on the scaling of the eigenvalue spectrum $\{\sigma^2_{N,1}, \dots, \sigma^2_{N,N}\}$ of $\mathbf{\Sigma}_{0,N}$.

For the following, we will assume that each neuron in the population features some small amount of "private" noise that is not correlated with the variability of other neurons. This private noise introduces a lower bound,

σ_0^2, on the variances, that is $\sigma_{N,n}^2 \geq \sigma_0^2$ for all n. Together with the above expression, this allows us to derive a lower bound on the scaling of non-limiting information. In particular, by Jensen's inequality

$$I_{0,N} \propto N \left(\frac{1}{N} \sum_{n=1}^{N} \frac{1}{\sigma_{N,n}^2} \right) \geq N \frac{1}{\frac{1}{N} \sum_{n=1}^{N} \sigma_{N,n}^2} \propto N. \tag{82}$$

The second-to-last expression contains the average variance, which is lower-bounded by σ_0^2 and of order one. Therefore, the scaling of $I_{0,N}$ is at least $\mathcal{O}(N)$.

To gain further insight into the scaling of $I_{0,N}$, assume a sequence of non-limiting covariance matrices $\mathbf{\Sigma}_{0,M}, \mathbf{\Sigma}_{0,M-1}, \mathbf{\Sigma}_{0,M-2}, \dots$, starting with some large population with M neurons. Each consecutive matrix $\mathbf{\Sigma}_{0,N-1}$ is constructed from the next-larger matrix $\mathbf{\Sigma}_{0,N}$ by removing a single neuron, such that they share all entries except for one row and column associated with that neuron. If we order their eigenvalues according to $\sigma_{N,1}^2 \geq \sigma_{N,2} \geq \dots \sigma_{N,N}^2$ and $\sigma_{N-1,1}^2 \geq \sigma_{N-1,2}^2 \geq \dots \geq \sigma_{N-1,N-1}^2$, it is known that these eigenvalues obey the interleaved ordering

$$\sigma_{N,1}^2 \geq \sigma_{N-1,1}^2 \geq \sigma_{N,2}^2 \geq \sigma_{N-1,2}^2 \geq \dots \sigma_{N-1,N-1}^2 \geq \sigma_{N,N}^2. \tag{83}$$

Using $I_{0,N} \propto \sum_{n=1}^{N} \sigma_{N,n}^{-2}$, the information increase when moving from $N-1$ to N neurons becomes

$$I_{0,N} - I_{0,N-1} \propto \sum_{n=1}^{N-1} \left(\frac{1}{\sigma_{N,n}^2} - \frac{1}{\sigma_{N-1,n}^2} \right) + \frac{1}{\sigma_{N,N}^2}. \tag{84}$$

This information increase is $\mathcal{O}(1)$ if both terms on the left-hand side are $\mathcal{O}(1)$.

The second term is $\mathcal{O}(1)$ if there exists some positive constant C such that, for all N above some N_0, $\sigma_{N,N}^{-2} \leq C$. As $\sigma_{N,N}^2$ is always the smallest eigenvalue of the covariance matrix, this implies that $\mathcal{O}(1)$ can be guaranteed as long as $\sigma_{N,N}^2$ remains positive with increasing N, which is satisfied by our previous assumption that each neuron has some private noise. If it instead would go to zero, we would have $\lim_{N \to M} \sigma_{N,N}^{-2} = 0$, violating the requirement.

For the first term we observe that the hierarchical eigenvalue relationship of nested matrices implies that $\sigma_{N,n}^{-2} \leq \sigma_{N-1,n}^{-2}$ for all $n = 1, \dots, N-1$. This implies that every element in the sum is negative. However, the information increase $I_{0,N} - I_{0,N-1}$ cannot be negative. Therefore, the second term on the left-hand side has to be at least as large as the negative first term (i.e., the sum), that is

$$\frac{1}{\sigma_{N,N}^2} \geq - \sum_{n=1}^{N} \left(\frac{1}{\sigma_{N,n}^2} - \frac{1}{\sigma_{N-1,n}^2} \right). \tag{85}$$

As $\sigma_{N,N}^{-2}$ is $\mathcal{O}(1)$, the sum cannot be larger than $\mathcal{O}(1)$. Overall, as long as none of the variances become zero with increasing N, the increase in $I_{0,N}$ will be $\mathcal{O}(1)$, which implies that $I_{0,N}$ scales with $\mathcal{O}(N)$.

2.2 Models for $I_{0,N}$

The above argument shows that, under rather general conditions, $I_{0,N}$ can be expected to scale with $\mathcal{O}(N)$. However, it does not tell us about how $I_{0,N}$ behaves for small N, which depends on the details of the structure of $\mathbf{\Sigma}_{N,0}$.

To describe the details of this structure, we compared two models for $I_{0,N}$. The first, called the *lim* model, directly follows the scaling results and assumes that $I_{0,N} = cN$ with some parameter c that is independent of N. The second model, called the *lim-exp* model, allows the non-limiting information to initially grow supralinearly before converging to a linear growth. We derived this model by integrating $c\left(1 - e^{-N/\tau}\right)$ from zero to N, resulting in

$$I_{0,N} = c\left(N + \tau\left(e^{-\frac{N}{\tau}} - 1\right)\right), \tag{86}$$

with the additional parameter τ that controls the extent of the initial supralinearity (in units of N). We have chosen this particular model, as it turns out easier to fit than alternative models (such as, for example, integrating a re-scaled logistic sigmoid over the positive half-line) that provide qualitatively similar qualitative $I_{0,N}$ scaling. This model approaches $I_{0,N} = cN$ in the $\tau \rightarrow 0$ limit. Model comparison revealed the *lim* model to significantly outperform the *lim-exp* model (Supplementary Figure 6), such that we focused on the *lim* model in the main text.

2.3 Impact of measurement noise

Our recordings of neural activity might be noisy, introducing additional variability into our estimates of $\mathbf{\Sigma}_N$ and $\mathbf{f}'_N$. To estimate the effect of such measurement noise, we assume it to be of equal magnitude and independent across neurons, such that it adds an additional diagonal term to the covariance decomposition,

$$\mathbf{\Sigma}_N = \mathbf{\Sigma}_{0,N} + \frac{1}{I_\infty} \mathbf{f}'_N \mathbf{f}'^T_N + \sigma^2_{rec} \mathbf{I}, \tag{87}$$

where σ^2_{rec} denotes the variance of the measurement noise. We don't assume it to impact differential correlations, as those limit information *in the brain*, rather than our measurement thereof.

Following the same derivation as in the beginning of this section, the information in a population of N neurons becomes

$$I_N = \frac{1}{\frac{1}{I_{0,rec,N}} + \frac{1}{I_\infty}}, \tag{88}$$

where

$$I_{0,rec,N} = \mathbf{f}'^T_N \left(\mathbf{\Sigma}_{0,N} + \sigma^2_{rec} \mathbf{I} \right)^{-1} \mathbf{f}'_N, \tag{89}$$

is the non-limiting information, including measurement noise. We can, as before, use the spectral decomposition $\mathbf{\Sigma}_{0,N} = \sum_{n=1}^N \sigma^2_{N,n} \mathbf{z}_n \mathbf{z}_n^T$ and observe that $\mathbf{I} = \sum_{n=1}^N \mathbf{z}_n \mathbf{z}_n^T$, resulting in

$$\mathbf{\Sigma}_{0,N} + \sigma^2_{rec} \mathbf{I} = \sum_{n=1}^N \left(\sigma^2_{N,n} + \sigma^2_{rec} \right) \mathbf{z}_n \mathbf{z}_n^T. \tag{90}$$

This shows that measurement noise increases all eigenvalues of $\mathbf{\Sigma}_{0,N}$ by the same magnitude.

This has several consequences. First, the added variance baseline results in $I_{0,rec,N}$ to grow more slowly with N than $I_{0,N}$. Second, this baseline causes in the eigenvalues of $\mathbf{\Sigma}_{0,N} + \sigma^2_{rec}\mathbf{I}$ to be more similar to each other than those of $\mathbf{\Sigma}_{0,N}$ alone. As a consequence, the growth of $I_{0,rec,N} \propto \sum_{n=1}^N \left(\sigma^2_{N,n} + \sigma^2_{rec} \right)^{-1}$ with N is more linear than that of $I_{0,N} \propto \sum_{n=1}^N \sigma^{-2}_{N,n}$. This might make $I_{0,rec,N} = c_{rec} N$ a good model of non-limiting information growth, even if $I_{0,N} = cN$ is not. Third, as the measurement noise impacts only $I_{0,rec,N}$ but not I_∞, measurement noise only impacts our estimates of c but not of I_∞. Fourth, measurement noise will lower our estimates of c, and therefore increase our estimates of $N_a = a/(1-a) I_\infty / c$, which is the population size at which a fraction a of the asymptotic information I_∞ is reached.

2.4 Impact of eye movements

To estimate the impact of eye movements that might occur between trials, we assume that the only impact that they might have had was to have the stimulus appear outside some neurons' receptive field in some trials. We assumed that, under those circumstances, the neuron's activity would be set to zero. For tractability, our model does not consider details of the spatial structure of receptive fields. This leads to a simple model in which we assume one indicator variable z_t per trial t that is $z_t = 0$ if the stimulus appears in all measured neurons' receptive fields in that trial, and $z_t = 1$ if it might fall outside of some receptive fields. We furthermore introduce the indicator variable z_{nt} for neuron n in trial t that is $z_{nt} = 0$ if the stimulus falls into neuron n's receptive field in trial t, and $z_{nt} = 1$ if it might fall outside of it. We assume that $p(z_t = 1) = p_t$ and $p(z_{nt} = 1) = p_n$, independent across trials and neurons. Furthermore, p_n is the same across all neurons, and thus a scalar. Neuron n's activity in trial t is set to zero only if $z_t = 1$ *and* $z_{nt} = 1$. Formally, if $r_{nt,ori}$ is the neuron's unperturbed (original) spike count in that trial, its perturbed one becomes

$$r_{nt} = (1 - z_t z_{nt}) r_{nt,ori}. \tag{91}$$

For simplicity we here assumed z_t and z_{nt} to be independent, and are zeroing neurons only if $z_t z_{nt} = 1$, jointly. For $z_t = 0$, this occurs with probability $p(z_t z_{nt} = 1 | z_t = 0) = 0$, that is, never. For $z_t = 1$, it occurs with probability $p(z_t z_{nt} = 1 | z_t = 1) = p_n$.

We could have equivalently fixed $z_{nt} = 0$ for all trials in which $z_t = 0$, defined $p(z_{nt} = 1 | z_t = 1) = p_n$, and set $r_{nt} = (1 - z_{nt}) r_{nt,ori}$. This choice leads to $p(z_{nt} = 1 | z_t = 0) = 0$ and $p(z_{nt} = 1 | z_t = 1) = p_n$, illustrating that it yields the same result as our original assumptions. In either formulation, z_t mimics global fluctuations, whereas z_{nt} mimics fluctuations that are private to individual neurons.

2.4.1 Impact on population activity moments

As linear Fisher information only depends on $\mathbf{f}'$ and $\mathbf{\Sigma}$ we here assess the impact on occasionally zeroing out neurons on these moments. To do so, we denote $\mathbf{f}_{ori}$, $\mathbf{f}'_{ori}$, and $\mathbf{\Sigma}_{ori}$ as the unperturbed moments, and $\mathbf{f}$, $\mathbf{f}'$, and $\mathbf{\Sigma}$ as the same moments after perturbation. To find the perturbed mean, observe that

$$f_n = \langle r_{nt} \rangle = \langle (1 - z_t z_{nt}) r_{nt,ori} \rangle = (1 - p_t p_n) \langle r_{nt,ori} \rangle = (1 - p_t p_n) f_{n,ori}, \tag{92}$$

such that $\mathbf{f} = (1 - p_t p_n) \mathbf{f}_{ori}$. As $\mathbf{f}'$ is the difference between two $\mathbf{f}$'s, we furthermore have $\mathbf{f}' = (1 - p_t p_n) \mathbf{f}'_{ori}$.

