Choice can be predicted from populations of bursting neurons in superficial layers of monkey V1

Veronika Koren [1,2,4,*], Ariana R. Andrei [3], Ming Hu[3,5], Valentin Dragoi [3,*], Klaus Obermayer [1,2,6,*]

1 Institute of Software Engineering and Theoretical Computer Science, Technische Universität Berlin, 10587 Berlin, Germany **2** Bernstein Center for Computational Neuroscience Berlin, Germany **3** Department of Neurobiology and Anatomy, University of Texas Medical School, Houston, Texas, 77030, US **4** Present address: Institute of Mathematics, Technische Universität Berlin, 10623 Berlin, Germany **5** Present address: Baylor College of Medicine, Houston, TX 77030, US **6** Lead author

[*] Correspondence: valentin.dragoi@uth.tmc.edu, koren@math.tu-berlin.de, klaus.obermayer@tu-berlin.de

Abstract

Primary visual cortex (V1) is absolutely necessary for normal visual processing, but whether V1 encodes upcoming behavioral decisions based on visual information is an unresolved issue, with conflicting evidence. Further, no study so far has been able to predict choice from time-resolved spiking activity in V1. Here, macaque monkeys are tested in a decision-making task with a categorical variable: match versus non-match of two consecutive stimuli. Our aim is to predict the binary choice "same" versus "different" on trials that only differ in choice, but not in stimuli. Classical decoding schemes fail to predict the choice in such a setting, however, the choice can be predicted by learning the structure of population responses in trials with correct behavior. The learned structure of population responses transfers to the representation of the choice, even when the choice is incorrect. With such generalized learning scheme, the choice can be successfully predicted from spiking activity of neural ensembles in V1 in single trials. We show that the choice signal is primarily carried by bursting neurons in the superficial layer of the cortex. Results suggest that bursting of single neurons and noise correlations between neurons with similar decoding selectivity are dynamical mechanisms that help the formation of time-resolved choice signals.

Keywords: Macaque; visual cortex; cortical column; decoding; stimulus; choice; population code; generalized learning

Introduction

In real-world scenarios, animals frequently make decisions based on the observation of their environment. In the context of such decision-making, encoding of sensory information and utilizing this information for guiding the behavior are two distinct, but connected computational problems (Panzeri et al. 2017). So far, the primary visual cortex has been shown to be a site of relatively sophisticated sensory processing, from which other brain areas read-out the sensory information in order to perform further computations, such as decision-making based on sensory cues. Corroborating this view has been the evidence of the choice probability in the middle and high visual cortical areas, as "early" as V2, but not in V1 (Nienborg and Cumming 2006). Recent study, however, has found the choice probability in V1 during a coarse orientation-discrimination task with artificial stimuli (Nienborg and Cumming 2014), arguing that, in order to observe choice probability, neurons must be topologically organized with respect to the task feature.

The main hypothesis behind the framework of the choice probability is that it measures the choice related sensory evidence in single neurons. Choice probabilities (CPs) have been extensively studied for more than three decades (Shadlen and Newsome 1996; Britten et al. 1996; Shadlen et al. 1996; Logothetis and Schall 1989; Roitman and Shadlen 2002; Nienborg and Cumming 2009; Goris et al. 2017; Seidemann and Geisler 2018; Jasper et al. 2019; Krishna et al. 2021). The hypothesis behind the CP framework is a positive correlation between the activity of a neuron, tuned to a specific stimulus feature, and the choice of the animal. Within the framework of choice probabilities, an increase in firing rate of a neuron with the preference for the vertical orientation, for example, leads to animal's decision for "vertical" more likely than for"horizontal". Vice-versa, an increase of the firing rate of a neurons with preferred orientation for "horizontal" leads to the choice "horizontal" more likely than the choice "vertical". Collecting the evidence (or CPs) across all contributing sensory neurons, neurons in a decision-making area would then activate in order to signal the choice of the animal (Shadlen and Newsome 1996; Nienborg et al. 2012). Weak but significant positive correlation between the CP and the choice of the animal has been reported in a variety of sensory areas (Shadlen and Newsome 1996; Britten et al. 1996; Logothetis and Schall 1989; Uka and DeAngelis 2004; Roitman and Shadlen 2002). Recent studies, however, have questioned the classical interpretation of CPs. Goris and colleagues (Goris et al. 2017) reported a negative correlation between activity of V1 neurons and animal's choice, since neurons fired more vigorously when the animal chose the orientation that was the opposite to their preferred orientation. This result is in contradiction with the hypothesis of CP as primarily reflecting the sensory evidence.

If the choice probability does not necessarily reflect the sensory evidence, it might simply reflect evidence about animal's choice. Recent work in brain-wide electrophysiology has questioned the division of brain areas into clearly defined sensory, motor or internal variables. The activity in the early sensory cortex, including V1, was shown to encode variables that go well beyond sensory features (Stringer et al. 2019a;b). Moreover, in the mouse brain at least, the choice is represented in multiple brain areas simultaneously (Yang et al. 2016; Steinmetz et al. 2019), and there seems to be no evidence suggesting a different picture in primate brains. Recent years have seen the revival and the development of the idea that behaviorally relevant variables are encoded by collective dynamics of neural ensembles (Gao and Ganguli 2015; Saxena and Cunningham 2019; Boerlin et al. 2013), and population codes seem to be appropriate to study neural responses in the auditory (Harris et al. 2011; Luczak et al. 2013), motor (Churchland et al. 2012), prefrontal (Rigotti et al. 2013), and visual cortex (Ni et al. 2018). Stimulus features are most likely only a part of variables that neural populations in V1 encode during a decision-making task.

To better capture the representation of the choice in V1, we here analyze data from a decision-making study where the variable of interest is a categorical, rather than a continuous sensory variable. The variable of interest in the present study is a binary stimulus class "match/non-match", where "match/non-match" refers to matching of two consecutive stimuli. Traditionally, decision-making is studied with a sensory variable that is changed in a continuous fashion, leading to a categorical choice variable with a limited number (typicaly two) of options. In such a setting, the representation of a continuous sensory variable is presumably mapped on the representation of a categorical choice variable. In the present experimental setting with a binary stimulus

class and a binary choice, we test the representation of a categorical variable in V1. Previous work on decision-making with categorical variables suggested that such variables might be represented as contextual cues through a feedback signal (Muckli et al. 2015; Zipser et al. 1996; Juan and Walsh 2003). Moreover, correlated variability in V1 can change with task instructions (Bondy et al. 2018), strongly suggesting that correlated variability in their experiment was driven by a feedback from higher cortical area(s). The stimulus class and the choice could therefore be computed downstream of V1 through a comparison circuit (Engel and Wang 2011), and be projected back to V1. This leads us to hypothesize that choice signals in V1 might be (partially) generated by a feedback rather than by feed-forward processing (Seidemann and Geisler 2018).

Previous studies investigated the representation of the choice in V1, decoding from time-averaged spiking signals and from the Local Field Potential, a proxy of the time-resolved population dynamics. In early (V1) and mid-level (V4) visual cortices, pooling of information across neurons yielded robust choice signals in V4, but weak signals in V1 Jasper et al. (2019). However, robust choice signals have been reported in V1 when decoding from the Local Field Potential, in a coarse orientation difference task with drifting gratings (Krishna et al. 2021), suggesting that the activity of larger populations in V1 does carry the information about the behavioral choice. Neural representations of behaviorally relevant variables, however, are presumably build from spike trains of neural populations and takes place in real time to contribute to animal's behavior as it unfolds. To investigate time-resolved representations of behaviorally relevant variables in the spiking signal of neural populations, we recently proposed a simple decoding protocol where spikes are weighted with population decoding weights (Koren et al. 2019). With such a decoder, we were able to decode the correct choice behavior in the present dataset, namely from the superficial cortical layers. With binary stimulus class, the representation of correct choice coincides with the representation of the stimulus class, and the question remained if the choice can also be decoded from trials that do not differ in stimulus class, but only in choice.

In this study, we decode the choice in absence of direct information about the stimulus class. We find that, interestingly, learning on trials with correct choice generalizes to the representation of incorrect choice. Generalized learning was suggested to be the mechanism for forming mixed representations in the prefrontal cortex (Bernardi et al. 2020), but highly mixed representations have been recently demonstrated also in V1 (Stringer et al. 2019a). In our case, we utilize a generalization of learning from a mixed variable *stimulus* + *choice* to the variable *choice*. We implement such generalization by training a linear decoder in the context of *stimulus* + *choice*, and then apply population decoding weights onto spike trains in single trials that only differ in choice, but not in stimulus class. We show that, contrary to previous beliefs, upcoming behavioral choice can be decoded from the spiking activity in V1 in a time-resolved population code. We show that the choice-related information is mostly present in superficial layers of the cortex and primarily carried by bursting neurons. Finally, our results suggest that bursting and correlated spiking of neurons with similar decoding selectivity as a dynamical mechanism that helps the formation of the choice signal.

Results

The stimulus class and the binary choice have overlapping representations

Two adult male macaques were trained on a delayed match-to-sample visual task. The subject visualized the target and the test stimuli, with a delay period in between (Figure 1A). The stimuli were complex naturalistic images in black and white, depicting an outdoor scene, and their identity changed from one recording session to another (Shahidi et al. 2019). The target and test stimulus were either identical (condition match) or not (condition non-match), with the only possible difference being the change in the orientation of the test stimulus. In condition non-match, the angle of rotation of the test stimulus was random (clockwise or anti-clockwise), close to psychophysical threshold (between 3 and 12 deg) and was kept fixed throughout the recording session. The task of the subject was to communicate its decision about the stimulus class ("same" or "different"; Figure 1B). The multiunit signal was captured with linear arrays, inserted perpendicularly to the cortical surface and recording the activity across the cortical depth (see methods). After spike sorting, we obtained on the order of

10 units in each recording session, and the total of 160 units in 20 recording sessions. Example spike rasters for three simultaneously recorded neurons, belonging to the superficial (supragranular or SG), middle (granular or G) and deep (infragranular or IG) cortical layer, are shown on Figure 1C. The behavioral performance was good, with an average of 73 % correct responses on non-matching stimuli (Figure 1D). Single neurons were sensitive to the non-match of stimuli (Figure 1E), with the sensitivity index of 133 out of 160 neurons being significant to the change in orientation between the target and the test stimulus (permutation tests with 1000 permutations; see methods), and with the average sensitivity index of $\langle d' \rangle = 0.6$ ($p < 0.001$, permutation test).

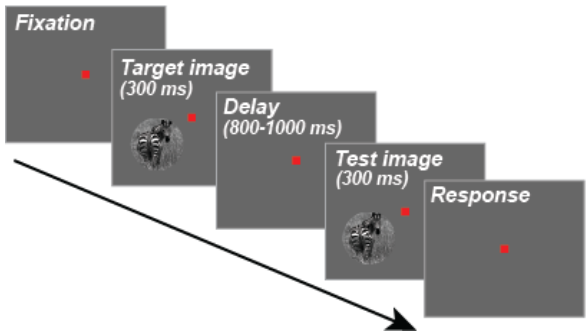
Single neuron statistics were not predictive of the variable *choice*. Firing rates of single neurons varied strongly across units, but were very similar for the choice "same" and "different" (Figure 1F). Coefficient of variation of single neurons (CV_2, see methods) varied around 1, which is the CV_2 of the Poisson process, and was strongly positively correlated across conditions (Figure 1G). Choice probabilities were both bigger and smaller than the baseline ($CP = 0.5$), indicating an increase ($CP > 0.5$) as well as a decrease ($CP < 0.5$) in the firing rate with the choice "same". After correction for multiple testing, only 4 out of 160 neurons had significant CP (permutation test with 1000 permutations; Figure 1H). The distribution of CPs was slightly imbalanced, with an excess of neurons with $CP > 0.5$ ($p = 0.0140$, permutation test).

A classical multivariate decoding paradigm predicts the variable *stimulus* + *choice*, but not the variable *choice*. We applied a linear decoder (linear Support Vector Machine or SVM; see methods) on spike counts of the population of simultaneously recorded neurons (Koren et al. 2020). Following our hypothesis that the stimulus class and the choice are represented through a feedback signal from downstream areas, we decode from the test stimulus alone, and not from the difference of activity between the target and the test. Prediction of the variable *choice* was not significantly different from chance (average prediction accuracy of 0.507; not significant, permutation test with 1000 permutations of class labels; Figure 1I, black crosses). The variable *stimulus* + *choice*, meanwhile, is predicted with better than chance accuracy (Figure 1I, orange crosses), demonstrating that patterns of spike counts do differ between the two choices when both choices are correct. Note that with the variable *choice*, one of the choices is incorrect (see Figure 1B). Since the variable *stimulus* + *choice* contains the information about the stimulus class ("match"/"non-match") along with the information about the binary choice ("same"/"different"), one is tempted to conclude that the activity of neural populations in V1 contains the information about the stimulus class, but not about the choice.

An alternative possibility that we explore in the remaining of the paper is that the variable *choice* cannot be predicted with classical decoding paradigms, because it comes with incorrect behavior (Figure 1B). When the choice is correct, we expect it to be, at least in large extent, caused by correct interpretation of sensory evidence by the brain. Causes of an incorrect choice, however, can be many, for example, fluctuations of attention (Cohen and Maunsell 2009), a failure of transmission of the sensory evidence downstream (Shahidi et al. 2019), the noise in the incoming sensory evidence (Brunton et al. 2013) or an error in the process of inference Beck et al. (2012). If different causes co-exist in the same dataset, it is unlikely that the classifier can extract patterns from the neural activity that would allow prediction of the choice. Nevertheless, if animal's choice is a latent variable that influences neural activity, we expect that neural activity in trials that lead to a particular choice might be to some degree similar, irrespectively of the stimulus. In particular, conditions "incorrect non-match" and "correct match" both lead to the animal reporting the decision "same" (Figure 1B).

Due to a common latent variable, we hypothesize that the information used in classification problems *stimulus* + *choice* and *choice* is partially overlapping. We test the overlap in the informational content between *stimulus* + *choice* and *choice* by comparing decoding weights across the two classification problems. We separately apply linear classifiers (SVMs) on the variable *stimulus* + *choice* and on the variable *choice* and estimate decoding weights (w_n^{S+C} and w_n^C, n is the neuron index) of single neurons. We find a weak but significant positive correlation between the two sets of weights, with the correlation coefficient $r(w^{S+C}, w^C) = [0.13, 0.12, 0.13]$ for length of decoding window of $K = [300, 400, 500]$ ms, respectively; Figure 1J). A positive correlation of decoding weights indicates that the two classification problems indeed have overlapping information. Another, complementary way to verify the overlap between *stimulus* + *choice* and *choice* is to consider the angle between population vectors. A population vector is a normalized vector of feature weights of the classifier, for neurons

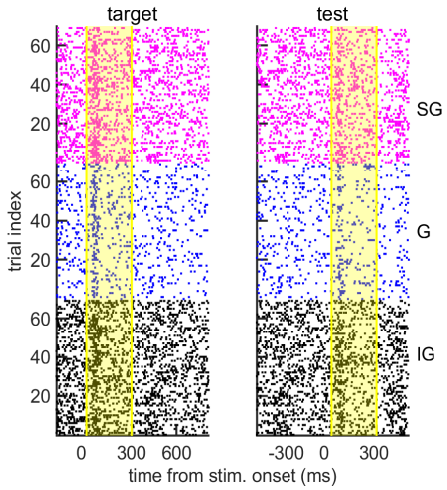
A



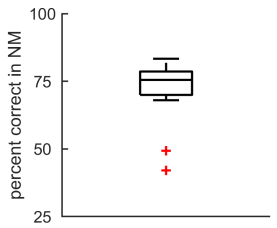
B

Condition	Stimulus	Choice	Info. content
correct non-match	non-match	"different"	stimulus + choice
correct match	match	"same"	
correct non-match	non-match	"different"	choice
incorrect non-match	non-match	"same"	

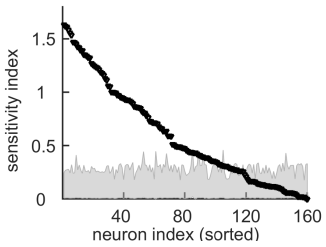
C



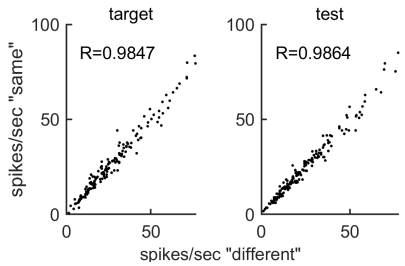
D



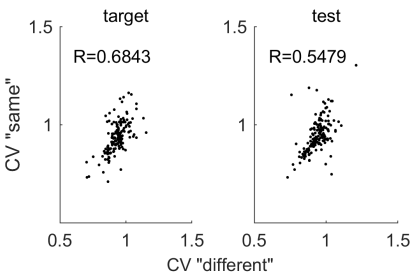
E



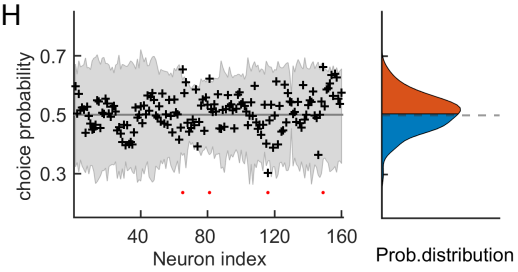
F



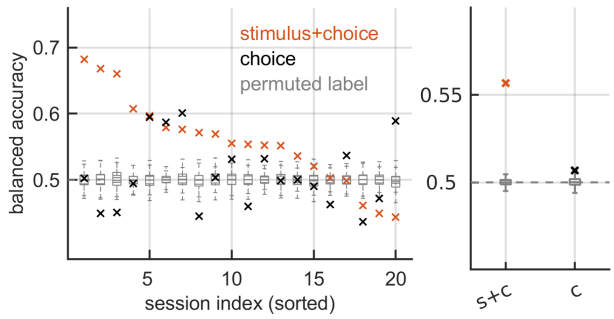
G



H



I



J

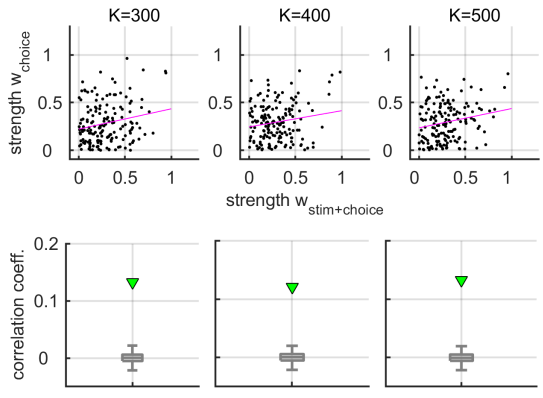


Figure 1. The structure of population responses in *stimulus + choice* and *choice* is partially overlapping. (A) Experimental paradigm. The subject visualizes the target and the test stimulus, with a delay period in between. **(B)** Informational content of two classification problems. We did not analyze condition "incorrect match" due to the small number of trials. **(C)** Spike rasters of three representative single neurons from the superficial cortical layer (SG, magenta), middle layer (G, blue) deep layer (IG, black). Neural activity has been captured on the same electrode. **(D)** Percent correct trials on non-matching stimuli. **(E)** Sensitivity index (d') of single neurons to non-match of stimuli, sorted from strongest to weakest. The gray area shows 95 % of results with permuted labels for "target" and "test".

Figure 1. **(F)** Scatter plot of the mean firing rate for the choice "different" (x-axis) and "same" (y-axis). R is the Pearson correlation coefficient. **(G)** Same as in **(F)**, but showing the coefficient of variation with time-dependent firing rates (CV_2). **(H)** Left: Choice probability of single neurons (black), and the distribution of choice probabilities for models with permuted class labels (gray). Right: Distribution of regular (non-permuted) choice probabilities from the left. **(I)** Prediction accuracy of a linear classifier for the classification problem on *stimulus* + *choice* (orange) and *choice* (black). Left plot shows results in recording sessions and right plot shows session-averaged results. Gray boxplots show the distribution of results of models with permuted class labels. Parameters for E-H: Time window of $K = [0, 400]$ milliseconds with respect to the onset of the test stimulus. **(J)** Top: Scatter plot of the absolute value of weights for the classification problem on *stimulus* + *choice* and on *choice*, utilizing time windows $K = [300, 400, 500]$ milliseconds (ms). Weights are collected across recording sessions. Bottom: Correlation coefficient (green) for the data on the top. Gray box-plots show the distribution of results when weights are computed with randomly permuted class labels.

from the same recording session (see methods). We compute population vectors for each classification problem, $\tilde{\mathbf{w}}^{S+C}$ and $\tilde{\mathbf{w}}^{C}$. The angle between population vectors (Figure 2A) tells about the similarity of the population vectors and, consequently, the similarity of separating boundaries in the two classification problems. If population vectors were independent, the angle between them would be orthogonal, $\alpha^{\text{independent}} = \pi/2$, while deviation from orthogonality means that the two vectors are partially overlapping. We find that the average angle between the two population vectors deviates significantly from orthogonality (Figure 2B). The deviation from orthogonality is robust to the length of the time window that we use for decoding ($\alpha = [80.2,\ 80.8,\ 80.0]$ deg for time windows $K = [300,\ 400,\ 500]$ ms, respectively) and is significant in all cases ($p < 0.001$, in all cases, permutation tests with 1000 permutations; see methods).

Behavioral choice can be predicted from spiking activity through generalized learning of neural representations

After examining the structure of population spike counts and noticing the similarity between classification problems on *stimulus* + *choice* and on *choice*, we exploit this similarity to decode the variable *choice* as a time-resolved and low-dimensional signal. Our decoding procedure consists of the learning and the validation step, introducing several modifications to the classical decoding scheme in the validation step. In the learning phase, the linear classifier (linear SVM) receives samples of population spike counts, where one sample is a vector of spike counts of simultaneously recorded neurons in a single trial (see methods). The classifier finds a linear separation boundary in the N-dimensional space of neural activities that is optimal for separating conditions "correct match" and "correct non-match" (variable *stimulus* + *choice*). From the classification model, we compute decoding weights (one number per neuron). Within a classical validation scheme, learned patterns of spike counts would then be tested on a hold-out set of trials, using the same input statistics and the same categories as the ones used for learning (Figure 2D, "regular validation scheme"). To decode the variable *choice*, we deviate from the standard decoding procedure in two ways. First, we train the model on spike counts, while we validate it on spike trains, due to the separation of time scales justified below. Second, the learning and the validation set also differ in categories that they decode, since the training utilizes the variable *stimulus* + *choice* while the test set is applied on the variable *choice* (Figure 2D, "transfer of learning"). The result of our decoding paradigm is a low-dimensional choice signal, a quantity proportional to the synaptic current at a hypothetical read-out neuron.

To decode from spike trains, we use a method that we proposed previously (Koren et al. 2019), briefly summarized in the following. There are two elements to our decoding protocol, learning of decoding weights and the computation of the choice signal, and we assume that the time scales of these two processes can be separated. Cortical neurons were shown to change their firing rates on the time scale of a couple of milliseconds (Shadlen and Newsome 1998). We therefore assume that the computation of the choice takes place on fast timescale

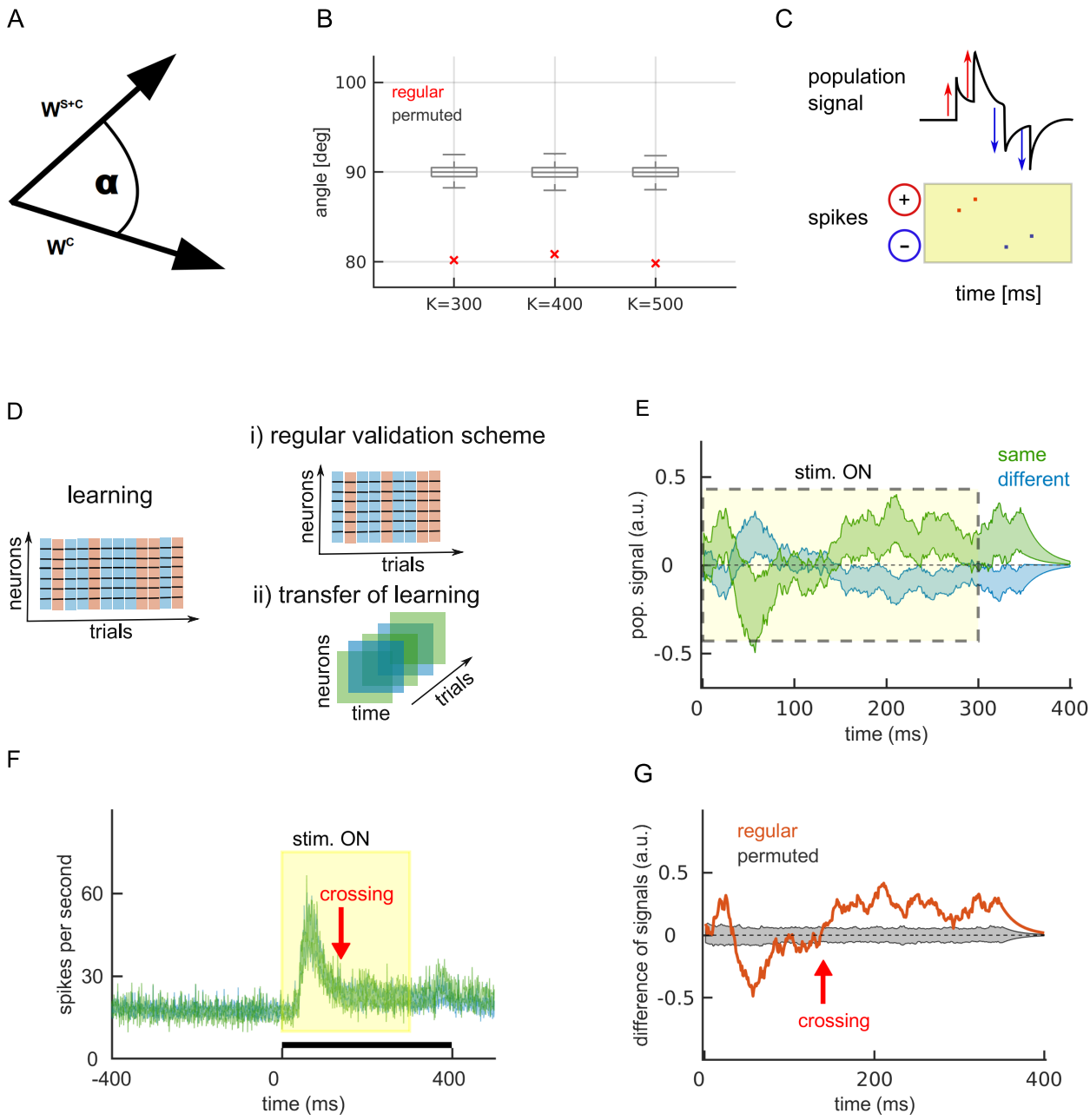


Figure 2. Behavioral choice can be discriminated with generalized learning of neural representation. **(A)** Schema of the angle between the population vectors in the context of *stimulus + choice* and *choice*. Population vectors are computed in each recording session. **(B)** Average angle between population vectors (red cross) and the distribution of same results for models with random weights (gray box-plot), for the length of the time window $K = 300$ (left), $K = 400$ (middle) and $K = 500$ (right). Parameter: $N_{perm} = 1000$. **(C)** Schema of decoding by weighting spikes. Neurons with positive weights push the signal up and neurons with negative weights pull the signal down. **(D)** Learning scheme. In the learning phase, the classifier learns from samples of parallel spike trains (left). Following the regular validation scheme, the same type of data is used for validating the predictive power of the model on a hold-out set (top right). Here, we test the generalization of learning (bottom right), where the features of the learned classification model are applied on spike trains and on a different classification problem. **(E)** The choice signal for the choice "same" (green) and "different" (blue). We show the mean ± SEM for the variability across sessions. Parameter: $\lambda^{-1} = 20$. **(F)** Population PSTHs for choices "same" (green) and "different" (blue), with the mean ± SEM. The yellow window indicates the presence of the stimulus and the black bar indicates the time window used for analysis. The red arrow marks the time of the zero crossing of the choice signal in **E**. **(G)** Session-averaged difference of choice signals for the regular model (red) and the model with permuted class labels (gray). Parameters: $\lambda^{-1} = 20$ ms, $N_{perm} = 1000$.

and compute the choice signal without temporal averaging, keeping the time resolution of the recorded spiking signal ($\Delta t = 1$ ms). The choice signal is computed as a linear sum of spikes, convolved with an exponential filter (Figure 2C; see methods) to model causal filtering of synaptic inputs at the synapse of the read-out neuron. Summing of the activity across neurons is justified by the relevance of pooling of neuronal signals for perception (Andrei et al. 2019). Decoding weights, on the other hand, model a quantity proportional to synaptic weights, that presumably change during the training phase of the experiment due to synaptic plasticity. In the brain, synaptic weights change on a time scale of hours or days Abbott and Nelson (2000), a time scale that is much slower than the time scale of the population signal. This allows us to use a fixed set of population decoding weights during the computation of the choice signal. Fixing decoding weights is well justified at least in the primary visual cortex, where decoding weights on the same dataset were shown to be similar between the first and the second half of the trial (see Koren et al. 2020, Figure 3).