The perturbed covariance follows a similar derivation, for $n \neq m$,

$$\begin{aligned}
\Sigma_{nm} &= \langle r_{nt} r_{mt} \rangle - \langle r_{nt} \rangle \langle r_{mt} \rangle \\
&= \langle (1 - z_t z_{nt})(1 - z_t z_{mt}) r_{nt,ori} r_{mt,ori} \rangle - \langle (1 - z_t z_{nt}) r_{nt,ori} \rangle \langle (1 - z_t z_{mt}) r_{mt,ori} \rangle \\
&= \langle (1 - z_t (z_{nt} + z_{mt} - z_{nt} z_{mt})) r_{nt,ori} r_{mt,ori} \rangle - (1 - p_t p_n)^2 f_{n,ori} f_{m,ori} \\
&= (1 - p_t p_n (2 - p_n)) \langle r_{nt,ori} r_{mt,ori} \rangle - (1 - p_t p_n)^2 f_{n,ori} f_{m,ori} \\
&= (1 - p_t p_n (2 - p_n)) (\langle r_{nt,ori} r_{mt,ori} \rangle - \langle r_{nt,ori} \rangle \langle r_{mt,ori} \rangle) + p_t (1 - p_t) p_n^2 f_{n,ori} f_{m,ori} \\
&= (a^2 + b^2) \Sigma_{nm,ori} + b^2 f_{n,ori} f_{m,ori},
\end{aligned} \tag{93}$$

where we have defined

$$a = 1 - p_t p_n, \qquad b = p_n \sqrt{p_t (1 - p_t)}. \tag{94}$$

A similar derivation for $n = m$ results in

$$\Sigma_{nn} = (a^2 + b^2) \Sigma_{nn,ori} + b^2 f_{n,ori}^2 + p_t p_n (1 - p_n) \left(\Sigma_{nn,ori} + f_{n,ori}^2 \right). \tag{95}$$

Together, this yields

$$\Sigma_{nm} = (a^2 + b^2) \Sigma_{nm,ori} + b^2 f_{n,ori} f_{m,ori} + \delta_{nm} p_t p_n (1 - p_n) \left(\Sigma_{nn,ori} + f_{n,ori}^2 \right), \tag{96}$$

where $\delta_{nm} = 1$ if $n = m$, and $\delta_{nm} = 0$ otherwise.

Overall, this results in

$$\mathbf{\Sigma} = \left(a^2 + b^2 \right) \mathbf{\Sigma}_{ori} + b^2 \mathbf{f}_{ori} \mathbf{f}_{ori}^T + p_t p_n (1 - p_n) \mathbf{S}, \tag{97}$$

where $\mathbf{S}$ is a diagonal matrix with $\Sigma_{nn,ori} + f_{n,ori}^2$ as the nth element of its diagonal. Therefore, the perturbed covariance $\mathbf{\Sigma}$ is a scaled-down version (as $a^2 + b^2 \leq 1$) of the unperturbed covariance $\mathbf{\Sigma}_{ori}$, with the scaling factor $a^2 + b^2$ decreasing monotonically in both p_n and p_t. Zeroing out neurons results in two additional perturbations. First, it causes the addition of the rank-one matrix $\mathbf{f}_{ori} \mathbf{f}_{ori}^T$ whose magnitude increases with p_n and the variance of z_t, $\text{var}(z_t) = p_t (1 - p_t)$, which is largest for $p_t = 1/2$. Second, it adds a diagonal component whose magnitude increases with p_t and the variance of z_{nt}, $\text{var}(z_{nt}) = p_n (1 - p_n)$.

Both $\mathbf{f}$ and $\mathbf{\Sigma}$ are computed conditional on a specific stimulus. The $\mathbf{\Sigma}$ we used to compute linear Fisher information is the average across the two considered stimuli. Then, the perturbed $\mathbf{\Sigma}$ becomes a linear combination of the average $\mathbf{\Sigma}_{ori}$, $\mathbf{f}_{ori} \mathbf{f}_{ori}^T$, and $\mathbf{S}$, for both stimuli.

2.4.2 Impact on information scaling

Let us consider the impact of the above perturbations on how information scales with population size, and its consequences for the estimated scaling parameters c and I_∞. If we ignore the rank-one and diagonal additions to the covariance matrix, the overall Fisher information for each N is re-scaled by

$$I_N \approx \frac{(1-p_n p_t)^2}{a^2+b^2} I_{N,ori}, \tag{98}$$

with a scaling factor that is close-to one for small p_n and p_t, and shrinks monotonically in p_n. Furthermore, it is smallest for interim values of p_t, but becomes one for $p_t = 0$ and $p_t = 1$. Based on the expression of our information scaling model, Eq. (78), as $I_{N,ori}^{-1} = c_{ori}^{-1} N^{-1} + I_{\infty,ori}^{-1}$, estimates of both c and I_∞ will be down-scaled by the same factor.

Adding a diagonal term to $\mathbf{\Sigma}$ has an analogous effect as measurement noise, as discussed in the previous section: it causes information to grow more slowly with N, thus lowering c further, but leaves I_∞ unperturbed.

The impact of the rank-one component $\mathbf{f}_{ori}\mathbf{f}_{ori}^T$ depends on the alignment between $\mathbf{f}_{ori}$ and $\mathbf{f}'_{ori}$. If they are perfectly aligned, that is, if $\mathbf{f}_{ori} \propto \mathbf{f}'_{ori}$, this component introduces differential correlations, thus lowering asymptotic information I_∞. If alignment is only partial, then this component does not limit information [1], but might still lower c. Non-alignment is likely if a change in the stimulus results in a simple shift in the population activity pattern, as is the case in our experiments. Perfect alignment could, for example, occur, if the stimulus modulates the gain of population activity, making the change in population activity due to a change in stimulus well-aligned with the average population activity. This might, for example, occur in contrast discrimination tasks, if contrast modulates the population activity gain.

In summary, zeroing out some neurons in a fraction of trials results in lowering our estimate of c, depending on a complex interplay of p_t and p_n. If $\mathbf{f}_{ori}$ is not perfectly aligned to $\mathbf{f}'_{ori}$, then I_∞ is down-weighted by $(1-p_n p_t)^2/(a^2+b^2)$. In case of perfect alignment, the asymptotic information estimate is suppressed further. We confirmed these effects in simulations, as shown in Supplementary Figure 14.

Supplementary Note 3. Estimating the information scaling moments from neural data

Here, we fix the discrimination (i.e., the pair of drift directions, θ_1 and θ_2) and discuss how we estimate the moments of Fisher information for different population sizes. To do so, we assume a large population with M neurons of which we subsample N neurons, and where $N \ll M$. Rather than focusing on the moments of the Fisher information I_n for population size $n \leq N$, we will instead focus on the moments of the Fisher information increase, $\Delta I_n = I_n - I_{n-1}$ (with $I_0 = 0$), when increasing the population size from $n-1$ to n neurons, for reasons that become apparent later. Our aim is to estimate the mean, $\mathbb{E}\left(\Delta I_n\right)$, the variance, $\text{var}\left(\Delta I_n\right)$, and the covariance, $\text{cov}\left(\Delta I_n, \Delta I_m\right)$, for different population sizes n and m.

3.1 Generative model and desired moments

To describe the stochasticity of ΔI_n, we assume the following generative process. Assume that neurons in the large population have indices 1 to M, and that we uniformly draw a subset of N different neurons with indices $i_1, i_2, \dots, i_N$, denoted $i_{1:N}$. This subpopulation has moments $\mathbf{f}'_{i_{1:N}}$ and $\mathbf{\Sigma}_{i_{1:N}}$, that in turn can be used to compute its associated Fisher information. However, we do not directly observe these moments, but instead record the population activity across T trials for each stimulus, θ_1 and θ_2, from which we compute the empirical moments $\gamma_{i_{1:N}}$ and $\mathbf{\Omega}_{i_{1:N}}$. These empirical moments are in turn used to compute the Fisher information increases $\Delta \hat{I}_{1:N}$, using the bias-corrected estimates discussed further above. In summary, the generative process follows the Markov chain

$$i_{1:N} \to \mathbf{f}'_{i_{1:N}}, \mathbf{\Sigma}_{i_{1:N}} \to \gamma_{i_{1:N}}, \mathbf{\Omega}_{i_{1:N}} \to \Delta \hat{I}_{1:N}. \tag{99}$$

In this Markov chain, the first and last transition are deterministic, and the center transition is stochastic. Therefore, we can write the generative model as

$$p\left(\Delta \hat{I}_{1:N}\right) = \sum_{i_{1:N}} p\left(\Delta \hat{I}_{1:N}\left(\gamma_{i_{1:N}}, \mathbf{\Omega}_{i_{1:N}}\right) | i_{1:N}\right) p\left(i_{1:N}\right), \tag{100}$$

where the Fisher information increases are a deterministic function of the empirical moments, and the sum is over different subpopulations drawn from the larger population. We assume these draws to be uniform, that is $p\left(i_{1:N}\right) \propto 1$.

To find the moments of $\Delta \hat{I}_n$, we use iterated expectation, variance, and covariance, which, for a Markov chain $Z \to X_1, X_2$ is given by

$$\mathbb{E}_{X_1}\left(X_1\right) = \mathbb{E}_Z\left(\mathbb{E}_{X_1|Z}\left(X_1\right)\right), \tag{101}$$

$$\text{var}_{X_1}\left(X_1\right) = \mathbb{E}_Z\left(\text{var}_{X_1|Z}\left(X_1\right)\right) + \text{var}_Z\left(\mathbb{E}_{X_1|Z}\left(X_1\right)\right), \tag{102}$$

$$\text{cov}_{X_1,X_2}\left(X_1, X_2\right) = \mathbb{E}_Z\left(\text{cov}_{X_1,X_2|Z}\left(X_1, X_2\right)\right) + \text{cov}_Z\left(\mathbb{E}_{X_1|Z}\left(X_1\right), \mathbb{E}_{X_2|Z}\left(X_2\right)\right). \tag{103}$$

Applied to our generative model, that yields the decompositions

$$\mathbb{E}_{\Delta \hat{I}_n}\left(\Delta \hat{I}_n\right) = \mathbb{E}_{i_{1:N}}\left(\mathbb{E}_{\Delta \hat{I}_n | i_{1:N}}\left(\Delta \hat{I}_n\right)\right), \tag{104}$$

$$\text{var}_{\Delta \hat{I}_n}\left(\Delta \hat{I}_n\right) = \mathbb{E}_{i_{1:N}}\left(\text{var}_{\Delta \hat{I}_n | i_{1:N}}\left(\Delta \hat{I}_n\right)\right) + \text{var}_{i_{1:N}}\left(\mathbb{E}_{\Delta \hat{I}_n | i_{1:N}}\left(\Delta \hat{I}_n\right)\right), \tag{105}$$

$$\text{cov}_{\Delta \hat{I}_n, \Delta \hat{I}_m}\left(\Delta \hat{I}_n, \Delta \hat{I}_m\right) = \mathbb{E}_{i_{1:N}}\left(\text{cov}_{\Delta \hat{I}_n, \Delta \hat{I}_m | i_{1:N}}\left(\Delta \hat{I}_n, \Delta \hat{I}_m\right)\right) + \text{cov}_{i_{1:n}}\left(\mathbb{E}_{\Delta \hat{I}_n | i_{1:N}}\left(\Delta \hat{I}_n\right), \mathbb{E}_{\Delta \hat{I}_m | i_{1:N}}\left(\Delta \hat{I}_m\right)\right), \tag{106}$$

where both variance and covariance are decomposed into (i) the (co)variance of the information increase for a fixed subpopulation $i_{1:N}$, averaged across different subpopulations, and (ii) how the average information increase for a fixed subpopulation (co)varies across different subpopulations.

Our data does not allow us to directly estimate these moments for two reasons. First, we don't observe the larger population, and so can't use it to draw different subpopulations from this larger population. We will address how we handle this limitation in the next subsection. Second, we only observe a single set of empirical moments, μ and S, for the subpopulation that we record from. We will address how we handle this limitation in the remaining subsections.

3.2 Simulating samples from a large, unobserved population

Our generative model assumes that we are subsampling N neurons from a large neural populations of M neurons. Our data, in contrast, are population recordings from a single neural population with N neurons. To use these recordings to simulate sampling from various subpopulations of the larger population, we assume these subpopulations to be statistically similar to the recorded population. That is, the different sampled subpopulations will contain neurons with similar activity statistics as the recorded population. Thus, each sampled subpopulation will contain all neurons from the recorded population, but in a different order for each sampled subpopulation. We will simulate this by introducing a new index set $j_{1:N} = j_1, j_2, \dots, j_N$ that, for each sampled subpopulation $i_{1:N}$, contains a random order of the indices $1, \dots, N$ of neurons in the recorded population. With this, all of the above moments across $i_{1:N}$ will become moments across $j_{1:N}$, while taking into account that the recorded subpopulation is used as a proxy for sampling different subpopulations from a larger populations. We will describe the consequences of this for each of the moments separately.