With such generalized decoding scheme, the choice "same" vs. "different" can be successfully discriminated during the majority of the trial (Figure 2E). Our results are cross-validated with 100 cross-validations and the learning and validation set are non-overlapping (see methods). Since the variable *choice* does not per se contain the information about the stimulus class (both conditions in the variable *choice* are with non-matching stimuli), we argue that the decoded signal must truly reflect the information about the choice of the animal. Learning from the correct behavior is necessary for a successful prediction. When the learning and the validation step both utilize the variable *choice* (instead of learning on $stimulus + choice$), resulting signals are barely different from zero and the choice cannot be predicted (Supplementary Figure S1). We conclude that prediction of the choice from the spiking activity in V1 is possible when weights are learned on trials with correct choice behavior.

Weighting of spikes is essential for successful discrimination. The population peristimulus time histograms (PSTH) can be seen as time-resolved signals with equal weighting across all neurons. Population PSTHs corresponding to the choice "same" and "different" are highly overlapping (Figure 2F) and do not allow discrimination. To asses the significance of choice signals, we compute the session-averaged difference of signals ($\Delta x(t) = x^{\text{same}}(t) - x^{\text{different}}(t)$) and compare the difference with the same result computed on models with randomly permuted class labels (see methods). Choice signals are significantly different during the majority of the trial (Figure 2G) and depend only weakly on the time scale of the convolution, with longer time scales resulting in a smoother signal (Figure 3A). The discrimination of the choice is also robust to the change in the length of the time window that we use for decoding, as the choice can also be discriminated with time windows of [0,300] (Supplementary Figure S1) and [0,500] ms with respect to the onset of the test stimulus (Figure 3B). As a check of our protocol, we note that during the target time window, the information that is necessary for discrimination of matching from non-matching stimuli is not yet available to the visual system. Consistently, prediction of the choice during the visualization of target stimulus is at chance (Supplementary Figure S2).

The population signal reflects an internal signal about the choice of the animal. At 140 ms after the onset of the test stimulus, the read-out of the choice switches sides (Figure 2E), resulting in crossing of the baseline of the difference signal (Figure 2G, red arrow). We reason that the switch in decoded choice might reflect the change in representation from the choice "different" to the choice "same" (i.e., from correct to incorrect representation) within the trial. The timing of 140 ms after the onset of the stimulus coincides with the end of transient firing at the stimulus onset (Figure 2F, red arrow), suggesting the influence of top-down or lateral inputs, as opposed to bottom-up inputs, for the choice signal. We also notice that after the "switch", the population signal is roughly stable and persists beyond the offset of the test stimulus (Figure 3B). The maximum deviation of the difference signal from the baseline occurs 105 ms after the offset of the test stimulus. Therefore, the choice can be read-out not only during the presentation of the test stimulus, but also after the offset, suggesting that the decoded signal reflects an internal variable rather than a purely feed-forward sensory variable.

The sign of the weight gives necessary information for discrimination of the choice (Figure 3C, left), while the amplitude of the weight is not required (Figure 3C, right). In spite of its simplicity, our read-out model has several components that may or may not be necessarily required for discrimination. In order to test the necessity of each component, we remove the information of a specific component and test the effect of such a perturbation on discrimination. In our decoding procedure, we estimated decoding weights with the linear

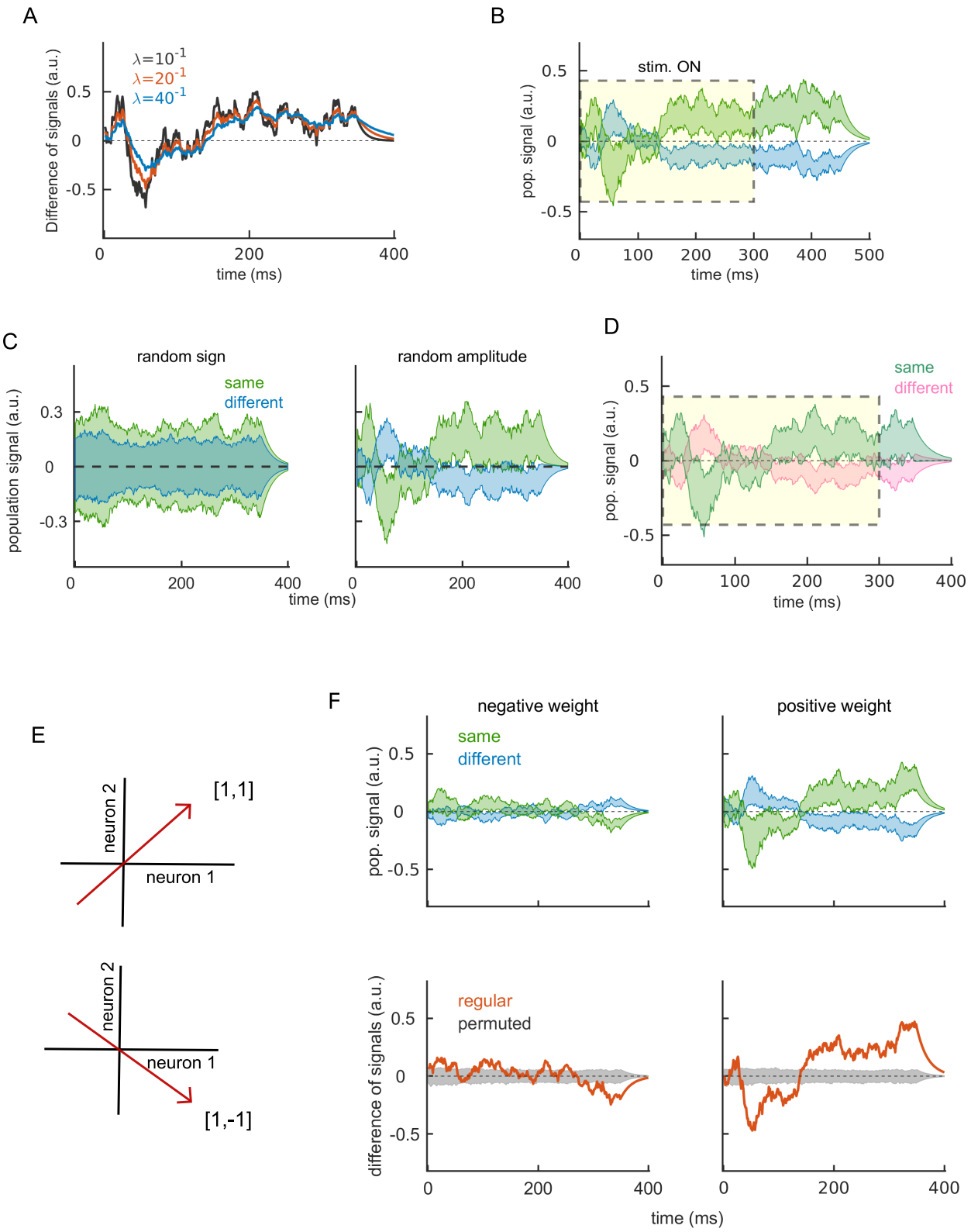


Figure 3. Correct sign of decoding weights is necessary for predicting the choice. (A) Difference of choice signals for three values of the parameter λ. **(B)** The choice signal computed with the time window of length $K = 500$ ms. The plot shows the mean $\pm$ SEM across recording sessions for the choice "same" (green) and "different" (blue). **(C)** The distribution of choice signals with random sign of weights (left) and random amplitude of weights (right) in conditions "same" (green) and "different" (blue). **(D)** The choice signal where decoding weights are computed as centered and normalized area under the ROC curve.

Figure 3. (E) Effect of the sign of the weight on the orientation of the separation boundary (schematic). In a toy model with 2 neurons, the upper schema shows the case where the weights of both neurons are positive, and the lower schema illustrates the case with one positive and one negative weight. **(F)** Top: Choice signal for neurons with negative (left) and positive weight (right). Bottom: Session-averaged difference of signals for the true model (red) and for models with permuted class labels (gray area). Parameters: $\lambda^{-1} = 20$ ms, $N_{perm} = 1000$.

Support Vector Machine, a linear decoder with optimal generalization performance (Belousov et al. 2002). As only the sign, but not the amplitude of decoding weights is necessary for the read-out, we reason that our result might not rely on the particular choice of the decoder we used. To test this, we replace population decoding weights (extracted from the linear SVM) with the weights that utilize the area under the ROC curve of single neurons (see methods). We decode a surprisingly similar choice signal (Figure 3D), showing that the specific choice of the decoder is not essential, since even an univariate method for estimating decoding weights is sufficient to discriminate the choice.

The sign of the weights is the essential part of the decoding model, since neurons with the opposite sign of the weight rotate the separation boundary of the classifier in opposite directions (Figure 3E). Neurons with positive weights tend to increase the firing rate for the choice "same" and neurons with negative weight tend to increase the firing rate for the choice "different". However, due to correlations across neurons and the sparsity constraint on the decoder (see methods), the population decoding weights are only partially similar to patterns of firing rates of single neurons. The latter are captured by the Area under the ROC curve (see methods), and we observe only a modest positive correlation between decoding weights and the Area under the ROC curve ($r(w_n, w_n^{\mathrm{ROC}}) = 0.14$). Population decoding weights rather represent the weightings of spike counts that maximize the prediction accuracy of the classifier (Koren 2021).

Neurons with positive weights convey more information about the choice than neurons with negative weights. We separate the population of simultaneously recorded neurons with respect to the sign of the weight into neurons with positive weights ($w_n > 0$) and negative weights ($w_n < 0$). As we compute the choice signal independently with each subnetwork (see methods), we find that neurons with positive weights discriminate the choice much better than neurons with negative weights (Figure 3F). The number of neurons with positive weights ($N^+ = 84$) is slightly higher than the number of neurons with negative weights ($N^- = 76$), however, this in unlikely to be important for the choice signal. As we compute decoding weights on the classification problem *stimulus* + *choice* with all neurons, the distribution of weights is centered around zero ($p = 0.235$ for K=400, $p = 0.0842$ for K=500, t-test testing that the mean of the distribution is different from 0; Figure 4A). Rather, we find that weights of *plus* neurons are more informative for classification than the weights of *minus* neurons. As we classify the variable *stimulus* + *choice* from parallel spike counts utilizing the information from only the *plus* or the *minus* subnetwork (see methods), we find that the predictive power of *plus* neurons is stronger compared to *minus* neurons ($p < 10^{-17}$, two-tailed Wilcoxon signed-rank test; Figure 4B).

The choice signals of *plus* and *minus* subnetworks are anticorrelated. Measuring the cross-correlation function of choice signals of *plus* and *minus* subnetworks in single trials (see methods), we find that choice signals of are negatively correlated (Figure 4C left; Supplementary table S4.1). This effect is significant and remains consistent across several time constants of the convolution (Figure 4C right). Negative correlation between *plus* and *minus* subnetworks is a signature of competitive dynamics and possibly recurrent inhibition between the two subpopulations (Engel and Wang 2011).

Noise correlations help the accumulation of the choice signal in a time-resolved population code

Spike timing on intermediate time scales carries the information about the choice. As we randomly permute the spike timing across the entire time window of interest (400 ms), we decode highly overlapping signals that

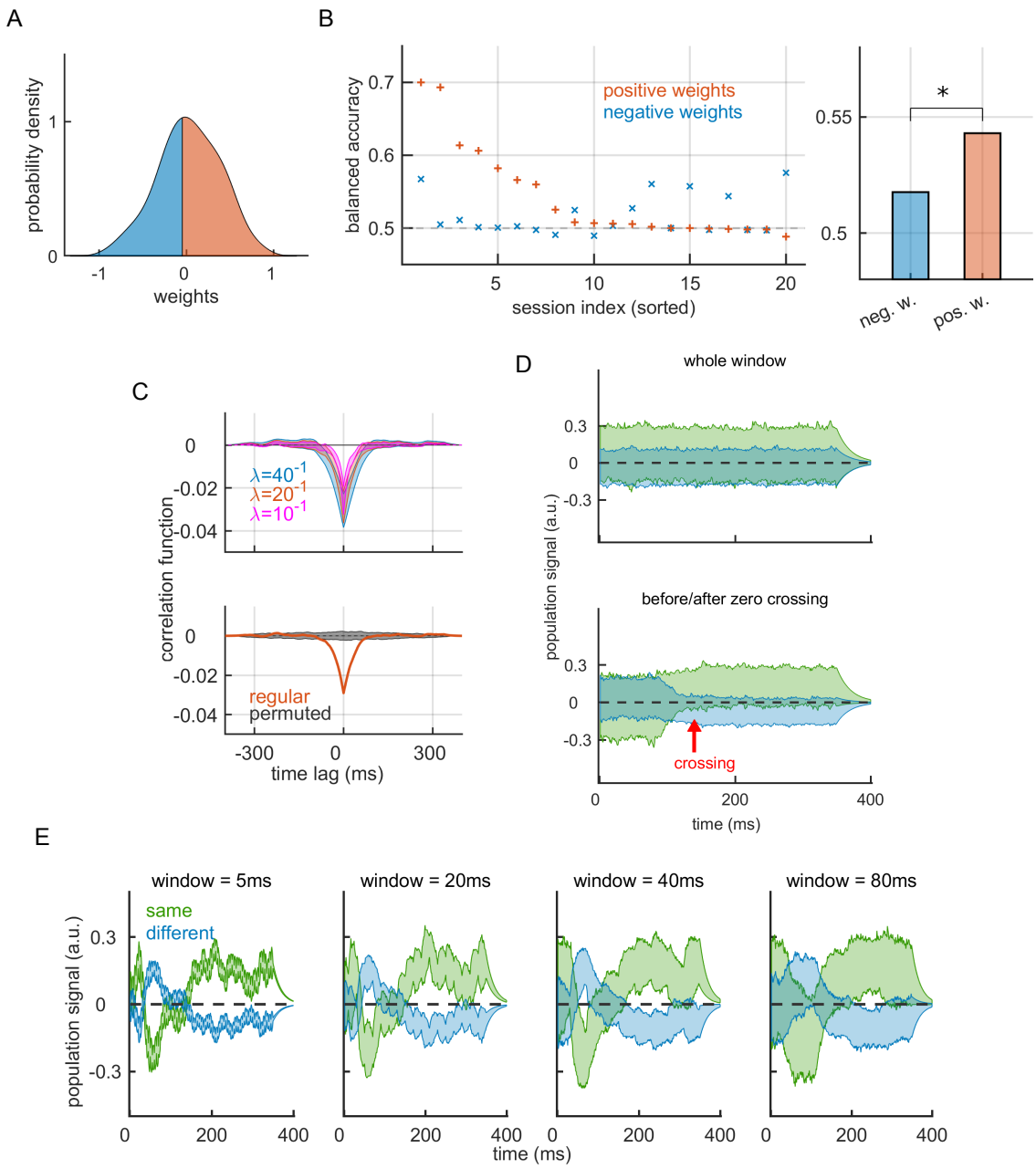


Figure 4. Population signals of neurons with positive and negative weights are anticorrelated. **(A)** Distribution of population decoding weights. **(B)** Prediction accuracy on the variable *stimulus* + *choice* utilizing exclusively the information from the *plus* subnetwork (red) and from the *minus* subnetwork (blue). Left plot shows balanced accuracy in sessions, while the right plot shows session-averaged balanced accuracy. **(C)** Top: Correlation function of choice signals between *plus* and *minus* subnetworks, for different values of the time constant of the convolution λ. Bottom: Session-averaged correlation function for the regular model with $\lambda^{-1} = 20$ ms (red) and the distribution of results for models with permutation (gray area). **(D)** Distribution of choice signals with randomly permuted spike timing across the entire time window (top) and in two time windows (bottom), before and after the switch in decoded choice at 140 milliseconds (red arrow). **(E)** The choice signal with temporal jitter, for different length of the jittering window. Parameters: $\lambda^{-1} = 20$ ms, $N_{perm} = 1000$.

preclude discrimination (Figure 4D, top). However, if the spike timing is permuted in the time window only before and after the switch in decoded choice (140 ms after the stimulus onset), the overlap is smaller (Figure 4D, bottom), again suggesting different sources of the choice signal before and after the switch. The temporal jitter of spike trains in short time windows of [5,20, 40, 80] ms makes the choice signals noisier, but still allows discrimination (Figure 4E). Robustness of the read-out to jitter on short time scales suggests that the choice signal is integrated over time.

A plausible mechanism for the integration of the choice signal over short time scales is correlated spiking of neurons with the same sign of the weight. If neurons with the same sign of the weight spike synchronously or are positively correlated on a short time scale, this helps the population signal to accumulate (Figure 5A, top). If, on the contrary, neurons with the opposite sign fire in alternation, the population signal remains close to zero and cannot be distinguished from random noise (Figure 4A, bottom). Measuring noise correlations of spike timing with the cross-correlogram (see methods, and Bair et al. (2001)), we find that neurons in the *plus* subnetwork are more strongly synchronized ($p < 0.001$, permutation test; Supplementary Table S5.1) and correlated than neurons in the *minus* subnetwork (Figure 5B). Together with more informative decoding weights, correlated spike timing explains better predictive power of *plus* neurons compared to *minus* neurons. Correlations of spike timing among *plus* neurons might therefore be a dynamical mechanism to enhance the choice signal in this subpopulation.

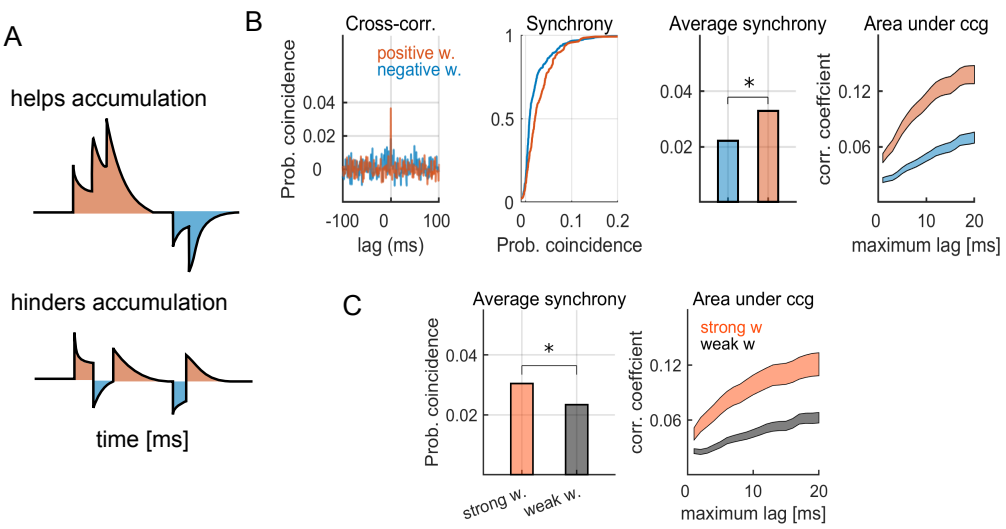


Figure 5. Noise correlations among neurons with positive weights help the formation of the choice signal. (A) Schematic of spike patterns that either help or hinder the accumulation of the signal. **(B)** Synchrony and correlations within the group of *plus* (red) and *minus* (blue) neurons. Left: example cross-correlograms for a pair of neurons with positive weights (red) and negative weights (blue). Middle left: empirical cumulative distribution functions of the cross-correlograms at zero time lag, measuring pair-wise synchrony in the time window of ± 1 ms. The distribution shows all neuronal pairs, collected across recording sessions (288 and 143 pairs of *plus* and *minus* neurons, respectively). Middle right: average synchrony between pairs of neurons with positive and negative weights. Right: Area under cross-correlograms as a function of the maximal lag used for integration. **(C)** Same as **(B)** middle right and right, but for pairwise interactions within the group of neurons with strong weights (orange, N=147) and weak weights (dark gray, N=176). Parameters: $\lambda^{-1} = 20$ ms, $N_{perm} = 1000$.

Informative neurons have stronger spike-timing correlations than uninformative neurons. If correlated spiking between neurons with similar weight is beneficial for the formation of a time-resolved choice signal, we would expect informative neurons to be more strongly correlated than uninformative neurons. We label "informative" neurons with a large amplitude of decoding weight $|\tilde{w}_n|$. The amplitude of decoding weight tells the neuron's contribution to the classification model, and neurons with strong weight amplitude are those that contribute

the most to classification. We separate the population of simultaneously recorded units in neurons with strong (79 neurons collected across the recording sessions) and weak weights (81 neurons; see methods) and compute the population signal for each subnetwork. As expected, the subnetwork with strong weights is the one that carries the choice signal, while the subnetwork with weak weights performs at chance (Supplementary Figure S3). As we compute spiking synchrony and correlations within informative and uninformative neurons, we find that informative neurons have stronger synchrony and correlation than uninformative neurons ($p = 0.0140$ permutation test, Figure 5C; Supplementary Table S5.2). This result is an additional argument in favor of the hypothesis that synchrony and correlations on short time scales are beneficial for the formation of the choice signal.

Activity of bursting neurons and of neurons in the superficial layer is particularly informative about the choice

Besides correlations and synchrony, another way to help the formation of the choice signal might be bursting of single neurons. Bursting can help to increase the amplitude of a population signal for the same reason as correlated spiking of neurons with the same sign of the weight. We divide neurons into bursting and non-bursting, based on characteristics of the power spectrum of spike trains. The power spectrum of bursting neurons is characterized by a reduced power for middle range frequencies (Compte et al. 2003) (Figure 5A). We capture this effect by computing the normalized area under the power spectrum. We classify a neuron as bursting if it has the area under the power spectrum smaller than the lower bound of models with permuted spike timing (Figure 5B; see methods). As we compute the choice signal for bursting and non-bursting neurons, we find that bursting neurons are largely better at predicting the choice (Figure 5C). Choice signals of bursting and non-bursting subnetworks are positively correlated (Figure 5D; supplementary table S4.3).

Similarly to *plus* neurons, bursting neurons are better at predicting the choice because they have more informative weights and because they are more strongly temporally correlated than non-bursting neurons. Weights of bursting subnetworks predict the variable *stimulus+choice* better than the non-bursting subnetworks ($p < 10^{-17}$, two-tailed Wilcoxon signed-rank test; Figure 5E). Consistently, the more bursting the neuron, the stronger weight it tends to have. The area under the power spectrum correlates negatively with the strength of the weight, but only for neurons with positive weight ($p = 0.012$ for neurons with positive weights, $p = 0.610$ for neurons with negative weight, permutation test; Figure 5F). Bursty neurons have stronger firing rate than non-bursting neurons (Figure 5G), stronger spiking synchrony ($p < 0.001$, permutation test; Figure 5H; supplementary table S5.3) and stronger noise correlations of spike timing than non-bursting neurons (Figure 5H, right).

At last, we divide neurons with respect to their location along the cortical depth in three layers. Cortical layers in V1 have been shown to importantly differ in their correlation structure (Hansen et al. 2012), which can have direct implications on the population code. Using a method based on the current source density (Koren et al. 2019), we distinguish the superficial layer (supragranular, SG, N=48), the middle layer (granular, G, N=51) and the deep layer of the cortex (infragranular, IG, N=61). As we compute the choice signal in each layer, we find that the superficial layer is by far the most important for discriminating the choice (Figure 6A). The superficial layer also reflects the switch in decoded choice at around 140 ms after the stimulus onset, seen in the overall response. Choice signals are weakly positively correlated across all pairs of layers, consistently for all tested time constants of the convolution (Figure 6B; supplementary table S4.4).

Measuring pair-wise synchrony of spike timing with the cross-correlogram, we find that the synchrony is stronger in the middle and the superficial layer compared to the deep layer ($p < 0.001$ for both SG/IG and G/IG, permutation test; Figure 6C, middle plots). The strength of synchrony in the middle and the superficial layer is not significantly different ($p = 0.573$; Supplementary table S5.4), however, the superficial layer clearly shows stronger correlations of spike timing (Figure 6C, right). Besides strongest spike-time correlation, the fact that there is excess of bursting neurons and of neurons with positive weights in the superficial layer (Figure 6D) explains why the superficial layer has stronger discriminatory capacity compared to other layers. In the

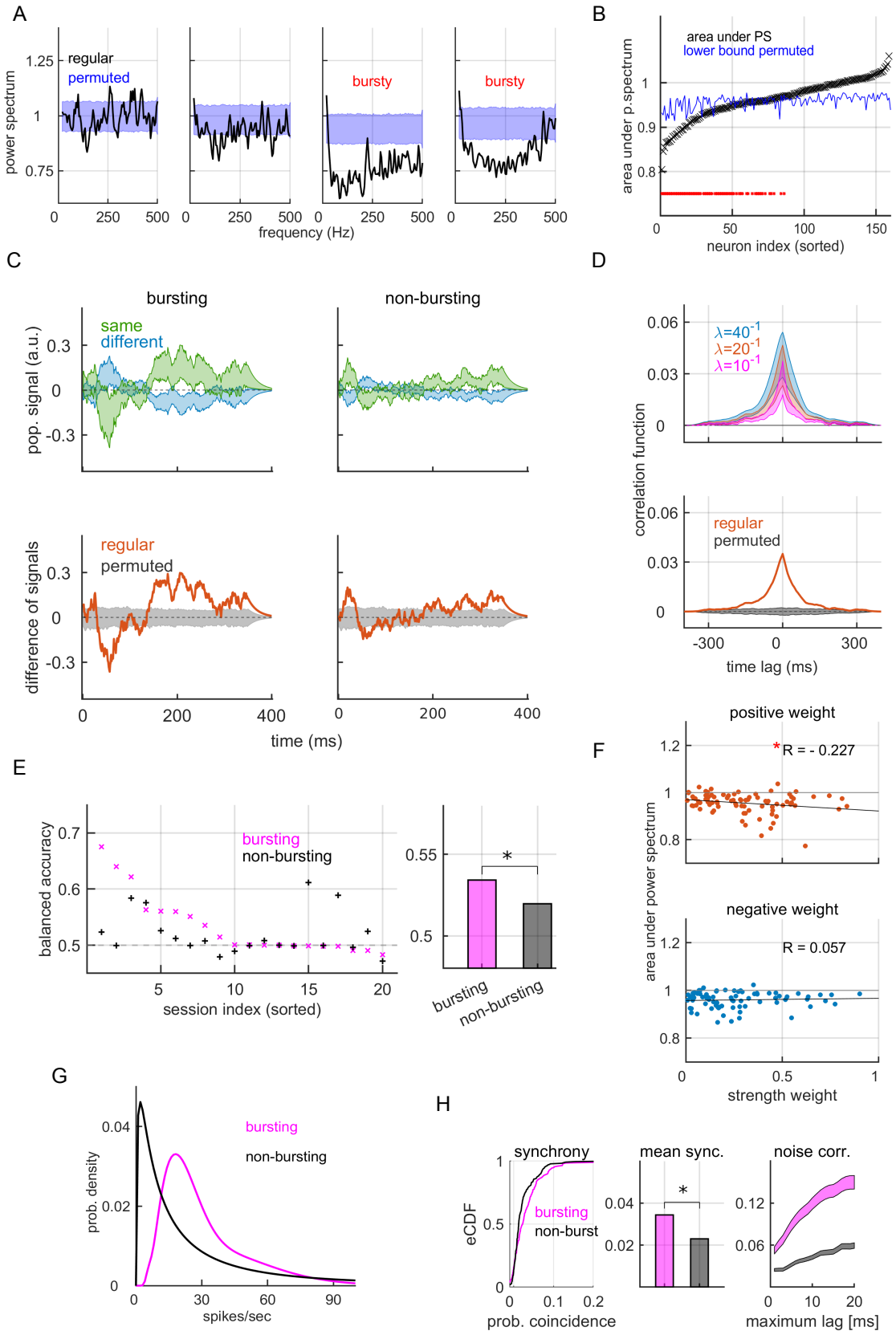

A
power spectrum
regular
permuted
bursty
bursty
frequency (Hz)
B
area under p.spectrum
area under PS
lower bound permuted
neuron index (sorted)
C
bursting
non-bursting
pop. signal (a.u.)
same
different
difference of signals
regular
permuted
time (ms)
D
correlation function
λ=40^-1
λ=20^-1
λ=10^-1
regular
permuted
time lag (ms)
E
balanced accuracy
bursting
non-bursting
session index (sorted)
bursting
non-bursting
F
positive weight
R = - 0.227
negative weight
R = 0.057
area under power spectrum
strength weight
G
prob. density
bursting
non-bursting
spikes/sec
H
synchrony
eCDF
bursting
non-burst
prob. coincidence
mean sync.
noise corr.
maximum lag [ms]