3.3 Estimating the mean

The desired mean of the information increase $\Delta\hat{I}_n$ is, by Eq. (104) the average information increase for a particular set of empirical moments, $\gamma_{i_{1:N}}$ and $\mathbf{\Omega}_{i_{1:N}}$, for a particular subpopulation $i_{1:N}$, averaged across different subpopulations. We deal with not being able to sample different subpopulations by replacing $i_{1:N}$ by a randomly ordered recorded population $j_{1:N}$. Furthermore, we cannot draw different empirical moments, $\gamma_{i_{1:N}}$ and $\mathbf{\Omega}_{i_{1:N}}$ for a given subpopulation, as would be required to compute $\mathbb{E}_{\Delta\hat{I}|i_{1:N}}\left(\Delta\hat{I}_n\right)$. We will replace this expectation with our best estimate thereof, which is the Fisher information increase estimate based on the bias-correctet Fisher information, estimated from the empirical moments of the recorded population, μ and S. Overall, this leads to the approximate estimate,

$$\mathbb{E}_{i_{1:N}}\left(\mathbb{E}_{\Delta\hat{I}_n|i_{1:N}}\left(\Delta\hat{I}_n\right)\right) \approx \mathbb{E}_{j_{1:N}}\left(\Delta\hat{I}_n\left(\mu_{j_{1:N}}, \mathbf{S}_{j_{1:N}}\right)\right), \tag{107}$$

where $\mu_{j_{1:N}}$ and $\mathbf{S}_{j_{1:N}}$ denote the empirical moments with neurons ordered according to $j_{1:N}$. As our Fisher information estimate is unbiased, the above estimate will be unbiased as well. In practice, we approximate the expectation over $j_{1:N}$ by 10000 random ordering.

3.4 Estimating the variance

The variance, Eq. (105), is decomposed into two terms. The first, $\mathbb{E}_{i_{1:N}}\left(\text{var}_{\Delta\hat{I}_n|i_{1:N}}\left(\Delta\hat{I}_n\right)\right)$, is the variance of the Fisher information increase for a fixed subpopulation, averaged across many subpopulations. This term captures the uncertainty in $\Delta\hat{I}_n$ due to using the empirical moments to estimate it. The second term, given $\text{var}_{i_{1:N}}\left(\mathbb{E}_{\Delta\hat{I}_n|i_{1:N}}\left(\hat{I}_n\right)\right)$, captures how the average Fisher information increase for a given subpopulation varies across different subpopulations. Our data doesn't allow us to compute either of these terms directly. However, it turns out that they are both well-approximated by how the Fisher information increase estimated from the empirical moments, μ and S, varies across different population orders, $j_{1:N}$, that is

$$\mathbb{E}_{i_{1:N}}\left(\text{var}_{\Delta\hat{I}_n|i_{1:N}}\left(\Delta\hat{I}_n\right)\right) + \text{var}_{i_{1:N}}\left(\mathbb{E}_{\Delta\hat{I}_n|i_{1:N}}\left(\Delta\hat{I}_n\right)\right) \approx \text{var}_{j_{1:N}}\left(\Delta\hat{I}_n\left(\mu_{j_{1:N}}, \mathbf{S}_{j_{1:N}}\right)\right). \tag{108}$$

To understand why this approximation works, we need to consider two components that contribute to the empirical moments of the recorded neurons. The first is that, for each neuron and each neuron pair, these empirical moments are noisy, as they are estimated from a limited number of trials. Thus, we can approximate the effect of using empirical rather than true moments, as captured by the first term in Eq. (105), by computing the variance across different neurons in the population, as achieved by the variance across different orderings, $j_{1:N}$. The second factor is that different neurons contribute different amounts of information to the population. This comes into play in the second term in Eq. (105), and is again well-approximated by the variance across different orderings, $j_{1:N}$. As it seems paradoxical that the same variance can capture both kinds of effects at the same time, we have demonstrated it in simulations of neural populations, shown in Supplementary Figure 16a.

3.5 Estimating the covariance

As the variance, the covariance, Eq. (106) can be decomposed into two terms that capture different sources of uncertainty. The first term, $\mathbb{E}_{i_{1:N}}\left(\text{cov}_{\Delta\hat{I}_n,\Delta\hat{I}_m|i_{1:N}}\left(\Delta\hat{I}_n,\Delta\hat{I}_m\right)\right)$, captures the uncertainty associated with estimating empirical moments from a limited number of trials. To find this covariance, assume $n \neq m$ and note that,

$$\begin{aligned}\text{cov}\left(\Delta\hat{I}_n,\Delta\hat{I}_m\right) &= \text{cov}\left(\hat{I}_n-\hat{I}_{n-1},\hat{I}_m-\hat{I}_{m-1}\right)\\ &= \text{cov}\left(\hat{I}_n,\hat{I}_m\right)-\text{cov}\left(\hat{I}_n,\hat{I}_{m-1}\right)-\text{cov}\left(\hat{I}_{n-1},\hat{I}_m\right)+\text{cov}\left(\hat{I}_{n-1},\hat{I}_{m-1}\right)\end{aligned} \tag{109}$$

where all covariances are conditional on $i_{1:N}$. Without loss of generality we can assume that $n > m$, and use Eq. (75) from Sec. 1.4 to find

$$\text{cov}\left(\Delta\hat{I}_n,\Delta\hat{I}_m\right) = \text{var}\left(\hat{I}_m\right)-\text{var}\left(\hat{I}_{m-1}\right)-\text{var}\left(\hat{I}_m\right)+\text{var}\left(\hat{I}_{m-1}\right)=0. \tag{110}$$

This shows, that, conditional on $i_{1:N}$, the information increase estimates are uncorrelated.

The second term, $\text{cov}_{i_{1:N}}\left(\mathbb{E}_{\Delta\hat{I}_n|i_{1:N}}\left(\Delta\hat{I}_n\right),\mathbb{E}_{\Delta\hat{I}_m|i_{1:N}}\left(\Delta\hat{I}_m\right)\right)$, captures how the average Fisher information increase associated with adding the nth neuron correlates with that when adding the mth neuron across different subpopulation samples. On average, these increases will be negatively correlated, for the following reason. The variance of the information estimate $\hat{I}_n = \sum_{k=1}^{n}\Delta\hat{I}_k$ can be decomposed into

$$\text{var}\left(\hat{I}_n\right)=\sum_{k=1}^{n}\left(\text{var}\left(\Delta\hat{I}_k\right)+2\sum_{l=1}^{k-1}\text{cov}\left(\Delta\hat{I}_k,\Delta\hat{I}_l\right)\right), \tag{111}$$

which shows the impact of the individual variances, as well as the covariance between estimates associated with different population sizes. For a population of M neurons, the estimate of total information, $\hat{I}_M$, will be the same, irrespective of how the neurons are ordered within that subset. Therefore, var $\left(\hat{I}_M\right)=0$. However, as, by definition, var $\left(\Delta\hat{I}_n\right)\geq 0$, the above decomposition implies that the covariances need to be on average negative, to ensure that the sum of variances and covariances becomes zero.

The same principle applies if we estimate the variance of $\Delta\hat{I}_n$ by shuffling the order, $j_{1:N}$, of neurons in a smaller, recorded population. If this population has N neurons, then var $\left(\mathbb{E}_{\hat{I}_N|j_{1:N}}\left(\hat{I}_N\right)\right)=0$, irrespective of $j_{1:N}$, such that the information increase estimates will be negatively correlated.

Recall that we use population order shuffling as a proxy for repeatedly subsampling N neurons from a larger population of M neurons. The shuffling-induced negative correlations arise from using the same N recorded neurons across all estimates. If we instead subsample a larger population, the different sampled subpopulations are bound to share a smaller number of neurons. For two subpopulations that share no neurons, these estimates would be completely uncorrelated. However, even for $N \ll M$, two random subpopulations of size N are likely to share neurons of the larger population. Indeed, the same intuition underlying the birthday paradox [14] tells us that we are almost guaranteed to find such shared neurons. However, the correlations don't only depend on the presence of shared neurons, but also on how many of them are shared, and the latter will decrease significantly for larger M. To show that this significantly lowers the impact of negative correlations on the total variance, we compare this variance computed with and without accounting for these correlations for different

M's. As Supplementary Figure 16b shows, their impact drops significantly with growing M. Therefore, we will approximate them to be zero, that is

$$\mathrm{cov}_{\Delta \hat{I}_n, \Delta \hat{I}_m} \left(\Delta \hat{I}_n, \Delta \hat{I}_m \right) \approx 0. \tag{112}$$

This results in an overestimate of the variance of the Fisher information estimate, and make our fits less certain, and, as a consequence, more conservative.

Supplementary Note 4. Population activity models

We used two different models to simulate population activity, as described below.

4.1 A Gaussian population activity model with limited information

We used a simple Gaussian activity model to satisfy the assumptions of Gaussianity underlying the generalized linear Fisher information, and to test some of the properties of our estimates. This model violates some properties of neural activity, like non-negativity, but is convenient for our purposes, as it supports fine control over the eigenvalues of $\mathbf{\Sigma}_0$, the alignment of $\mathbf{f}'$ to $\mathbf{\Sigma}_0$, and the asymptotic information, I_∞. For a population size of N neurons, we generated $\mathbf{\Sigma}_0$ by drawing a random orthonormal matrix $\mathbf{Z}_0$ of size $N \times N$ that forms the eigenvectors of $\mathbf{\Sigma}_0$. We parameterized the eigenvalues by $\sigma_{n,0}^2 = \sigma_0^2 + \sigma_b m^{-\beta}$, which together form the diagonal matrix $\mathbf{D}_0 = \operatorname{diag}\left(\sigma_{0,1}^2, \dots, \sigma_{0,N}^2\right)$. $\mathbf{Z}_0$ and $\mathbf{D}_0$ together specify $\mathbf{\Sigma}_0$ by $\mathbf{\Sigma}_0 = \mathbf{Z}_0 \mathbf{D}_0 \mathbf{Z}_0^T$. For a given $\mathbf{f}'$, the full noise covariance is then given by $\mathbf{\Sigma} = \mathbf{\Sigma}_0 + I_\infty^{-1} \mathbf{f}' \mathbf{f}'^T$.

For Supplementary Figure 7, we drew a random $\mathbf{f}' \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and subsequently rescaled the vector such that $\|\mathbf{f}'\| = g$. This makes the alignment of $\mathbf{f}'$ to the eigenvectors of $\mathbf{\Sigma}_0$ roughly uniform on average. For this figure, we use parameters $I_\infty = 20$ (or $I_\infty = \infty$ for the unlimited-information case), $g = 20$, $\sigma_0^2 = 10^{-3}$, $\sigma_b^2 = 1$, and $\beta = 0.1$.

For Supplementary Figure 16, we specified the alignment of $\mathbf{f}'$ to the eigenvectors of $\mathbf{\Sigma}_0$ by $\alpha_n = \sigma_\alpha^2 + \propto e^{-n/\tau_\alpha}$, normalized such that $\sum_n \alpha_n = 1$. This yields $\tilde{\mathbf{f}}' = \sum_n \alpha_n \mathbf{z}_n$ ($\mathbf{z}_n$ is the nth eigenvector of $\mathbf{\Sigma}_0$), and $\mathbf{f}' = \sqrt{g_f N} \tilde{\mathbf{f}}' / \|\tilde{\mathbf{f}}'\|$. The magnitude of $\mathbf{f}'$ here scales with $\sqrt{N}$ to ensure roughly similar information across different N's. The used parameters were $I_\infty = 100$, $g_f = 0.008$, $\sigma_0^2 = 5 \times 10^{-5}$, $\sigma_g^2 = 3$, $\beta = 0.5$, $\sigma_\alpha^2 = 10^{-3}$, and $\tau_\alpha = 30$, which results in population statistics comparable to those shown in Figs. 3 and 8 in the main text.

4.2 A visual hierarchy population activity model

We relied on [15] for a more realistic model of V1 population activity that is driven by pixel-level inputs. Details of this model can be found in [15]. Briefly, a population of N neurons responded to a $P \times P$ pixelated images $\mathbf{J}$ of an oriented Gabor. The nth neuron's linear filter $\mathbf{F}_n$ was for each (x, y) pixel determined by

$$c e^{-\frac{(x^2+y^2)}{2\sigma^2}} \cos\left(\frac{2\pi x}{\lambda} \cos(\theta_n) + \frac{2\pi y}{\lambda} \sin(\theta_n) + \phi\right), \tag{113}$$

where c is the Michelson contrast, θ_n determines the neuron's tuning, σ^2 determines the size of the exponential envelope, and λ and ϕ are the Gabor's frequency and phase, respectively. The filter was computed by the above function for each (x, y) and then standardized to have mean zero and unit variance across all (x, y). Image templates, $\mathbf{J}(\theta)$, in response to stimulus θ were generated equally, with θ_n replaced by the template's orientation, θ. Each neuron's gain, a_n, was drawn from a log-normal distribution with unit mean and variance σ_a^2, and then multiplied by the overall gain, g.