Figure 5. Bursty neurons discriminate better and earlier in the trial. (A) Power spectrum for 4 representative neurons from the same recording session. We show the power spectrum for the true model (black) and the distribution of power spectra for models with randomly permuted spike timing (blue). Bursty neurons are those with reduced power in lower and medium range frequencies (middle right and right). **(B)** Area under the power spectrum for all neurons, sorted from the lowest to the highest value (black), and the lower bound of the area under the power spectrum for models with permuted spike timing (blue). Red asterisk marks significance (permutation test, $\alpha = 0.05$). There are 68 bursting and 92 non-bursting neurons. **(C)** Top: The choice signal for bursting (left) and non-bursting subnetworks (right), showing the mean $\pm SEM$ in recording sessions. Bottom: Session-averaged difference of signals for the true model (red) and for models with random permutation (gray area). Parameter: $\lambda^{-1} = 20$ ms. **(D)** Top: Correlation function between the choice signal of bursting and non-bursting subnetworks. We show the mean ± SEM for the variability across recording sessions. Bottom: Session-averaged correlation function for the regular model (red) and the distribution of results for the model with permutation (gray area). **(E)** Population signal for bursting (magenta) and non-bursting (black) subnetworks. **(F)** Scatter plot of the area under the power spectrum vs. strength of the weight, for neurons with positive weights (top) and neurons with negative weights (bottom). We used trials corresponding to the variable $stimulus + choice$. R marks the Pearson correlation coefficient. **(G)** Distribution of firing rates of bursting and non-bursting neurons. **(H)** Left: Cumulative distribution function of the coefficient of synchrony for bursting (magenta) and non-bursting neurons (black). Middle: Averages of the distribution on the left. Right: Noise correlation of spike timing. There are 198 pairs of bursting and 231 pairs of non-bursting neurons. Parameter: $N_{perm} = 1000$.

superficial layer, 58 % of neurons have positive weights, while in the middle and deep layers, the percentage of neurons with positive weights is 49 and 51 %, respectively. Moreover, 54 % of neurons in the superficial layer are bursting, while in the middle and deep layer, 39 and 36 % of neurons are bursting, respectively.

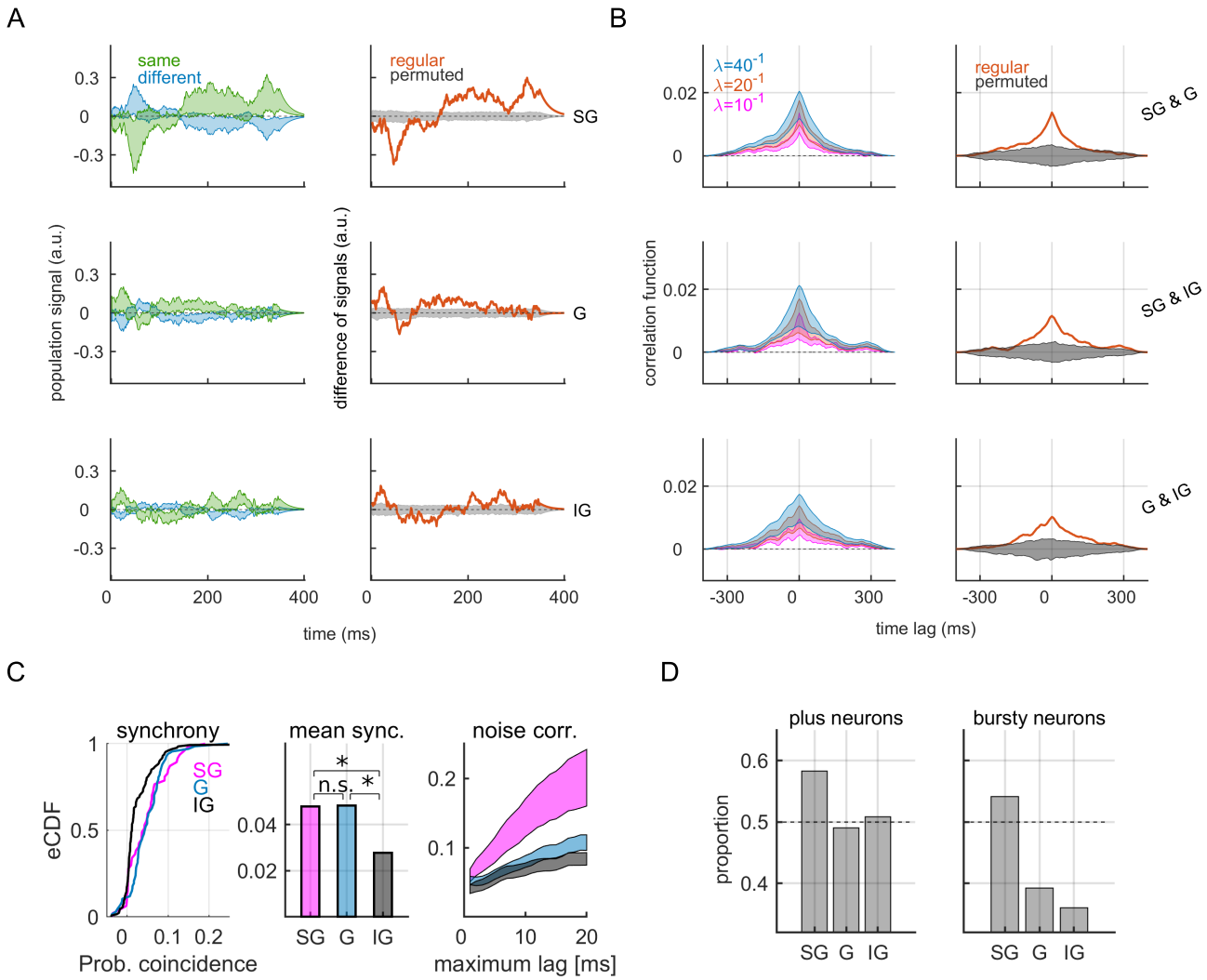


Figure 6. Spiking activity in the superficial layer of the cortex is the most discriminative of the choice behavior. (A) Left: The choice signal in the SG (top), G (middle) and IG layer (bottom), showing the mean ± SEM across recording sessions. Right: Session-averaged difference of choice signals for the true model (red) and the model with permuted class labels (gray). Parameters: $\lambda^{-1} = 20$ ms, $N_{perm} = 1000$. **(B)** Left: Correlation function of choice signals across pairs of layers. We show the mean ± SEM across recording sessions and plot results for three values of the time constant of the convolution. Right: Session-averaged correlation function for the regular model (red) and for models with permutation (gray area). We use the time constant $\lambda^{-1} = 20$ ms. **(C)** Same as in Figure 5H, but in cortical layers. **(D)** Proportion of neurons with positive weight(left) and the proportion of bursting neurons (right) in cortical layers. The proportion is computed relative to the tested group (numbers do not sum to 1 across layers).

Discussion

In this work, we decoded the future behavioral choice from the spiking activity in the primary visual cortex. We showed that the choice can be decoded as a linear sum of spikes of neural populations, and that decoding weights need to be learned in correct trials. We argued that the generalization of learning from a mixed variable $stimulus + choice$ to $choice$ is possible due to the partial overlap in the representation of the stimuli and the choice. Furthermore, we found that bursting neurons and neurons in the superficial layer of the cortex are particularly informative for the task. Our results suggested that bursting of single neurons and correlated spiking of neurons with similar decoding selectivity are dynamical mechanisms that help the formation of the time-resolved choice signal.

Current results raise a question, why is learning in correct trials necessary for the prediction of the choice? The following scenario during the condition "incorrect non-match" is compatible with all our observations: 1) the test stimulus is "non-match" and V1 responds with a spiking pattern corresponding to the stimulus class "non-match" during the first 100 or so milliseconds after the onset of the stimulus, 2) there is an error in transmission of the signal to downstream areas that will result in incorrect choice (Shahidi et al. 2019) 3) due to the error in transmission, a high-level area erroneously signals the stimulus class "match" and the choice "same" and projects this signal back to V1, and 4) as the projection is received by V1, the spiking pattern in V1 switches and signals the stimulus class "match" and the choice "same". This scenario allows the activity in V1 to be predictive of the choice simply through the alignment of the representation about the stimuli and the choice, activated by a top-down projection.

If the representation in the context of $stimulus+choice$ would rely exclusively on the difference in stimuli, we would not be able to predict the variable $choice$ by learning population weights on $stimulus+choice$. If, on the contrary, the representation of the choice and the stimulus class in V1 are mixed (Rigotti et al. 2013), learning in the context of $stimulus + choice$ can generalize to the representation of $choice$, due to the partial overlap in the information between the two classification problems (Panzeri et al. 2017). In the brain, such alignment can be instrumental, as it reinforces the representation of the upcoming behavioral choice across brain networks. The choice was shown to be represented in multiple sensory areas in parallel (Steinmetz et al. 2019). Such multiplicity of representations could serve the purpose of reverberating and amplifying the choice-related signal within the cortical circuitry.

Another result that merits the discussion is the role of noise correlations of spike timing for the accumulation of the choice signal. Noise correlations of spike counts between neurons with similar selectivity are known to decrease the quantity of information in a rate code (Averbeck et al. 2006), and they do so also in the present dataset (Koren et al. 2020). However, in the present decision-making task, rather than transmitting all the information about the stimulus, neural networks are faced with the task of extracting task-relevant information and creating a binary variable about the choice of the animal. A binary variable only consists of 1 bit of information and rather than the quantity of information, the robustness of coding might be a primary concern. Recently, the work by (Zylberberg et al. 2017) pointed out that noise correlations among neurons with similar tuning are useful because they improve the robustness of transmission of the rate code downstream. In the present work, we show how structured noise correlations of spike timing can be instrumental in a time-resolved population code. Correlated spiking of neurons with similar decoding weights allows the accumulation of the population signal and makes the signal rise from the noisy background. Note that if the spiking activity of neurons with the same sign of the weight was entirely uncorrelated, the read-out of such uncorrelated spiking would be a noisy oscillation around zero. From such a signal it would be indeed impossible to read-out the choice. Present results therefore suggest a simple and biologically relevant mechanism that leverages correlated spike timing to build the choice signal and help the transmission of the choice signal downstream.

Previous work has suggested that the structure of noise correlations in V1 is primarily determined by a common input (Nienborg et al. 2012), possibly from a top-down source (Bondy et al. 2018). In line with these results, we showed that the spiking synchrony and noise correlations of spike timing are stronger between neurons that are also carriers of the choice signal. As long as top-down projections are specific about the sign

of decoding weights of V1 neurons, such projections can at the same time drive the choice signal and create the observed pattern of noise correlations where neurons with similar weights are more strongly correlated. It is sufficient that the top-down input selectively drives neurons with positive weights in trials with decision "same" and neurons with negative weights in trials with decision "different". Precise long range microcircuit-to-microcircuit connections between the prefrontal cortex and V1, recently put in evidence (Ren et al. 2019), are a good candidate for conveying such a feedback signal.

Besides noise correlations of spike timing, we have suggested that bursting of single neurons is a useful dynamical feature for encoding of the choice signal. While bursting might help the accumulation of the choice signal, bursting of single neurons can also increase the reliability of transmission to downstream areas by counteracting the synaptic failure (Lisman 1997), or by driving synchronization in Gamma frequency (Onorato et al. 2020). Bursting neurons observed in our data might correspond to recently reported class of bursting excitatory neurons, located predominantly in the superficial layers of V1 of the macaque (Onorato et al. 2020). Moreover, consistently with our hypothesis, bursting neurons have been proposed to integrate bottom-up and top-down inputs (Larkum 2013; Naud and Sprekeler 2018), which might be useful in the present setting.

The role of feedforward and feedback inputs for the representation of the choice in sensory areas is a matter of ongoing discussion (Cumming and Nienborg 2016; Nienborg and Cumming 2009; Nienborg et al. 2012; Nienborg and Cumming 2014; Wang 2002). Several results of the present analysis suggest that both bottom up and top-down sources contribute to the choice signal. One of them is the switch in decoded choice at about 140 ms after the stimulus onset. We observed that the difference of signals $\Delta\ x(t) = x^{\text{same}}(t) - x^{\text{different}}(t)$ crosses the zero line at about 140 ms after the stimulus onset (Figure 2G), implying a switch in decoded choice before and after this moment within the trial. The timing of the switch coincides with the qualitative change of the population response from the strong transient at the stimulus onset to tonic firing thereafter (Figure 2F). The transient elevation of firing rates at the stimulus onset is dominated by the flow of feedforward current to the network, while feedback sources are active during tonic firing that follows (Callaway 2004). The switch in decoded choice therefore raises the possibility of conflicting information from feedforward and feedback sources to the network during trials with incorrect behavior. Note that the timing of the switch in decoded choice is consistent across recording sessions, indicating a precise and reliable timing of the feedback signal across recording sessions. Precise timing of the feedback signal has been shown to be important for performing computations in recurrent networks (Spoerer et al. 2020) and to be crucial for perception (Silvanto et al. 2005; Hupé et al. 1998). Supposing that in incorrect trials the top-down and bottom-up signals are incongruent, we would expect to have congruent top-down and bottom-up signals in correct trials. In our previous work, we have indeed shown that with correct behavior, the variable *stimulus* + *choice* is stable across the trial (Koren et al. 2019), and the representation does not switch as it does when decoding the *choice*.

Several other results exclude the possibility that the feed-forward signal is the only source of choice-related information in V1. If the feed-forward drive alone was decisive for decoding of the choice, we would expect to decode the choice already in the middle cortical layer, which is not supported by present results. Moreover, we have observed that permuting spike timing over the entire time window destroys discriminability of the choice, while limiting the permutation to the time window before and after the switch in decoded choice, discrimination is to some degree preserved. Another piece of evidence in favor of the top-down origin of the choice-related information is the negative correlation of population signals between neurons with positive and negative weights. The population signal between neurons with positive and negative weights were negatively correlated while a positive correlation was observed for all other groupings (e.g., bursting/non-bursting, strong/weak weights, layers). These effects can be explained by a top-down input that selectively drives either neurons with positive of negative weights, or by a specific lateral connectivity where neurons with similar weight are more strongly connected than average. A modeling study is necessary to test these possible mechanisms.

Recent study in monkey middle temporal area suggested that the stimulus- and the choice-related information lie in partially distinct subspaces, with a substantial component of the choice-related information in a subspace orthogonal to the stimulus subspace (Zhao et al. 2020). These results are not incompatible with our findings, since we found that the population vectors corresponding to variables *stimulus* + *choice* and *choice* are only

weakly similar, allowing for a large component of neural responses to lie in orthogonal subspaces. In the present data, population vectors deviate for about 10 degrees from orthogonality, and therefore deviate for about 80 degrees from alignment. In (Zhao et al. 2020), authors suggest that the benefit of the partial misalignment of the stimulus and the choice is the co-existence of two representations, as the feedback can target the "choice" subspace without corrupting the representation in the "stimulus" subspace. Our study, on the other hand, points out the benefits of the partial alignment of subspaces containing the information on the stimulus class and the choice. The amount and the function of the overlap between representations of the stimulus and the choice likely depends on the brain area. The primary visual cortex, the first stage of cortical processing, is unlikely to have a representation of the choice that is independent from the representation of the stimulus. In higher cortical areas, however, the representation of the stimulus and of cognitive variables such as the choice might become more independent.

Acknowledgments

This work was supported by Deutsche Forschungsgemeinschaft (grant number GRK 1589/2) and the Equal Opportunity Program of the Technische Universität Berlin.

Author contributions

Conceptualization: V.K. Methodology: V.K., A.A., M.H., V.D., K.O.; Software: V.K.; Validation: V.K.; Formal Analysis: V.K.; Investigation: A.A., M.H.; Resources: V.D., K.O.; Data curation: A.A. and V.K.; Writing-Original Draft: V.K.; Writing-Review and Editing: A.A., V.D., K.O., V.K.; Visualization: V.K.; Supervision: V.D., K.O.; Project Administration: V.K., A.A.; Funding Acquisition: K.O.

Declaration of interests

The authors declare no competing interests.

Methods

Resource availability

Lead contact Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Klaus Obermayer (klaus.obermayer@tu-berlin.de).

Materials availability This study did not generate new unique reagents.

Data and code availability Code is available in a public GitHub repository:
https://github.com/VeronikaKoren/transfer learning

Experimental model and subject details

Ethics statement All experiments were conducted in accordance with protocols approved by The Animal Welfare Committee (AWC) and the Institutional Animal Care and Use Committee (IACUC) for McGovern Medical School at The University of Texas Health Science Center at Houston (UTHealth), and met or exceeded the standards proposed by the National Institutes of Health's Guide for the Care and Use of Laboratory Animals.

Animal subjects

The experimental method and animal subject details were described in our previous work (Koren et al. 2020). In the following, we provide a brief summary of the experiment. Two male rhesus macaques (*Macaca mulatta*; M1, 7 years old, 15kg; M2, 11 years old, 13kg) were used in this study. Animals were trained on a delayed match-to-sample task with visual stimuli. As the animal successfully fixated within the fixation area, this automatically triggered the trial. Two naturalistic images appeared consecutively, with a delay period in between. The target and the test stimulus were shown for 300 ms each, with a variable delay period in between (800 to 1000 ms). The identity of stimuli, complex naturalistic images depicting an outdoor scene, changed from one recording session to the other. Within the same session, the target and the test stimuli could be either the same (condition "match") or else the test stimulus was rotated with respect to the target (condition "non-match"). The difference in orientation was such as to keep the performance stable at around 70 percent correct on non-matching stimuli, and ranged between 3 and 10 degrees. Animal subjects were trained to communicate their choice by holding the bar for the choice "different" and releasing the bar for "same". For the trial to be valid, the behavioral response had to be communicated between 200 and 1200 ms after the offset of the test stimulus.

Method details

Multi-unit signal was measured with laminar arrays comprising 16 channels with 100 $\mu-$ meter spacing between adjacent contacts. Electrodes captured neuronal activity across the cortical depth, from superficial to deep layers. The multi-unit signal was spike sorted and we analyzed all cells that responded to the stimulus with at least a 4-fold increase of the firing rate with respect to the baseline. The data was collected in 20 recording sessions, which gave 160 neurons. We analyzed 3 conditions, "correct match" (CM), "correct non-match" (CNM) and "incorrect non-match" (INM), where "correct/incorrect" refers to the behavioral performance and "match/non-match" refers to the stimulus class. In condition "incorrect match", there were not enough trials to perform the analysis. Table 1 reports the summary statistics on the number of trials (biological replicates) in each condition.

Condition	Nb. trials min.	Nb. trials max.	Nb trials average
CM	51	290	118
CNM	42	230	107
INM	21	98	39

Table 1. The minimal, maximal, and average number of trials across recording sessions, for each condition.

Quantification and statistical analysis

The analysis was done with Matlab R2019, Mathworks.

Parallel spike trains The spike train of a single neuron n in trial j is defined as a binary vector of zeros and ones,

$$o_{n,j}(t_k) = \begin{cases} 1, & \text{if neuron } n \text{ in trial } j \text{ spikes during the } k\text{-th millisecond} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $n = 1, ..., N$ is the neural index, $k = 1, ..., K$, is the time index with step of 1 millisecond, and $j = 1, .., J_1, J_1 + 1, ..., J_2, J_2 + 1, ...J$ is the trial index. Trials were collected in conditions correct match ($j = 1, ..., J_1$), correct non-match ($j = J_1 + 1, ..., J_2$) and incorrect non-match ($j = J_2 + 1, ..., J$).

Coefficient of variation The coefficient of variation for the neuron n in trial j is defined as follows,

$$CV2_{n,j} = 2\frac{|ISI_{i+1} - ISI_i|}{ISI_{i+1} + ISI_i}$$

where ISI is the inter-spike interval with index $i = 1, ..., N_{int}$. We report trial-averaged results, distinguishing trials from conditions "correct non-match" (decision "same") and "incorrect non-match" (decision "different").

Sensitivity index The sensitivity index of neuron n is defined as follows:

$$d'_n = \frac{|\mu_n^{\text{match}} - \mu_n^{\text{non-match}}|}{\sigma_n} \tag{2}$$

where μ_n^{match} ($\mu_n^{\text{non-match}}$) is the average spike count in condition "match" ('non-match'), and σ_n is the standard deviation of spike counts. To estimate the significance of d', we randomly permute (without repetition) trial labels "match" and "non-match", compute d' with randomly permuted data and repeat the procedure $N^{perm} = 1000$ times. The p-value is then calculated by ranking the d' on original data among the distribution of N^{perm} values of d' with permuted indices. Note that random permutations are technical replicates (Blainey et al. 2014), here as well as throughout the paper.

Parallel spike counts Spike counts were computed in the time window of $[0, K]$ ms with respect to the onset of the stimulus (target or test), $s_{n,j} = \sum_{t_k=1}^{K} o_{n,j}(t_k)$. For the parameter K we used $K = [300, 400, 500]$, as reported in the Results section. Spike counts were z-scored:

$$\tilde{s}_{n,j} = \frac{s_{n,j} - \langle s_{n,j} \rangle_j}{\sqrt{\text{Var}_j\ (s_{n,j})}}, \tag{3}$$

where $\langle s_{n,j} \rangle_j$ and $\text{Var}_j\ (s_{n,j})$ are the mean and the variance across trials from all conditions.

Prediction accuracy of a linear classifier Z-scored spike counts of N neurons that have beer recorded in parallel are utilized as input to the linear classification model (linear Support Vector Machine, SVM). The prediction accuracy of the classifier is computed as the balanced accuracy (BAC) on the hold-out test set:

$$BAC = \frac{TP}{2(TP + FN)} + \frac{TN}{2(TN + FP)}, \tag{4}$$

with TP, FN, TN and FP the number of true positive, false negative, true negative and false positive classifications on the test set, respectively. The balanced accuracy is cross-validated $N_{cv} = 100$-times and reported values are averaged across cross-validations. The significance of the balanced accuracy is evaluated with the permutation test, utilizing $N_{perm} = 1000$ random permutations of class labels without repetition.

When the population is divided in groups (see below), the prediction accuracy for the group of interest is computed by utilizing all simultaneously recorded neurons, while removing the information from neurons that do not belong to the group of interest (see Koren et al. (2020)). While class labels for the group of interest are correct, they are randomly assigned for the remaining of neurons, and those therefore only contribute chance-level prediction. The advantage of this method is to keep the dimensionality of the dataset intact, thus removing the difference in the number of neurons as a possible cause of the difference in prediction accuracy. The difference in accuracy between neurons with positive and negative weights is tested with the t-test on 100 instances of the balanced accuracy with random permutation of the class label (without repetition).

Learning of structural features from parallel spike counts

Our decoding procedure comprises two steps, learning of decoding weights, and the computation of the population signal. Decoding weights are estimated as feature vector of the linear classifier, utilizing spike counts of simultaneously recorded neurons. We then use decoding weights to compute the population signal reflecting

animal's choice. The population signal is computed on spike trains of simultaneously recorded neurons in single trials, without temporal averaging.

Decoding weights are estimated from trials with condition CM and half of the trials with condition CNM (trials with index $j = 1, ..., J_1, J_1 + 1, ..., \frac{J_2 - J_1}{2}$). The computation of the choice signal is computed on a hold-out set, utilizing the remaining half of the trials from condition CNM and all trials from condition INM (trials with index $j = \frac{J_2 - J_1}{2} + 1, ..., J_2, J_2 + 1, ..., J$). The split of trials in training and reconstruction set in condition CNM is cross-validated with Monte-Carlo method, using $N_{cv} = 100$ random splits. The split of trials in training and reconstruction set is non-overlapping, such that no trials that have been used for training appear in the reconstruction set. All reported results are averaged across cross-validations.

Estimation of the population vector In the N-dimensional space of inputs, one sample is the vector of z-scored spike counts of N simultaneously recorded neurons in trial j, $\mathbf{s}_j = [\tilde{s}_{1,j}, \tilde{s}_{2,j}, ..., \tilde{s}_{N,j}]^T$. Linear SVM searches for an $N-1$-dimensional plane (a hyperplane) that optimally separates points in conditions CNM and CM. The hyperplane is defined as follows,

$$H_0 : \mathbf{w}^T \mathbf{s}_j + b = w_1 s_{1,j} + w_2 s_{2,j} + ... + w_N s_{N,j} + b = 0 \tag{5}$$

where $\mathbf{w}$ is the vector of feature weights and b is the offset of the hyperplane from the origin. On each side of H_0, we can define a hyperplane that verifies the following:

$$\begin{aligned} H_1 : \mathbf{w}^T \mathbf{s}_j + b = -1 \\ H_2 : \mathbf{w}^T \mathbf{s}_j + b = 1 \end{aligned} \tag{6}$$

If the problem is linearly separable, all training samples verify the following the inequality,

$$\begin{aligned} \mathbf{w}^T \mathbf{s}_j + b \leq -1 \text{ if } y_j = -1 \\ \mathbf{w}^T \mathbf{s}_j + b \geq 1 \text{ if } y_j = 1 \end{aligned} \tag{7}$$

where $y_j \in \{-1, 1\}$ is the class label ($y_j = -1$ in condition CNM and $y_j = 1$ in condition CM). Training the linear SVM consists in maximizing the number of correctly classified samples and, at the same time, minimizing the distance between H_1 and H_2, which can be expressed with the Lagrangian.

$$L_p = \frac{1}{2} \mathbf{w}^T \mathbf{w} - \sum_{j=1}^{\frac{J_2 - J_1}{2}} \lambda_j [y_j (\mathbf{w}^T \mathbf{s}_j + b) - 1] \tag{8}$$

The first term on the right hand side ensures the maximal distance between hyperplanes H_1 and H_2, and the second term on the right ensures correct classification. As the derivative of the Lagrangian with respect to $\mathbf{w}$ is set to zero, we get the expression for the vector of weights.

$$\mathbf{w} = \sum_{j=1}^{\frac{J_2 - J_1}{2}} \lambda_j y_j \mathbf{s}_j \tag{9}$$

Since $\lambda_j \neq 0$ only for trials that define the margin (points that lie on $H1$ or on $H2$), the weight vector only depends on support vectors,

$$\mathbf{w} = \sum_{q=1}^{Q} \lambda_q y_q \mathbf{v}_q \tag{10}$$

where $q = 1, 2, ..., Q$ are the support vectors, with $Q < \frac{J_2 - J_1}{2}$. The weight vector is normalized with the L2 norm,

$$\tilde{\mathbf{w}} = \frac{\mathbf{w}}{||\mathbf{w}||} \tag{11}$$

with $||\mathbf{w}||$ defined as follows:

$$||\mathbf{w}|| = (w_1^2 + ... + w_N^2)^{\frac{1}{2}}. \tag{12}$$

We refer to the vector $\tilde{\mathbf{w}} = [\tilde{w}_1, \tilde{w}_2, ..., \tilde{w}_n, ..., \tilde{w}_N]^T$ as the population vector and to the $n-$th entry of $\tilde{\mathbf{w}}$, $\tilde{w}_n$, as the decoding weight of the neuron n.