Neural population activity is assumed to arise from the image template with Gaussian pixel noise (zero mean, variance σ_0^2), followed by application of the per-neuron linear filters, $\mathbf{F}_n$, multiplied by their gain a_n, and a Poisson step. For Supplementary Figure 7, we estimated information from a set of trials, in each of which neural activity was generated from a different pixel noise instantiation. For Supplementary Figure 10, we skipped the Poisson step, as it introduced additional noise and was not required for the point we were trying to make. Instead, we estimated Fisher information from approximations to the neural mean responses and their covariance matrix, following [15]. We computed the mean response of neuron n to image $\mathbf{J}$ by $f_n(\theta) = \left[a_n \sum_{xy} F_{n,xy} J_{xy}(\theta)\right]_+$, where $[\cdot]_+$ is the threshold-linear function that sets negative values to zero. The population noise covariance was computed by

$$\mathbf{\Sigma}(\theta) = \sigma_0^2 \left(\mathbf{a}\mathbf{a}^T\right) \otimes \left[\mathbf{F}^T \mathbf{F}\right]_+ + \operatorname{diag}\left(\mathbf{a} \otimes \mathbf{f}(\theta)\right), \tag{114}$$

where $\otimes$ denotes the (element-wise) Hadamard product, $\mathbf{a} = (a_1, \dots, a_N)^T$ is the column vector of per-neuron gains, $\mathbf{F}$ is the $P^2 \times N$ filter matrix with per-neuron filters unrolled as vectors along its columns, and $\mathbf{f}(\theta)$ is the mean population activity in response to stimulus θ. The information was computed from $\mathbf{\Sigma}(\theta)$ and $\mathbf{f}(\theta)$.

For Figs. 7, 14, and 15 we used the parameters $\sigma = P/5$, $\lambda = P/1.5$, $\phi = 0$, $c = 1$, $g = 20$, and $\sigma_a = \sqrt{2}$, as in [15], and additionally different N's for different figures, $P = 32$ and $\sigma_0 = 0.25$. To simulate infinite information, we removed pixel noise by setting $\sigma_0 = 0$. For Supplementary Figure 10, we used the same parameters except $N = 1000$, $g = 10$, and $\sigma_0 = 0.11$, to achieve the desired level of information, and approximate information saturation within the simulated population size. In all simulations, neural tuning, θ_n, was uniformly distributed over $[-\pi, \pi]$, and pixels (x, y) were uniformly distributed over locations $[-(P-1)/2, (P-1)/2]$ in both dimensions.

5 Supplementary Tables

		Session						
Mouse	Contrast	A	B	C	D	E	F	G
1	10%	2.03%	2.65%					
2	10%	2.03%	3.11%					
3	10%	3.21%	2.44%	5.49%	6.23%	3.55%		
4	10%	5.19%	4.70%	4.46%	5.13%	3.66%	2.30%	5.38%
5	10%	4.13%	2.29%	2.84%	1.00%			
	25%	3.36%	1.72%	0.52%	0.33%			
6	10%	3.57%	2.34%	2.37%				
	25%	2.30%	1.34%	4.44%				

Supplementary Table 1: Percentage of neurons that show significant adaptation at the $p = 0.05$ level, for all sessions/mice. We assessed adaptation by asking if the response of each neuron to a drifting stimulus was significantly modulated by the drift direction of the preceding stimulus. We did so by asking if we could reject a non-adaptive model when comparing it to an adaptive model. The non-adaptive model fit neural responses r_i across trials $i = 1, 2, \dots$ by linear regression, using the model

$$r_i \sim \sum_{j=1}^{8} 1_{x_i=\theta_j} \beta_j, \tag{115}$$

where x_i is the stimulus' drift direction in trial i, θ_j is the j's of the eight drift direction used in our experiments, and 1_a is the indicator function that results in $1_a = 1$ if a is true, and $1_a = 0$ otherwise. In this model, β_j will be the neuron's average activity in response to stimulus θ_j. The adaptive model fits neural responses across trials by the following linear model,

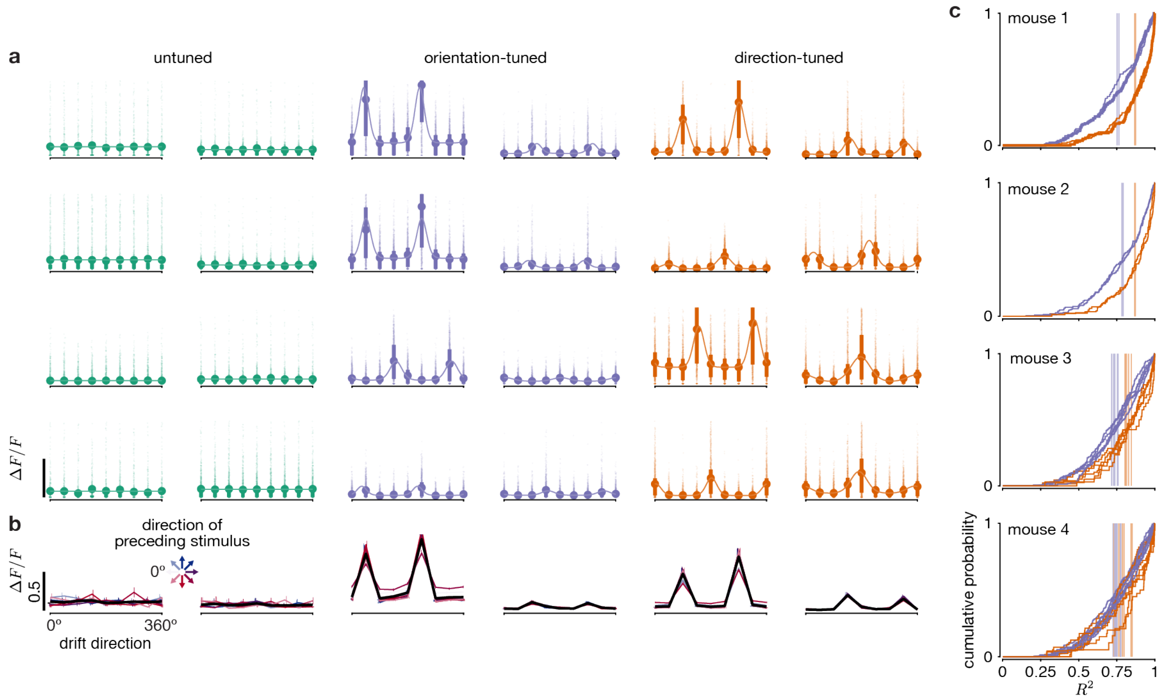
$$r_i \sim \sum_{j=1}^{8} \left(1_{x_i=\theta_j} \beta_j + \sum_{k=2}^{8} 1_{x_i=\theta_j, x_{i-1}=\theta_k} \beta_{jk} \right). \tag{116}$$

In this model, the neuron's mean activity in response to a stimulus with drift direction $x_i = \theta_j$ preceded by a stimulus with drift direction $x_{i-1} = \theta_1$ is β_j. The β_{jk}'s then model how this activity changes relative to β_j if the current trial's stimulus is instead preceded by a stimulus with drift direction $x_{i-1} = \theta_k$. For sessions with stimuli of multiple contrasts, the contrast in the table refers to that of the preceding stimulus (i.e., that in trial $i - 1$ of the *adapting* stimulus). As these models are nested (i.e., setting $\beta_{jk} = 0$ for all j and k turns the adaptive model into a non-adaptive one), we used an F-test to test the null hypothesis that the non-adaptive model fits the data better than the adaptive model. The above table shows the percentage of neurons for which this null hypothesis could be rejected at the $p = 0.05$ significance level. We performed a one-sided binomial test (H_0: fraction = 5%, testing for fraction $> 5\%$) to determine if any of the observed fractions are unlike to have arisen by chance, and found that none are significantly above 5% (one-sided $p > 0.124$ for all sessions/mice) [16]. This made us conclude that none of our datasets featured significant adaptation effects.

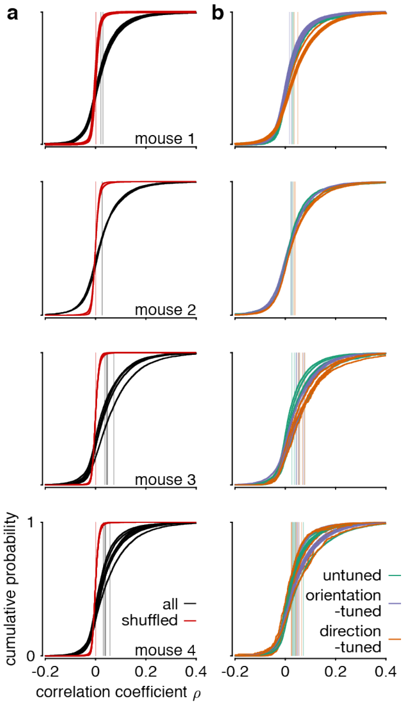
			Session						
Mouse	Contrast	Avg.	A	B	C	D	E	F	G
1	10%	0.12 ± 0.02	0.14	0.11					
2	10%	0.07 ± 0.03	0.11	0.04					
3	10%	0.16 ± 0.02	0.22	0.20	0.09	0.13	0.14		
4	10%	0.12 ± 0.01	0.10	0.11	0.12	0.10	0.13	0.13	0.16
5	10%	0.22 ± 0.06	0.12	0.16	0.20	0.40			
	25%	0.23 ± 0.03	0.19	0.20	0.22	0.32			
6	10%	0.20 ± 0.02	0.16	0.23	0.22				
	25%	0.22 ± 0.01	0.20	0.23	0.24				

Supplementary Table 2: Average Fisher information per neuron in $rad^{-2}/neuron$, across all sessions/mice, averaged across all $\delta\theta = 45^\circ$ discriminations. The average Fisher information was computed from the Fisher information scaling for trial-shuffled data that removed across-neuron correlations. For individual neurons, it can be computed by $2\left(\langle r|\theta_1\rangle - \langle r|\theta_2\rangle\right)^2 / \left(\delta\theta^2\left(\text{var}\left(r|\theta_1\right) + \text{var}\left(r|\theta_2\right)\right)\right)$, where $r|\theta_j$ is the neural response to stimulus θ_j. The *Avg.* column provides the average across sessions (mean ± 1 SEM).