The population signal as a read-out of parallel spike trains

Decoding weights are utilized to compute the population signal as a weighted sum of spikes of simultaneously active neurons. While the learning step is identical to decoding protocol proposed in (Koren et al. 2019), we now compute the population signal on the variable *choice* instead of *stimulus* + *choice*.

Consider the vector of spike trains of N simultaneously recorded neurons.

$$\mathbf{o}_j(t_k) = [o_{1,j}(t_k), o_{2,j}(t_k)..., o_{N,j}(t_k)]^T \tag{13}$$

The choice signal is a projection of the vector of spike trains on the population vector,

$$x_j(t_k) = F(\tilde{\mathbf{w}}^T \mathbf{o}_j(t_k)) - \langle F(\tilde{\mathbf{w}}^T \mathbf{o}_j(t_k))\rangle_j \tag{14}$$

where $\langle F(\tilde{\mathbf{w}}^T \mathbf{o}_j(t_k))\rangle_j$ is the trial average, utilizing all trials of the reconstruction set. By subtracting the trial average, we compute the deviation of the signal from the mean. We argue that the deviation of the signal from its mean, rather than the absolute value of the signal, is a signal of biological relevance. As the transfer function, $F(y(t_k))$, we use a convolution with an exponential kernel,

$$F(y) = \sum_{\tau \in T} y(t_k - \tau) u(\tau) \tag{15}$$

with $u(\tau) = exp(-\lambda\tau),\ \tau \in T$, with support $T = \{-100, ..., 100\}$ ms. Convolution with an exponential kernel models the causal effect of the presynaptic spike on the neural membrane of the read-out neuron. Notice that $x_j(t_k)$ is a time-resolved, low-dimensional representation of parallel spike trains in trial j.

To test the discriminability of the choice, we average the choice signal across trials, distinguishing conditions CNM (decision "different") and INM (decision "same"),

$$\begin{aligned} x^{\text{same}}(t_k) &= \frac{2}{J_2 - J_1} \sum_{j=J'}^{J_2} x_j(t_k) \\ x^{\text{different}}(t_k) &= \frac{1}{J - J_2} \sum_{j=J_2+1}^{J} x_j(t_k) \end{aligned} \tag{16}$$

where $J' = \lceil \frac{1}{2}(J_1 + J_2) \rceil + 1$, and $\lceil z \rceil$ is the ceiling function . The significance of the difference between choice signals for the decision "same" and "different" is evaluated with the permutation test. We compute the difference of the choice signals in every recording session,

$$\Delta x(t_k) = x^{\text{same}}(t_k) - x^{\text{different}}(t_k) \tag{17}$$

and average across sessions. We then rank $\Delta x(t_k)$ among $\Delta x_p^{perm}(t_k)$, where the latter has been computed with random weights, and the class labels in the validation set have been randomly permuted. Random weights were drawn from the uniform distribution with the same range as the regular weights. The permutation procedure is repeated N_{nperm}-times and gives a distribution of results for each time step. When the result of the true model, $\Delta x(t_k)$, appears outside of the distribution of results of the null model, $\Delta x_p^{perm}(t_k)$ for $p = 1, 2, ..., N_{perm}$, we consider that signals $x^{\text{same}}(t_k)$ and $x^{\text{diff}}(t_k)$ have been successfully discriminated.

Criteria for division into subpopulations

Sign of the weight We separate the population with respect to the sign of entries of the population vector, distinguishing neurons with positive weight ($\tilde{w}_n > 0$, *plus* neurons) and negative weight ($\tilde{w}_n < 0$, *minus* neurons). The sign of the weight is determined with respect to the variable *stimulus* + *choice* (utilizing conditions correct match and correct non-match). To determine the sign as accurately as possible, we compute the population vector utilizing all available trials from conditions correct match and correct non-match.

Strength of the weight We distinguish neurons with strong and weak weights by ranking the absolute value of the weight, $|\tilde{w}_n|$, among the distribution of weights of models with permuted class labels. The strength of the weight of a single neuron n is a scalar, and the same result for models with permuted class labels is a distribution of N_{perm} values, where N_{perm} is the number of random permutations of class labels. If the strength of the weight of the neuron n is ranked within the first 25 % of weights from the null model, we assume that the neuron n has strong weight, and we assume it has weak weight otherwise. Similarly as with the sign of the weight, the strength of the weight is computed using conditions correct match and correct non-match and utilizing all trials from these conditions.

Burstiness We distinguish bursting and non-bursting neurons utilizing a criterion based on the power spectrum of spike trains. It has been shown that bursting neurons have decreased power spectrum in low and middle frequency ranges (Compte et al. 2003). We compute the power spectrum of spike trains for every single neuron, using multiplication of the spike train with Slepian tapers to increase the reliability of the estimation (Pesaran et al. 2002). We use 5 Slepian tapers and the time window of $K = 400$ ms. Power spectra are normalized with neuron's firing rate. We then compute the normalized sum of the power spectrum for frequencies between 10 and 200 Hz, discarding frequencies below 10 Hz due to the short time window K that prevents reliable estimation at low frequencies. As a reference, homogeneous Poisson process has a constant power spectrum of 1, yielding a normalized sum under the power spectrum of 1, while the sum under the power spectrum of a bursting neuron is typically below 1.

Significance of the result is estimated with the permutation test, comparing the sum under the power spectrum of the regular model with the distribution of same results for the null model. The null model is computed by following the same procedure as for the regular model, but randomly permuting, without repetition, the time index of spike trains. We iterate the null model $N_{perm}-$ times and rank the result of the regular model among the distribution of the same results given by null models. We assume that the neuron n is bursting if the normalized sum under its power spectrum is significantly below the same result of models with permutation ($\alpha = 0.05$). If the neuron does not fulfill this criterion, it is assumed to be non-bursting.

Cortical layers We distinguish three cortical layers, the superficial (supragranular, SG), the middle (granular, G) and the deep layer (infragranular, IG) (Hansen et al. 2012). The method for determining cortical layers utilizes the covariance matrix of the current source density and has been published previously (Figure 6 and methods in (Koren et al. 2019)).

The choice signal in subpopulations

Sign-specific choice signal is computed by removing the information from the spike train of neurons with the opposite sign. This method ensures that the imbalance in the number of *plus* and *minus* neurons does not influence the population signal. We remove the information from one of the groups by randomly permuting class labels "same" and "different" in the reconstruction step. When the class label in the reconstruction step is permuted, the association between the weight and the class label of the spike train is random, resulting in a signal that is close to zero at all times. As an example, the choice signal of *plus* neurons is computed with the spike train $\mathbf{o}_{j,p}^{perm,+}(t_k)$, where the label of the spike train is correct for *plus* neurons, and random (i.e., correct

or incorrect with equal probability) for *minus* neurons.

$$x^{+}_{j,p}(t_k) = F(\tilde{\mathbf{w}}^T \mathbf{o}^{perm,+}_j(t_k)) - \langle F(\tilde{\mathbf{w}}^T \mathbf{o}^{perm,+}_j(t_k))\rangle_j \tag{18}$$

Similarly, the signal of *minus* neurons is computed by utilizing the spike train $\mathbf{o}^{perm,-}_{j,p}(t_k)$, where the label of the spike train is correct for *minus* neurons, and random for *plus* neurons.

$$x^{-}_{j,p}(t_k) = F(\tilde{\mathbf{w}}^T \mathbf{o}^{perm,-}_j(t_k)) - \langle F(\tilde{\mathbf{w}}^T \mathbf{o}^{perm,-}_j(t_k))\rangle_j \tag{19}$$

Random permutation is repeated $N_{perm}-$times, with $p = 1, 2, ..., N_{perm}$ random permutations, without repetition, of the order of trials. Next, we average each of the signals across trials, distinguishing conditions "same" and "different".

$$\begin{aligned} x^{+,\text{same}}_p(t_k) &= \frac{2}{J_2 - J_1} \sum_{j=J'}^{J_2} x^{+}_{j,p}(t_k) \\ x^{+,\text{different}}_p(t_k) &= \frac{1}{J - J_2} \sum_{j=J_2+1}^{J} x^{+}_{j,p}(t_k) \end{aligned} \tag{20}$$

Same follows for *minus* neurons. The signal is then averaged across permutations, getting $x^{+,\text{same}}(t_k)$ and $x^{+,\text{different}}(t_k)$ as the signal for *plus* subnetwork and $x^{-,\text{same}}(t_k)$ and $x^{-,\text{different}}(t_k)$ as the signal for the *minus* subnetwork. The significance is evaluated with the permutation test. The test statistic is the sign-specific difference of signals in conditions "same" and "different".

$$\begin{aligned} \Delta x^{+}(t_k) &= x^{+,\text{same}}(t_k) - x^{+,\text{different}}(t_k) \\ \Delta x^{-}(t_k) &= x^{-,\text{same}}(t_k) - x^{-,\text{different}}(t_k) \end{aligned} \tag{21}$$

The null model is computed with the random permutation of class labels 1) when training the classification model and 2) in the reconstruction step. In addition, we use a random assignment to the class of *plus* and *minus* neurons by randomly permuting neural indexes.

The same methods is used to compute the choice signal for subnetworks of neurons with strong and weak weights, for bursting and non-bursting subnetworks and in cortical layers.

Correlation function between the choice signals

The correlation function between the choice signals of *plus* and *minus* subnetworks in trial j and for the permutation instance p is defined as follows:

$$R^{+-}_{j,p}(\tau) = \begin{cases} \displaystyle\sum_{k=0}^{K-\tau-1} x^{+}_{j,p}(t_k) x^{-}_{j,p}(t_k + \tau) & \text{for } \tau \geq 0 \\ \\ R^{-+}_{j}(-\tau) & \text{for } \tau < 0 \end{cases} \tag{22}$$

with time lag $\tau = 1, 2, ..., 2K - 1$. The correlation function is normalized with autocorrelation functions at zero time lag,

$$\tilde{R}^{+-}_{j,p}(\tau) = \frac{R^{+-}_{j,p}(\tau)}{R^{--}_{j,p}(0) R^{++}_{j,p}(0)} \tag{23}$$

where R^{++} (R^{--}) is the autocorrelation function for *plus* (*minus*) neurons.

$$R^{++}_{j,p}(\tau) = \begin{cases} \sum_{k=0}^{K-\tau-1} x^{+}_{j,p}(t_k) x^{+}_{j,p}(t_k+\tau), & \text{for } \tau \geq 0 \\ \\ R^{++}_{j,p}(-\tau) & \text{for } \tau < 0 \end{cases} \tag{24}$$

The correlation function is computed in single trials and then averaged across trials and across permutations. Since there was no difference in the correlation across conditions, we used all trials from the reconstruction set (conditions CNM and INM, trials with index $j = J_1 + 1, ..., J$).

$$\tilde{R}^{+-}(\tau) = \frac{1}{N_{perm}(J - J_1)} \sum_{p=1}^{N_{perm}} \sum_{j=J_1+1}^{J} \tilde{R}^{+-}_{j,p}(\tau) \tag{25}$$

The significance of the correlation function is estimated with the permutation test. We compute the choice signal with random weights and random class label of spike trains. In addition, we use a random assignment to the group of *plus* and *minus* neurons as we compute the correlation function. The same method is used to compute the correlation function between subnetworks with strong and weak weights, and between bursting and non-bursting subnetworks.

Correlation function of choice signals in cortical layers

Similarly, we compute the cross-correlation of choice signals between pairs of cortical layers. The correlation function is computed for the choice signals from two cortical layers,

$$R^{c_1 c_2}_{j,p}(\tau) = \begin{cases} \sum_{k=0}^{K-\tau-1} x^{c_1}_{j,p}(t_k) x^{c_2}_{j,p}(t_k+\tau), & \text{for } \tau \geq 0 \\ \\ R^{c_2 c_1}_{j,p}(-\tau), & \text{for } \tau < 0 \end{cases} \tag{26}$$

with $(c_1, c_2) \in \{(SG, G), (SG, IG), (G, IG)\}$. The rest of the procedure is the same as for *plus* and *minus* neurons. The significance of results is evaluated with the permutation test, where signals are computed with random weights and random class of the spike train. During the computation of the correlation function, the null model in addition uses random assignment to one of the three cortical layers.

Comparison of population vectors across informational contexts of *stimulus+choice* and *choice*

We compare the population vectors in the context of *stimulus+choice* and in the context of *choice*. We train the linear SVM and compute the population vector in each of the two contexts, getting two vectors, $\tilde{\mathbf{w}}^{S+C}$ and $\tilde{\mathbf{w}}^{C}$, in each recording session. Decoding model in the context of *stimulus+choice* utilizes trials from conditions CM and CNM, while decoding model in the context of *choice* utilizes trials from conditions CNM and INM. The number of trials is imbalanced across conditions (namely, there are less trials in condition INM compared to CM), and such imbalance can affect the population vector. We balance the number of trials with the bootstrap method. In each recording session, we find the number of trials of the condition with most trials, and randomly sample, with repetition, the same number of trials from the two other distributions. All reported results are averaged across bootstraps.