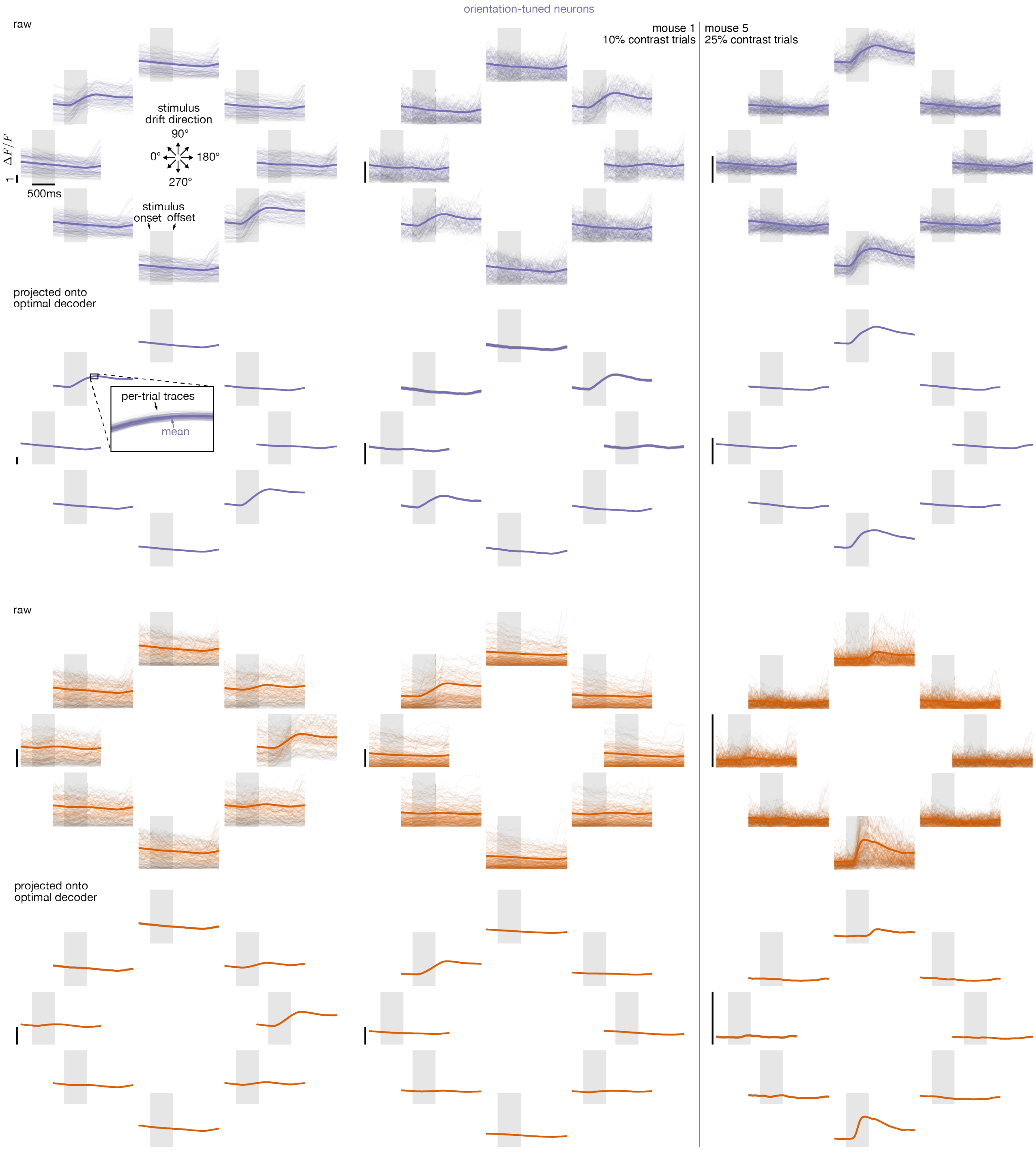
6 Supplementary Figures



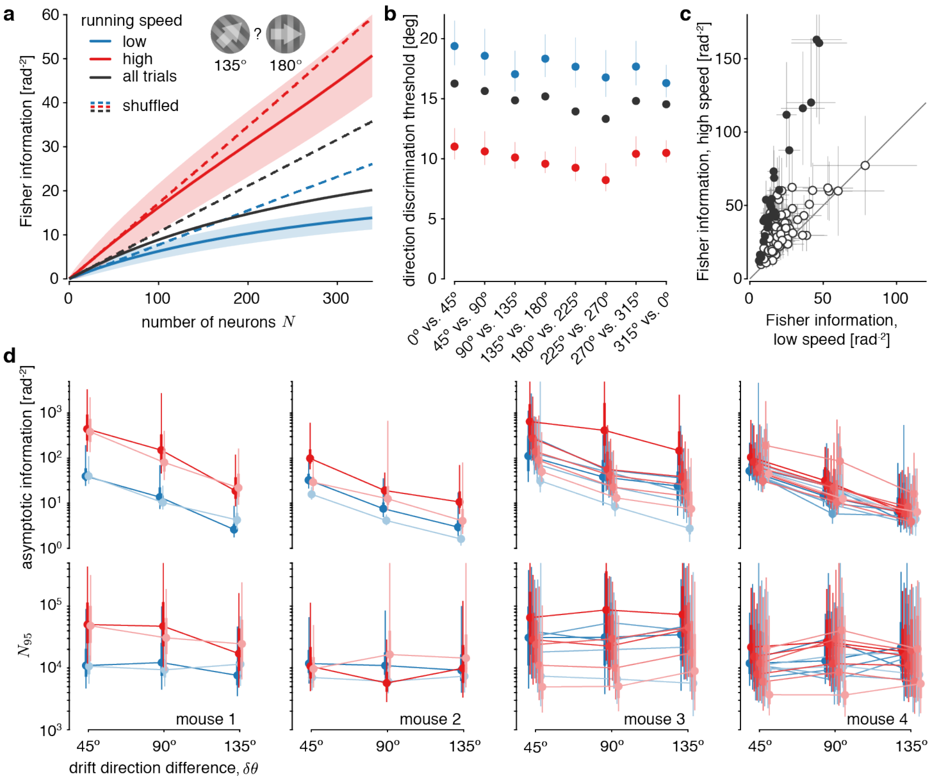
Supplementary Figure 1: Example tuning curves, absence of adaptation, and fitted tuning curve R^2's, for 10% contrast trials. (a) Eight examples of untuned, orientation-tuned, and direction-tuned neurons. We defined direction-tuned neurons (see Method) as having significantly higher responses for the tuned direction than the opposite direction. Orientation-tuned neurons are those for which this difference is not significant. The response of untuned neurons is not significantly modulated by drift direction. The pale, small dots show responses in individual trials. The large dots show mean responses for each drift direction, and the solid, vertical lines connect the 25th and 75th percentile. The pale lines show the fitted tuning curves. See Methods for how tuning was determined. Plots are truncated at $\Delta F/F = 1$. (b) Example direction tuning averaged across all trials (black), and across trials following a stimulus of a specific drift direction (colors, mean $\pm$ 1SEM, error bars horizontally shifted for visibility), demonstrating little to no adaptation to the preceding stimulus (see also Supplementary Supplementary Table 1). The shown example neurons are the same as in the first row of (a). (c) The cumulative distribution of coefficients of variations R^2 for different mice (rows) and sessions (line) for orientation-tuned (purple) and direction-tuned (orange) neurons. The pale vertical lines show the average R^2 for each session and neuron type. The R^2 for untuned neurons is not shown, as it is, by definition, $R^2 = 0$. (a) and (b) used data from one session of mouse 1 whose corresponding R^2 values are shown in bold in (c). All data at $0°$ is replicated at $360°$ to show the tuning curve across all possible drift directions. Note that fitted tuning curves were not used to estimate information, and are provided for reference only.



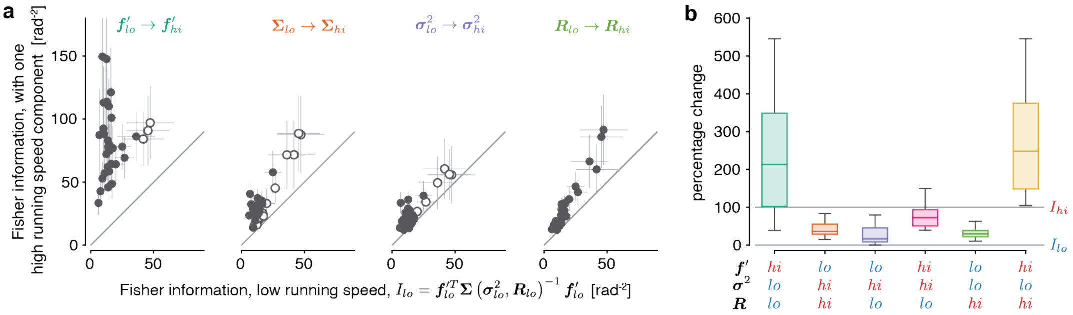
Supplementary Figure 2: Pairwise noise correlations across all neurons, and for neurons with specific tuning. (a) The cumulative distribution of pairwise noise correlations across all neurons collected within individual sessions (black lines) for mice 1-4 (rows). For reference, we also show these noise correlations for trial-shuffled data (red) that, on average, removes noise correlations. (b) The cumulative distribution of pairwise noise correlations for pairs of neurons with specific tuning (colors). See Methods for how tuning was determined. In both (a) and (b), the vertical lines show the mean pairwise correlations for the respective session and type of analysis. These mean correlations where comparable to those found in previous studies [17]. For mouse 1, the bold lines correspond to the session for which tuning curve examples are shown in Supplementary Figure 1(a)/(b). Note that the average pair-wise correlations were not used to estimate information, and are provided for reference only.



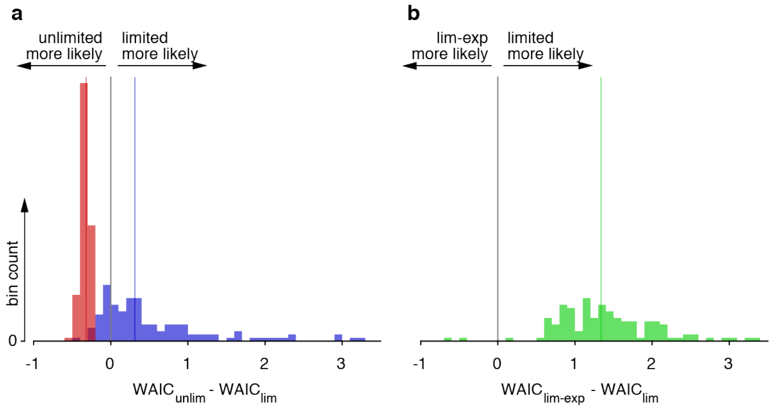
Supplementary Figure 3: Examples of raw $\Delta F/F$ traces and traces projected onto the optimal decoder. Each circle of 8 panels shows 200 raw $\Delta F/F$ time-course examples (thin lines) as well as their mean (thick line) for one example neuron, grouped by different stimulus drift directions (different panels; grey-shaded area = stimulus presentation period; black horizontal bar = 1 $\Delta F/F$). The top six panel circles show three orientation-tuned example neurons, and the bottom six panel circles represent three direction-tuned example neurons (left two columns: mouse 1, all trials 10% contrast; right column: mouse 5, only 25% contrast trials; see Supplementary Figure 1 for definition of orientation/direction-tuning). The top row in each neuron group shows the raw traces, and the bottom row shows the same traces projected onto the optimal decoder, $\mathbf{w} \propto \mathbf{\Sigma}^{-1}\mathbf{f}'$. The low variability of per-trial trace examples after this projection (frequently obscured by the across-trial mean; see enlarged inset) illustrates that most variability of the raw traces does not impact information, as it is orthogonal to the signal direction. The optimal decoder was for each stimulus drift direction computed (from temporally deconvolved traces) as the best discriminator between this and the next-closest drift direction.



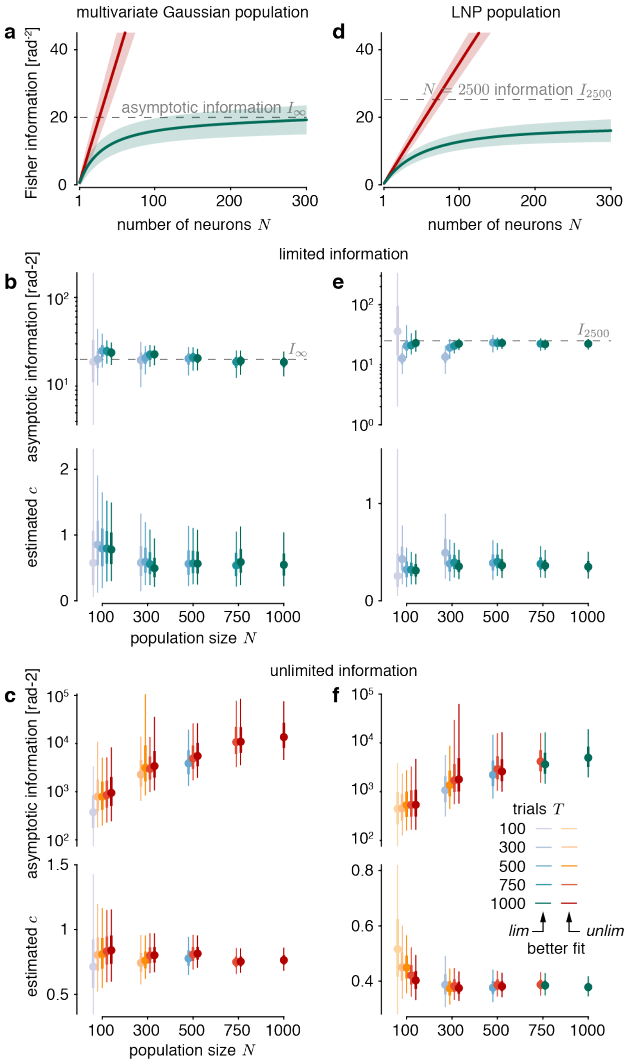
Supplementary Figure 4: Impact of running speed on information. For each session, we grouped trials into low (blue) and high (red) running speed by (i) subsampling trials to ensure a similar running speed distribution across all drift directions, and (ii) performing a median split by running speed for each drift direction. To maximize the number of analyzed trials, we did not aim to achieve the same average running speeds within each speed group across sessions — these average running speeds within each group might thus differ across sessions. Furthermore, mice 3 and 4 were overall running less than mice 1 and 2. (a) Information increases more rapidly with population size for higher running speeds (same mouse/session as in Supplementary Figure 3c; mean ± 1SD across random orderings of neurons within the population). The black line shows the mean information growth across all trials (as in Supplementary Figure 3c). The dashed lines show trial-shuffled data that removes pairwise noise correlations, and illustrate that, in all cases, these correlations lower information across all population sizes. (b) The drift direction discrimination threshold (80% correct) inferred from the information estimated in the recorded population is consistent across different drift direction pairs with $\delta\theta = 45^\circ$, and is lower for high running speeds. The black dots show the inferred thresholds across all trials (as in Supplementary Figure 3f). (c) Higher running speed increases information in the recorded population. Each dot (mean ± 1SD of information estimate; filled = significant increase, bootstrap, p<0.05) shows the information estimated for one discrimination with $\delta\theta = 45^\circ$. Across all sessions, higher running speed significantly increased information ($t_{63} = 6.69$, two-sided $p \approx 7 \times 10^{-9}$, across non-overlapping $\delta\theta = 45^\circ$ discriminations). (d) Both estimated asymptotic information and N_{95} appear impacted by running speed (lines connect median estimates of individual sessions; horizontally shifted to ease comparison; posterior densities as is Supplementary Figure 4c). Across sessions, we found a significant increase in asymptotic information (signed-rank on median estimates; $\delta\theta = 45^\circ$: $z = 2.64$, two-sided $p \approx 8.36 \times 10^{-3}$; $\delta\theta = 90^\circ$: $z = 2.69$, two-sided $p \approx 7.17 \times 10^{-3}$; $\delta\theta = 135^\circ$: $z = 3.36$, two-sided $p \approx 7.76 \times 10^{-4}$; not adjusted for multiple comparisons across $\delta\theta$), but not N_{95} (signed-rank on median estimates; $\delta\theta = 45^\circ$: $z = 0.672$, two-sided $p \approx 0.501$; $\delta\theta = 90^\circ$: $z = 1.14$, two-sided $p \approx 0.255$; $\delta\theta = 135^\circ$: $z = 1.500$, two-sided $p \approx 0.134$; not adjusted for multiple comparisons across $\delta\theta$) with running speed.



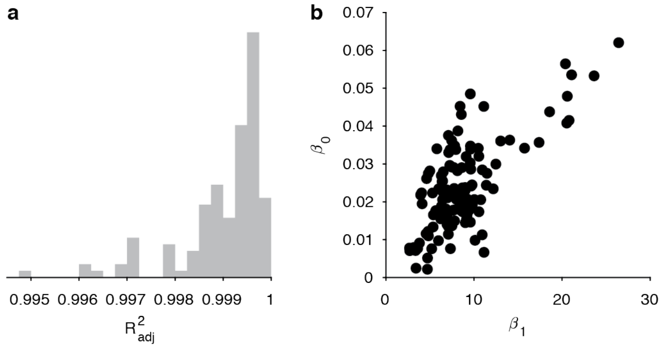
Supplementary Figure 5: Information boost due to increased running speed results from a combination of multiple factors. To identify which changes to the population response statistics are responsible for the boost in information for increased running speeds (see Supplementary Figure 4), we focused on $\delta\theta = 45^\circ$ discriminations in each session for which we observed a significant information increase (filled dots in Supplementary Figure 4c). We focused on three factors: (i) a change in mean responses, $\mathbf{f}'$, (ii) a change in per-neuron noise variances, σ^2, and (iii) a change in pairwise noise correlations, $\mathbf{R}$. Per-neuron variances and pairwise noise correlations completely determine the noise covariance matrix $\mathbf{\Sigma}\left(\sigma^2, \mathbf{R}\right)$. For each considered discrimination, we computed these factors separately for trials in which the running speed was low, denoted $\cdot_{lo}$, and in which it was high, denoted $\cdot_{hi}$. With this notation, the information in the recorded population for low and high running speeds, as shown in Supplementary Figure 4, is given by $I_{lo} = \mathbf{f}'^{T}_{lo}\mathbf{\Sigma}^{-1}_{lo}\mathbf{f}'_{lo}$ and $I_{hi} = \mathbf{f}'^{T}_{hi}\mathbf{\Sigma}^{-1}_{hi}\mathbf{f}'_{hi}$, where we have used the short-hand notation $\mathbf{\Sigma}_x = \mathbf{\Sigma}\left(\sigma^2_x, \mathbf{R}_x\right)$ with $x \in \{lo, hi\}$. (a) All factors boost information individually (mean ± 1SD of information estimate; filled = significant increase, bootstrap, p<0.05). To see the impact on information if only a single of these factors changes, we compared I_{lo} for all considered discriminations to the information when a single factors (except for $\mathbf{\Sigma}$, for which we changed both σ^2 and $\mathbf{R}$) was changed from $\cdot_{lo}$ to $\cdot_{hi}$. Across all considered discriminations we observed a significant information increase due to all factors (two-sided t-test, $t_{32} < -7.22$, $p < 3.34 \times 10^{-8}$). (b) Relative information boost on the I_{lo} — I_{hi} scale. To see how the information due to separate factors compares to that due to all factors, we computed the percentage of information boost for each considered discrimination and factor combination (different bars) where the information boost lies on the scale from I_{lo} (0%) to I_{hi} (100%). Each box plot shows the median, the 25% and 75% percentiles (box) and the extremes (whiskers) of the percentage information boost for a given combination of factors across considered discriminations. The relative information boost can exceed 100% in cases in which changing the remaining $\cdot_{lo}$ factor to $\cdot_{hi}$ results in a drop of information, as seen for $(\mathbf{f}'_{hi}, \mathbf{\Sigma}_{lo})$ and $\left(\mathbf{f}'_{hi}, \sigma^2_{lo}, \mathbf{R}^2_{hi}\right)$. Both in combination show that the change of $\mathbf{f}'$ from $\mathbf{f}'_{lo}$ to $\mathbf{f}'_{hi}$, which results in an information boost beyond I_{hi}, is compensated by a change of the per-neuron variances σ^2 from σ^2_{lo} to σ^2_{hi}. n=4 mice with a total of 16 sessions, resulting in 128 discriminations that were utilized for this analysis.



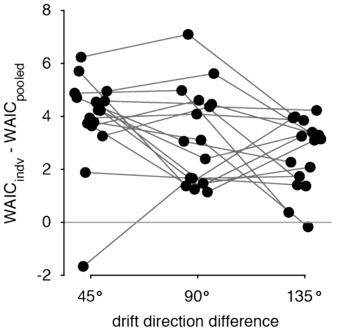
Supplementary Figure 6: Model comparison of different information scaling models. Both panels show histograms of differences in the Watanabe-Akaike Information Criterion (WAIC) for two different models fitted to the measured information scaling curves across all eight discriminations with $\delta\theta = 45°$, sessions, and mice. (a) shows the WAIC difference for fitting a model that assumes no information limitation (*unlim*) to one that does (*lim*), for regular (blue) and shuffled (red) data. For regular data this difference is in most cases positive, indicating that the information-limiting model fits the data better. In fact, even for individual negative WAIC differences, the average across all eight WAIC difference within a session remains positive. For shuffled data, a model assuming no information limitation fits the data better in all instances. This confirms that our model comparison is not biased towards the model assuming limited information. (b) shows the WAIC difference for fitting two models that assume limited information (see Sec. 2.2), one with linear scaling of the non-limiting component (*lim*), and one assuming initial supralinear scaling of that component (*lim-exp*). The latter only fits the data better in few instances. In those, the average WAIC difference across all discriminations within that session is nonetheless positive. The colored lines in (a) and (b) show the median WAIC difference across all comparisons.



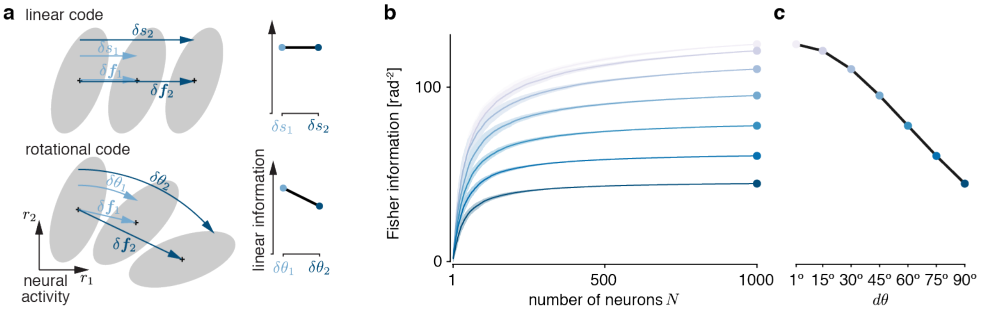
Supplementary Figure 7: Recovering asymptotic information from simulated population activity. We simulated neural population activity, using either a multivariate Gaussian population model (a-c; see Sec. 4.1 for details) or a linear-nonlinear Poisson model (d-f; see Sec. 4.2 for details) and fitted a linear scaling (*unlim*) and a limited information scaling model (*lim*). For each model type, we generated two large datasets (limited information and unlimited information; $\delta\theta = \theta_2 - \theta_1 = 45^\circ$ in both cases) and then subsampled neurons and trials to perform the fits. (a,d) Example information scaling for $N = 300$ and $T = 500$ (mean ± 1SD information estimation; green/red = limited/unlimited information). For the Gaussian model we could specify the asymptotic information I_∞ (dashed grey line). For the LNP model we estimated it from the information I_{2500} at $N = 2500$ neurons. (b-c,e-f) Estimated asymptotic information and non-limiting information scaling for the *lim* model from data with different population sizes N and numbers of trials T per stimulus. The posterior estimates are shown as in Supplementary Figure 4c in the main text. Blue/green and orange/red colors indicate a better fit by the *lim* and *unlim* model (WAIC for model comparison), respectively. Asymptotic information is well-estimated by the *lim* model (b,e), and more certain for larger N and T. Model comparison in most cases (28 out of 30 for Gaussian model, 26 out of 30 for LNP model) correctly identifies if information was limited or unlimited (colors).



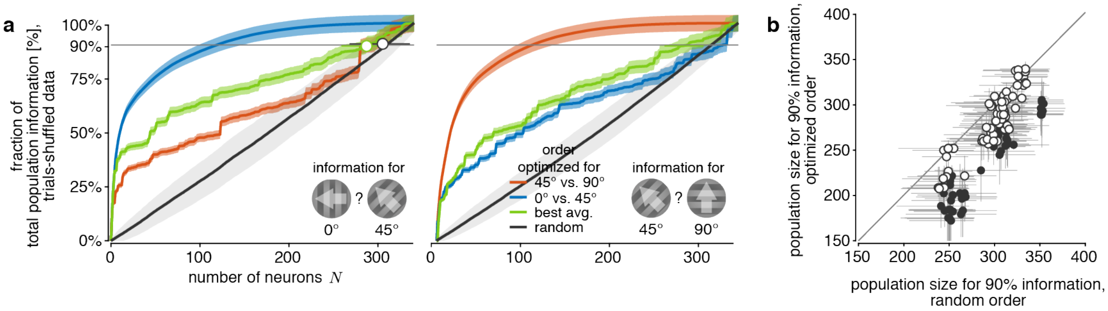
Supplementary Figure 8: Statistics of a linear fit $I_N^{-1} = \beta_0 + \beta_1 N^{-1}$ across all eight discriminations with $\delta\theta = 45°$, sessions, and mice. (a) The adjusted R^2 is close to one for all fits. (b) Both intercept, β_0, and slope, β_1, are significantly above zero for all discriminations. The plot shows these intercepts with 95% CIs, which are obscured by the dots.



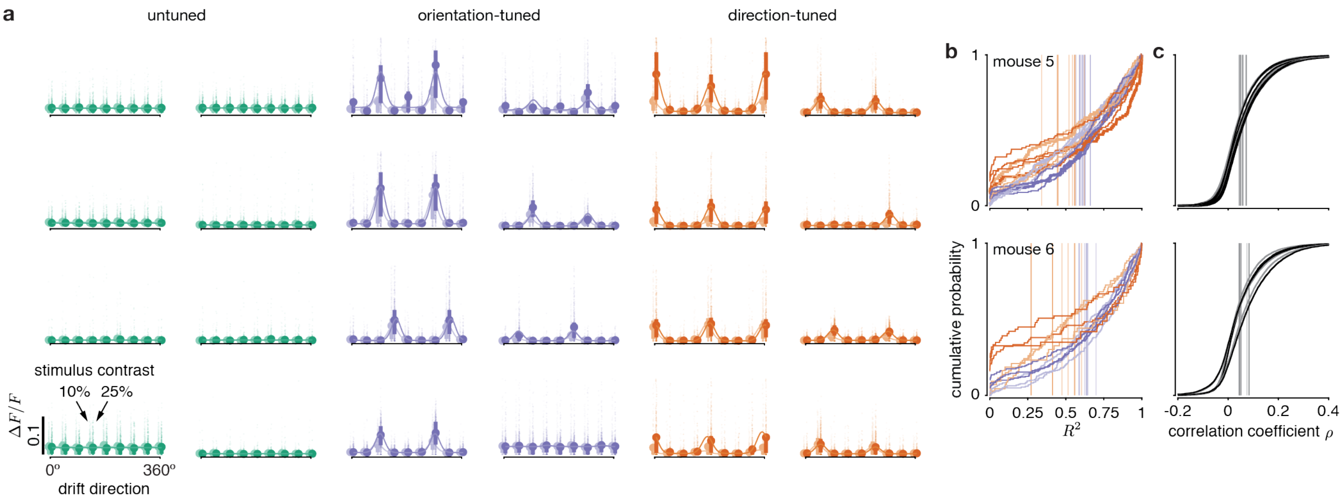
Supplementary Figure 9: Model comparison of per-discrimination fits vs. pooled fits across multiple discriminations. The figure shows for each session (individual sessions connected by grey lines; horizontally jittered for clarity) the WAIC difference of fitting the information scaling of individual discriminations (*indv*) vs. fitting all of these discriminations simultaneously (*pooled*). The mostly positive WAIC differences, preferring pooled fits, confirm that the information scaling across different discriminations with the same drift direction difference $\delta\theta$ were exceedingly similar. The tested discriminations were $45°$ vs. $90°$, $135°$ vs. $180°$, $225°$ vs. $270°$, and $315°$ vs. $0°$ ($\delta\theta = 45°$); $45°$ vs. $135°$, $90°$ vs. $180°$, $225°$ vs. $315°$, and $270°$ vs. $0°$ ($\delta\theta = 90°$); and $45°$ vs. $180°$, $90°$ vs. $315°$, and $225°$ vs. $0°$ ($\delta\theta = 135°$). The WAIC differences for $\delta\theta = 315°$ had overall smaller magnitudes, as they pooled across three rather than four discriminations.



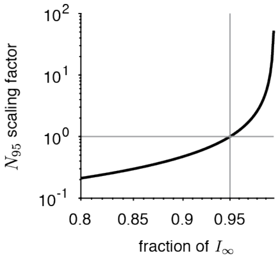
Supplementary Figure 10: Linear Fisher information is expected to drop with increasing $\delta\theta$. (a) Generalized linear Fisher information measures how easy it is to discriminate two stimuli from the population responses they evoke. This discriminabiliy is measured by the performance of a linear discriminator, normalized by the stimulus difference (δs or $\delta\theta$). For population responses (dots = mean population activity for one stimulus, shaded areas = 1SD of the noise covariance; $\delta \mathbf{f}_i$ = difference in mean population activity for different δs_i / $\delta\theta_i$) whose mean response changes linearly with the stimulus s, this information remains unchanged when δs changes (top; δs_1 vs. δs_2). Population activity that encodes a circular stimulus θ is bound to violate this linearity, and its associated linearly decodable information drops with an increase in $\delta\theta$ (bottom; $\delta\theta_1$ vs. $\delta\theta_2$). This occurs also if a non-linear decoder that accounts for the circularity of θ would recover the same information, irrespective of $\delta\theta$, and is not a bug of the linear decoder, which nonetheless correctly identifies all linearly decodable information (that drops with $\delta\theta$). (b) We demonstrate this effect by simulating V1 population in response to oriented Gabor pattern, and estimate the information encoded about their orientation. We show how information grows with population sizes for stimulus pairs with different $\delta\theta$ (colors; mean ± 1SD across different orders with which neurons are added to the population). (c) The information at $N = 1000$, which we use as a proxy for I_∞, drops with $\delta\theta$, for the reason illustrated for the rotational code in (a). Details of the simulations to generate (b) and (c) are described in Sec. 4.2. The simulations quantify information about oriented Gabor pattern rather than the drift direction of drifting gratings, and so should only be qualitatively compared to the data in the main text.



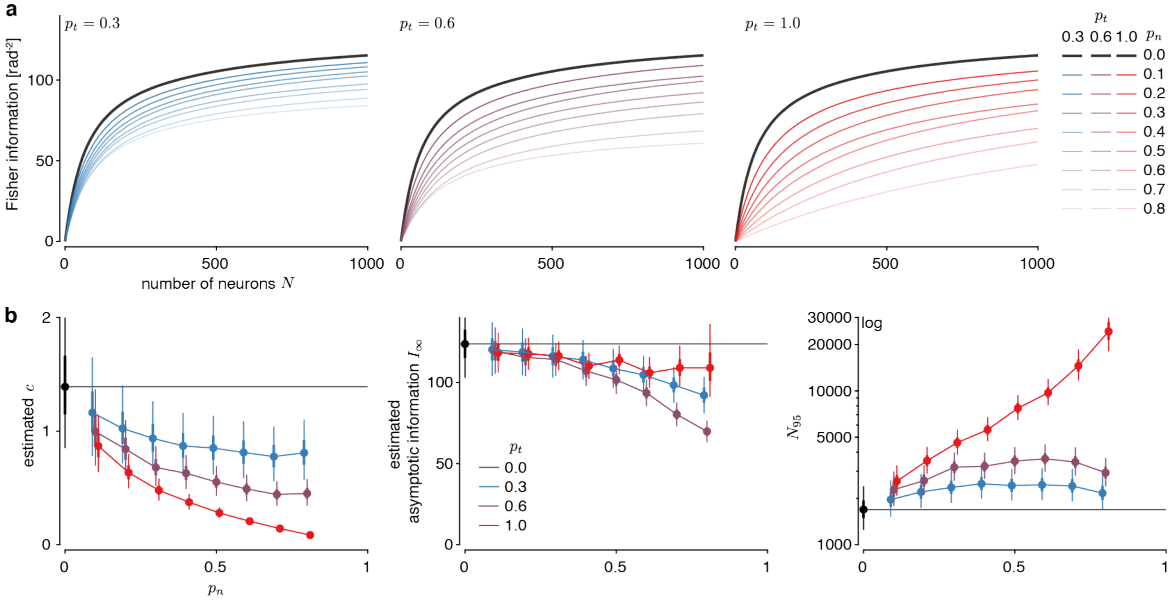
Supplementary Figure 11: Same as Supplementary Figure 5, but for trial-shuffled data. As in Supplementary Figure 5, we asked if a subpopulation appears to encode a disproportionate amount of information across all stimulus drift directions. In contrast to Supplementary Figure 5, we here removed the impact of noise correlations by, for each neuron and drift direction, randomly shuffling the trial identity. (a) Both panels show that information increase in the recorded population depends on the order with which neurons are added to the population (colors). The panels differ in the considered drift direction discrimination (left: $0°$ vs. $45°$; right: $45°$ vs $90°$). The neuron order was optimized by incrementally adding the neuron that resulted in the largest information increase for a $0°$ vs. $45°$ (blue) or $45°$ vs $90°$ (orange) drift direction discrimination, or largest average increase across all discriminations with $\delta\theta = 45°$ (green). The optimal ordering for the $0°$ vs. $45°$ was also applied to the $45°$ vs $90°$ discrimination (blue line in right panel) and vice versa (orange line in left panel). The average information increase across random orders (black) is shown as baseline reference. Shaded error regions illustrate the uncertainty (mean ± 1SD) due to limited numbers of trials (all curves), and variability across random orderings (black only). The black and green open circle (bootstrapped median ± 95% CI) show the population sizes required to capture 90% of the information in the recorded population for the associated orderings. (b) Plotting population sizes required to capture 90% of the information in the recorded population (bootstrapped median ± 95% CI) for random ordering vs. orderings optimized to maximize average information across all discriminations revealed a significant difference between the two orderings for some datasets (filled dots). Each dot reflects one discrimination for one session. The difference in population sizes was also significant across all datasets (t-test, $t_{63} = 9.541$, two-sided $p = 6.1 \times 10^{-14}$, across non-overlapping $\delta\theta = 45°$ discriminations).



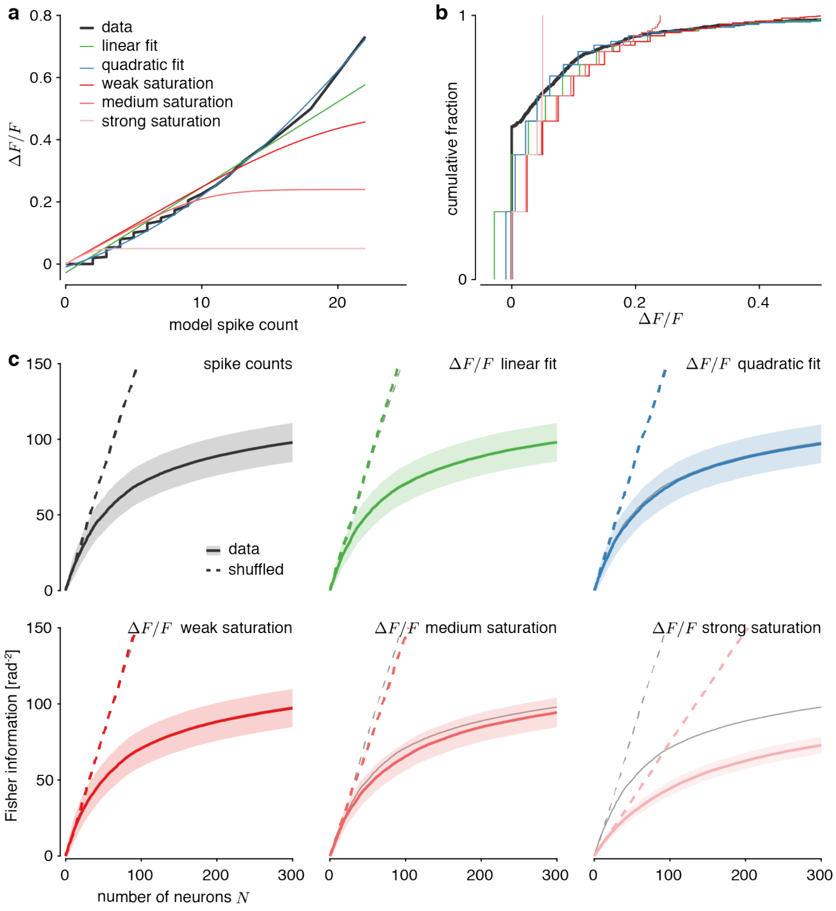
Supplementary Figure 12: Example tuning curves, tuning curve R^2's, and pair-wise correlations, for the 10% and 25% contrast trials of mice 5 and 6. We determined the tuning type (untuned, orientation-tuned, or direction-tuned) of each neuron by fitting the 25% contrast trials only, and chose the best-fit tuning curve function $f_{25}(\theta)$ (functional form depends on tuning type) as described in Methods. We then jointly fit the 10% and 25% contrast data by using the previously determined $f_{25}(\theta)$ function to fit the 25% contrast trials, and $f_{10}(\theta) = a + bf_{25}(\theta)$ to fit the 10% contrast trials. To do so, we jointly adjusted the parameters of the f_{25} function, as well as a and b. Except for untuned neurons, this resulted in tuning curve fits with fewer parameters than if we would have fitted the tuning curves for each contrast level separately. Bayesian model comparison that accounted for the different numbers of parameters revealed that, for most neurons, this joint fit across both contrast levels explained the data better than separate fits for each contrast level ($\text{BIC}_{\text{joint}} < \text{BIC}_{\text{separate}}$; mouse 5: direction-tuned 76.54% (310 of 405 neurons), orientation-tuned 94.21% (683 of 725 neurons); mouse 6: direction-tuned 70.35% (140 of 199 neurons), orientation-tuned 92.41% (597 of 646 neurons)). (a) Eight examples of untuned, orientation-tuned, and direction-tuned neurons. The pale, small dots show responses in individual trials. The large dots show mean responses for each drift direction, and the solid, vertical lines connect the 25th and 75th percentile. The pale lines shows the fitted tuning curves. Each panel shows data for both 10% and 25% contrast trials. Data, raw, and fitted tuning curves are darker for 25% contrast trials and slightly shifted to the right. Plots are truncated at $\Delta F/F = 0.2$. (b) Cumulative distributions of coefficients of variation R^2 for different mice (rows) and sessions (line) for orientation-tuned (purple) and direction-tuned (orange) neurons for 25% (dark) and 10% (bright) contrast trials. The R^2 values are computed separately for each contrast level, even though the tuning curves were fit jointly across the two contrast levels. (c) The cumulative distribution of pairwise noise correlations for pairs of neurons for 25% contrast (dark) and 10% contrast (bright trials), shown separately for each session. (a) used data from one session of mouse 5 whose corresponding R^2 values are shown in bold in (b) and (c). All data at 0° is replicated at 360° to show the fitted tuning curve across all possible drift directions. The pale vertical lines in (b) and (c) show the average R^2 and correlation coefficients for each sessions, tuning type, and contrast level. The observed mean correlations were comparable to those found in previous studies [17]. Note that fitted tuning curves and average pair-wise correlations were not used to estimate information, and are provided for reference only.



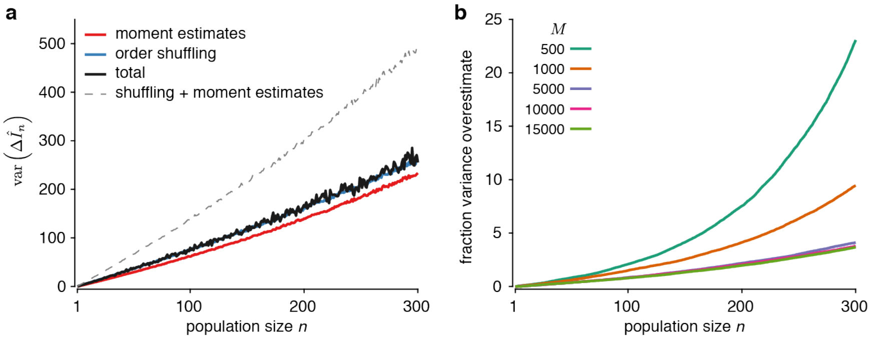
Supplementary Figure 13: The scaling of the estimated population size with the fraction of asymptotic information. Let N_a denote the population size required to encode a% of the total asymptotic information, I_∞. Changing a results in a simple re-scaling of N_a. This figures illustrates this re-scaling for different a, using N_{95} as a base measure. For example, if we would be interested in N_{90} instead of N_{95}, we would read off the scaling factor for 90%, and would re-scale the reported N_{95} to get estimates for N_{90}.



Supplementary Figure 14: Effects of modeled eye movement on information scaling. We assessed the effects of eye movements on information scaling using a simple eye movement model (see Sec. 2.4 for details). In this model, in a fraction of trials p_t the activity of a fraction of neurons p_n was set to zero. (a) We simulated population activity by adding a Poisson step to the model described in Sec. 4.2 to simulate spike counts in a 500ms window in a population of $N = 1000$ neurons in response to oriented Gabor pattern, and estimated bias-corrected linear Fisher information for different population sizes, different p_t's (panels/colors), and different p_n's (color shading). As can be seen, information grows more slowly with number of neurons N with increasing p_t and increasing p_n. (b) We fitted our information scaling model to the information scaling curves in (a) to estimate non-asymptotic scaling c (left) and asymptotic information I_∞ (center). We used these estimates to, in turn, estimate the number of neurons N_{95} required to encode 95% of asymptotic information (right). The posterior estimates are shown as in Supplementary Figure 4c in the main text. As predicted by our theory (see Sec. 2.4), asymptotic information scales with $p_t(1-p_t)$: it hardly drops if $p_t = 1$, and drops most strongly for $p_t = 0.6$. This also confirms that $\boldsymbol{f}_{ori}$ is most likely not perfectly aligned to $\boldsymbol{f}'_{ori}$ in our simulations. The estimated c, in contrast, drops monotonically with both an increase in p_t and in p_n. In combination, this yields an overestimation of N_{95} that grows with both p_t and p_n.



Supplementary Figure 15: A non-linear mapping between spike counts and $\Delta F/F$ signal does not qualitatively impact our results. We added a Poisson step to the model described in Sec. 4.2 to simulate spike counts in a 500ms window in a population of $N = 300$ neurons in response to oriented Gabor pattern. We then used linear and non-linear functions that map these spike counts to $\Delta F/F$ signals that were in turn used to estimate how information scales with population size. (a) The different utilized functions for mapping per-neuron spike counts to $\Delta F/F$ signals. The black line connects the percentiles (from the 1st to the 99th percentile) of the distribution of simulated spike counts to those of the distribution of $\Delta F/F$ responses of the data of session 1 of mouse 1. The near-linear relationship indicates that a linear remapping of these spike counts will well-replicate the observed $\Delta F/F$ distribution. We fitted this relationship with both a linear (green) and a quadratic function (blue). We furthermore tested various saturating functions (different shades of red, weak: $0.5\tanh\left((0.5r)^2\right)^{1/2}$, medium: $0.24\tanh\left((0.1r)^2\right)^{1/2}$, strong: $0.05\tanh\left((0.45)^2\right)^{1/2}$, where r is the spike count) to simulate the scenario in which the $\Delta F/F$ response saturates for higher spike counts. (b) The cumulative distribution function of the $\Delta F/F$ distributions of the data of session 1 of mouse 1 (black), the linear (green) and quadratic (blue) model, and the different saturating models (shades of red). In particular the strongly saturating model results in significant deviations from the data. (c) Information scaling computed from the $\Delta F/F$ signal (except for the black & grey lines, which are based on spike counts) for the different mappings between spike counts and $\Delta F/F$. The dashed lines show results when trial-shuffling the spike count data before mapping it to $\Delta F/F$ signals to destroys the spike count noise correlations (see Supplementary Figure 3). Any eventual difference between Fisher information computed from spike counts (black) and the linear model (green) are due to numerical precision, as invertible linear transformations, as used here, do not change the Fisher information. Most importantly, trial-shuffled spike count data yields linear information scaling even after non-linear mappings, as these mappings do not introduce new noise correlations. Similarly, the information scaling of non-shuffled data saturates even after perturbing the spike counts with a non-linear mapping, as this mapping does not remove the noise correlations. While strong non-linear mappings might lower our information estimates, they do not impact our finding that information saturates.



Supplementary Figure 16: The variance and covariance of Fisher information scaling. We simulated virtual populations of different sizes M as described in Sec. 4.1, yielding $\mathbf{f}'$ and $\boldsymbol{\Sigma}$ for each M. (a) To demonstrate that the variance in the Fisher information increase estimate due to shuffling well-approximates the combined variance due to population subsampling and due to estimating the moments from a finite number of trials, we generated one population with $M = 10,000$ neurons. We in turn drew 100 empirical moments, $\gamma \sim \mathcal{N}\left(\mathbf{f}', 2\boldsymbol{\Sigma}/(T\delta\theta)^2\right)$ and $\boldsymbol{\Omega} \sim \mathcal{W}\left(\boldsymbol{\Sigma}/(2T-2), 2T-1\right)$, corresponding to estimating these moments from $T = 1,000$ trials each for two drift directions separated by $\delta\theta = 45^\circ$. We additionally subsampled $N = 300$ neurons of the full population ten times, resulting in ten $i_{1:N}$ neuron indices, and, for each $i_{1:N}$, containing a fixed set of neurons, shuffled their order 100 times, resulting in 100 $j_{1:N}$ per $i_{1:N}$. For the empirical moments, we computed the Fisher information increase for each subsampled, shuffled population $j_{1:N}$, resulting in 10^6 estimates for each population size $n \in 1, \dots N$. The figure shows the variance due to shuffling only (blue, averaged over different subsamples and empirical moments), and due to empirical moments only (red, averaged over different subsamples and shuffles). As comparison, we computed the total variance of the same estimate across 100 subsampled populations with $N = 300$ neurons for each set of empirical moment (black; variance across 10^4 estimates), which is the variance we aim to estimate. As the plot shows, the variance due to shuffling well-approximates this total variance. A naïve sum of the variance due to empirical moments and shuffling (grey dashed) would over-estimate the total variance. (b) To estimate the degree by which the variance of the Fisher information increase, $\mathrm{var}\left(\Delta\hat{I}_n\right)$, is overestimated when ignoring the negative correlations across different $\Delta\hat{I}_n$'s, we generate populations of different sizes, M, and their associated moments. For each population, we then estimated the covariance $\mathrm{cov}\left(\Delta\hat{I}_n, \Delta\hat{I}_m\right)$ across 1,000 different subsamples $i_{1:N}$ of populations of $N = 300$ neurons. In turn, we estimated the Fisher information variance once when taking into account this covariance, $\mathrm{var}\left(\hat{I}_n\right) = \sum_{j=1}^{n}\left(\mathrm{var}\left(\Delta\hat{I}_j\right) + 2\sum_{k=1}^{j-1}\mathrm{cov}\left(\Delta\hat{I}_k, \Delta\hat{I}_j\right)\right)$, and once when not doing so, $\tilde{\mathrm{var}}\left(\hat{I}_n\right) = \sum_{j=1}^{n}\mathrm{var}\left(\Delta\hat{I}_j\right)$. The plot shows the resulting fraction $\left(\tilde{\mathrm{var}}\left(\Delta\hat{I}_n\right) - \mathrm{var}\left(\Delta\hat{I}_n\right)\right)/\mathrm{var}\left(\Delta\hat{I}_n\right)$ for different n and M as an average across ten different generated populations, and shows that the variance overestimate becomes smaller for larger populations.

Supplementary References

[1] Rubén Moreno-Bote, Jeffrey Beck, Ingmar Kanitscheider, Xaq Pitkow, Peter Latham, and Alexandre Pouget. Information-limiting correlations. *Nature Neurosience*, 17(10):1410–1417, 2014. doi: 10.1038/nn.3807.

[2] Deep Ganguli and Eero P. Simoncelli. Efficient sensory encoding and bayesian inference with heterogeneous neural populations. *Neural Computation*, 26:2103–2134, 2014. doi: 10.1162/NECO_a_00638.

[3] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications and Signal Processing. Whiley-Interscience, 2nd edition, 2006.

[4] Wei Ji Ma, Jeffrey M Beck, Peter E Latham, and Alexandre Pouget. Bayesian inference with probabilistic population codes. *Nature Neuroscience*, 9:1432–1438, 2006. doi: 10.1038/nn1790.

[5] James O. Berger. *Statistical decision theory and Bayesian analysis*. Springer Series in Statistics. Springer, 2nd edition, 1993.

[6] David Marvin Green and John A. Swets. *Signal detection theory and psychophysics*. Peninsula Pub, 1989.

[7] Bruno B. Averbeck and Daeyeol Lee. Effects of noise correlations on informatino encoding and decoding. *Journal of Neurophysiology*, 95:3633–3644, 2006. doi: 10.1152/jn.00919.2005.

[8] Yuzhi Chen, Wilson S. Geisler, and Eyal Seidemann. Optimal decoding of correlated neural population responses in the primate visual cortex. *Nature Neuroscience*, 9(11):1412–1420, 2006. doi: 10.1038/nn1792.

[9] Ramon Nogueira, Nicole E. Peltier, Akiyuki Anzai, Gregory C. DeAngelis, Julio Martínez-Trujillo, and Rubén Moreno-Bote. The effects of population tuning and trial-by-trial variability on information encoding and behavior. *The Journal of Neuroscience*, 40(5):1066–1083, 2020. doi: 10.1523/JNEUROSCI.0859-19.2019.

[10] James O. Berger. *Pattern recognition and machine learning*. Information Science and Statistics. Springer, 2011.

[11] Ingmar Kanitscheider, Ruben Coen-Cagli, Adam Kohn, and Alexandre Pouget. Measuring fisher information accurately in correlated neural populations. *PLoS Computational Biology*, 11(6):e1004218, 2015. doi: 10.1371/journal.pcbi.1004218.

[12] Richard A. Johnson and Dean W. Wichern. *Applied multivariate statistical analysis*. Pearson, 5th edition, 2007.

[13] Robb J. Muirhead. *Aspects of multivariate statistical theory*. Wiley, 2nd edition, 2005.

[14] Kazuhiro Suzuki, Dongvu Tonien, Kaoru Kurosawa, and Koji Toyota. Birthday paradox for multi-collisions. In Min Surp Rhee and Byoungcheon Lee, editors, *Information Security and Cryptology — ICISC 2006*, pages 29–40. Springer-Verlag Berlin Heidelberg, 2006. doi: 10.1007/11927587_5.

[15] Ingmar Kanitscheider, Ruben Coen-Cagli, and Alexandre Pouget. Origin of information-limiting noise correlations. *Proceedings of the National Academy of Sciences*, 112(50):E6973–E6982, 2015. doi: 10.1073/pnas.1508738112.

[16] Ramon Nogueira, Juan M. Abolafia, Jan Drugowitsch, Emili Balaguer-Ballester, Maria Sanchez-Vives, and Rubén Moreno-Bote. Lateral orbitofrontal cortex anticipates choices and integrates prior with current information. *Nature Communications*, 8:14823, 2017. doi: 10.1038/ncomms1482.

[17] Marlene R Cohen and Adam Kohn. Measuring and interpreting neuronal correlations. *Nature Neurosience*, 14(7):811–819, 2011. doi: 10.1038/nn.2842.