We measure the similarity of the two population vectors by computing the angle between them,

$$\alpha = \arccos\left(\tilde{\mathbf{w}}^{S+C} \cdot \tilde{\mathbf{w}}^{C}\right) \tag{27}$$

where $(\cdot)$ is the dot product between the two vectors. Notice that, since vectors are normalized, we have

that $||\tilde{\mathbf{w}}^{S+C}|| = ||\tilde{\mathbf{w}}^{C}|| = 1$. If the vectors $\tilde{\mathbf{w}}^{S+C}$ and $\tilde{\mathbf{w}}^{C}$ are similar, they point in similar direction, and the angle between them is small. If, conversely, the two vectors are pointing in random directions between 0 and π, the angle between them is, on average, orthogonal (the average is across bootstrapped samples). The significance of the angle is evaluated with the permutation test, using the test statistics of the angle, averaged across recording sessions. To construct the null model, we draw random vectors from the uniform distribution that have the same range as the true population vectors, and compute the angle between the two random vectors, $\alpha_p = \arccos(\tilde{\mathbf{w}}_p^{S+C} \cdot \tilde{\mathbf{w}}_p^{C}),\ p = 1, ..., N_{perm}$. The p-value is computed by ranking the angle of the true model among the distribution of angles of N_{perm} models with random weights. The test is significant if p-val $< {}^{0.05}/_{N_{test}}$, where the division with the number of tests implements the Bonferroni correction for multiple testing.

Univariate decoding weights

To test decoding with univariate weights, we replace population decoding weights (eqs. 9 - 12) with weights that are computed from the area under the ROC curve of single neurons (Britten et al. 1996). The area under the ROC curve ($AUROC$) is computed from the distribution of spike counts for each neuron. From the $AUROC$, we subtract the baseline, $w_n^{\mathrm{ROC}} = AUROC_n - 0.5$, and normalize the weight across neurons from the same recording session. Collecting weights of neurons from the same recording session, we get the following weight vector:

$$\tilde{\mathbf{w}}^{\mathrm{ROC}} = \frac{\mathbf{w}^{\mathrm{ROC}}}{||\mathbf{w}^{\mathrm{ROC}}||}, \tag{28}$$

where $||\mathbf{w}^{\mathrm{ROC}}||$ is the L2 norm of weights from the same recording session (see eq. 12). By subtracting the baseline, we get positive and negative weights. Neurons with positive weights are those with increased spike count in condition "correct match"($CP > 0.5$), while neurons with negative weights increase the spike count in condition "non-match" ($CP < 0.5$). The normalization step is done to be able to compare the multivariate method (SVM weights) with the univariate method (ROC weights).

Noise correlation of spike timing and synchrony

Let $n, m \in \{1, ..., N\}$ be fixed indices of two neurons, $n \neq m$. We define the spike trains of neurons n and m in trial j by $f_j(t_k) := o_{n,j}(t_k)$ and $g_j(t_k) := o_{m,j}(t_k)$. The cross-correlation function measures the co-occurence of spikes:

$$ccg_j(\tau) = \frac{1}{2K} \sum_{t_k=-K}^{K} f_j(t_k) g_j(t_k + \tau). \tag{29}$$

To make the cross-correlation function independent of firing rates, we normalize it with autocorrelation functions of neurons n and m at zero time lag:

$$ccg_j^{norm}(\tau) = \frac{ccg_j(\tau)}{\sqrt{acg_{n,j}(0) acg_{m,j}(0)}}, \tag{30}$$

where the autocorrelation function is computed as follows:

$$acg_{n,j}(\tau) = \frac{1}{2K} \sum_{t_k=-K}^{K} f_j(t_k) f_j(t_k + \tau). \tag{31}$$

To measure the noise correlation, we now subtract the co-occuring spikes that are attributed to trial-invariant processes:

$$ccg^{noise}(\tau) = \frac{1}{J} \sum_{j=1}^{J} (ccg_j^{norm}(\tau) - ccg_j^{inv}(\tau)). \tag{32}$$

Trial-invariant correlation is computed by randomly permuting the trial index, without repetition, of one of the neurons,

$$ccg^{inv}_{j,p}(\tau) = \frac{1}{2K} \sum_{t_k=-K}^{K} f_{j'}(t_k) g_j(t_k + \tau),\ j' \neq j, \tag{33}$$

and $ccg^{inv}_j(\tau)$ is the average of N_{perm} instances of $ccg^{inv}_{j,p}(\tau)$. Note that every instance of $ccg^{inv}_j(\tau)$ utilizes an independent random permutation of the trial order. The synchrony is defined as the noise correlation at zero time lag, $C = ccg^{noise}(0)$.

To compute noise correlations and synchrony, we use spike trains at the resolution of experimental recordings (1 ms) and do not apply any further binning. When noise correlations are measured within a subpopulation, the entire procedure (eqs. 29-33) applies to neuronal pairs from the specific subpopulation. The subpopulations are formed taking into account decoding weights from the classification problem *stimulus + choice*. The correlation function, meanwhile, is computed for neural activity corresponding to the variable *choice*.

Acknowledgements

V.K. was supported by Deutsche Forschungsgemeinschaft, grant GRK 1589/2, and Berlin Equal Opportunity Programm (Berliner Chancengleichheitsprogramm).

Author contributions

Conceptualization: V.K. Methodology: V.K., A.A., M.H., V.D., K.O.; Software: V.K.; Validation: V.K.; Formal Analysis: V.K.; Investigation: A.A., M.H.; Resources: V.D., K.O.; Data curation: A.A. and V.K.; Writing-Original Draft: V.K.; Writing-Review and Editing: A.A., V.D., K.O.; Visualization: V.K.; Supervision: V.D., K.O.; Project Administration: V.K., A.A.; Funding Acquisition: K.O.

Declaration of interest

The authors declare no competing interests.

References

Abbott, L. F. and Nelson, S. B. (2000). Synaptic plasticity: taming the beast. *Nature neuroscience*, 3(11):1178–1183.

Andrei, A. R., Pojoga, S., Janz, R., and Dragoi, V. (2019). Integration of cortical population signals for visual perception. *Nature communications*, 10(1):1–13.

Averbeck, B. B., Latham, P. E., and Pouget, A. (2006). Neural correlations, population coding and computation. *Nature reviews neuroscience*, 7(5):358–366.

Bair, W., Zohary, E., and Newsome, W. T. (2001). Correlated firing in macaque visual area mt: time scales and relationship to behavior. *Journal of Neuroscience*, 21(5):1676–1697.

Beck, J. M., Ma, W. J., Pitkow, X., Latham, P. E., and Pouget, A. (2012). Not noisy, just wrong: the role of suboptimal inference in behavioral variability. *Neuron*, 74(1):30–39.

Belousov, A., Verzakov, S., and Von Frese, J. (2002). A flexible classification approach with optimal generalisation performance: support vector machines. *Chemometrics and intelligent laboratory systems*, 64(1):15–25.

Bernardi, S., Benna, M. K., Rigotti, M., Munuera, J., Fusi, S., and Salzman, C. D. (2020). The geometry of abstraction in the hippocampus and prefrontal cortex. *Cell*, 183(4):954–967.

Blainey, P., Krzywinski, M., and Altman, N. (2014). Points of significance: replication. *Nature methods*, 11(9):879.

Boerlin, M., Machens, C. K., and Denève, S. (2013). Predictive coding of dynamical variables in balanced spiking networks. *PLoS Comput Biol*, 9(11):e1003258.

Bondy, A. G., Haefner, R. M., and Cumming, B. G. (2018). Feedback determines the structure of correlated variability in primary visual cortex. *Nature neuroscience*, 21(4):598–606.

Britten, K. H., Newsome, W. T., Shadlen, M. N., Celebrini, S., Movshon, J. A., et al. (1996). A relationship between behavioral choice and the visual responses of neurons in macaque mt. *Visual neuroscience*, 13(1):87–100.

Brunton, B. W., Botvinick, M. M., and Brody, C. D. (2013). Rats and humans can optimally accumulate evidence for decision-making. *Science*, 340(6128):95–98.

Callaway, E. M. (2004). Feedforward, feedback and inhibitory connections in primate visual cortex. *Neural Networks*, 17(5-6):625–632.

Churchland, M. M., Cunningham, J. P., Kaufman, M. T., Foster, J. D., Nuyujukian, P., Ryu, S. I., and Shenoy, K. V. (2012). Neural population dynamics during reaching. *Nature*, 487(7405):51–56.

Cohen, M. R. and Maunsell, J. H. (2009). Attention improves performance primarily by reducing interneuronal correlations. *Nature neuroscience*, 12(12):1594.

Compte, A., Constantinidis, C., Tegnér, J., Raghavachari, S., Chafee, M. V., Goldman-Rakic, P. S., and Wang, X.-J. (2003). Temporally irregular mnemonic persistent activity in prefrontal neurons of monkeys during a delayed response task. *Journal of neurophysiology*, 90(5):3441–3454.

Cumming, B. G. and Nienborg, H. (2016). Feedforward and feedback sources of choice probability in neural population responses. *Current opinion in neurobiology*, 37:126–132.

Engel, T. A. and Wang, X.-J. (2011). Same or different? a neural circuit mechanism of similarity-based pattern match decision making. *Journal of Neuroscience*, 31(19):6982–6996.

Gao, P. and Ganguli, S. (2015). On simplicity and complexity in the brave new world of large-scale neuroscience. *Current opinion in neurobiology*, 32:148–155.

Goris, R. L., Ziemba, C. M., Stine, G. M., Simoncelli, E. P., and Movshon, J. A. (2017). Dissociation of choice formation and choice-correlated activity in macaque visual cortex. *Journal of Neuroscience*, 37(20):5195–5203.

Hansen, B. J., Chelaru, M. I., and Dragoi, V. (2012). Correlated variability in laminar cortical circuits. *Neuron*, 76(3):590–602.

Harris, K. D., Bartho, P., Chadderton, P., Curto, C., de la Rocha, J., Hollender, L., Itskov, V., Luczak, A., Marguet, S. L., Renart, A., et al. (2011). How do neurons work together? lessons from auditory cortex. *Hearing research*, 271(1-2):37–53.

Hupé, J., James, A., Payne, B., Lomber, S., Girard, P., and Bullier, J. (1998). Cortical feedback improves discrimination between figure and background by v1, v2 and v3 neurons. *Nature*, 394(6695):784–787.

Jasper, A. I., Tanabe, S., and Kohn, A. (2019). Predicting perceptual decisions using visual cortical population responses and choice history. *Journal of Neuroscience*, 39(34):6714–6727.

Juan, C.-H. and Walsh, V. (2003). Feedback to v1: a reverse hierarchy in vision. *Experimental brain research*, 150(2):259–263.

Koren, V. (2021). Uncovering structured responses of neural populations recorded from macaque monkeys with linear support vector machines. *STAR protocols*, 2(3):100746.

Koren, V., Andrei, A. R., Hu, M., Dragoi, V., and Obermayer, K. (2019). Reading-out task variables as a low-dimensional reconstruction of neural spike trains in single trials. *PloS one*, 14(10):e0222649.

Koren, V., Andrei, A. R., Hu, M., Dragoi, V., and Obermayer, K. (2020). Pairwise synchrony and correlations depend on the structure of the population code in visual cortex. *Cell Reports*, 33(6):108367.

Krishna, A., Tanabe, S., and Kohn, A. (2021). Decision signals in the local field potentials of early and mid-level macaque visual cortex. *Cerebral Cortex*, 31(1):169–183.

Larkum, M. (2013). A cellular mechanism for cortical associations: an organizing principle for the cerebral cortex. *Trends in neurosciences*, 36(3):141–151.

Lisman, J. E. (1997). Bursts as a unit of neural information: making unreliable synapses reliable. *Trends in neurosciences*, 20(1):38–43.

Logothetis, N. K. and Schall, J. D. (1989). Neuronal correlates of subjective visual perception. *Science*, 245(4919):761–763.

Luczak, A., Bartho, P., and Harris, K. D. (2013). Gating of sensory input by spontaneous cortical activity. *Journal of Neuroscience*, 33(4):1684–1695.

Muckli, L., De Martino, F., Vizioli, L., Petro, L. S., Smith, F. W., Ugurbil, K., Goebel, R., and Yacoub, E. (2015). Contextual feedback to superficial layers of v1. *Current Biology*, 25(20):2690–2695.

Naud, R. and Sprekeler, H. (2018). Sparse bursts optimize information transmission in a multiplexed neural code. *Proceedings of the National Academy of Sciences*, 115(27):E6329–E6338.

Ni, A. M., Ruff, D. A., Alberts, J. J., Symmonds, J., and Cohen, M. R. (2018). Learning and attention reveal a general relationship between population activity and behavior. *Science*, 359(6374):463–465.

Nienborg, H. and Cumming, B. G. (2006). Macaque v2 neurons, but not v1 neurons, show choice-related activity. *Journal of Neuroscience*, 26(37):9567–9578.

Nienborg, H. and Cumming, B. G. (2009). Decision-related activity in sensory neurons reflects more than a neuron's causal effect. *Nature*, 459(7243):89–92.

Nienborg, H. and Cumming, B. G. (2014). Decision-related activity in sensory neurons may depend on the columnar architecture of cerebral cortex. *Journal of Neuroscience*, 34(10):3579–3585.

Nienborg, H., R. Cohen, M., and Cumming, B. G. (2012). Decision-related activity in sensory neurons: correlations among neurons and with behavior. *Annual review of neuroscience*, 35:463–483.

Onorato, I., Neuenschwander, S., Hoy, J., Lima, B., Rocha, K.-S., Broggini, A. C., Uran, C., Spyropoulos, G., Klon-Lipok, J., Womelsdorf, T., et al. (2020). A distinct class of bursting neurons with strong gamma synchronization and stimulus selectivity in monkey v1. *Neuron*, 105(1):180–197.

Panzeri, S., Harvey, C. D., Piasini, E., Latham, P. E., and Fellin, T. (2017). Cracking the neural code for sensory perception by combining statistics, intervention, and behavior. *Neuron*, 93(3):491–507.

Pesaran, B., Pezaris, J. S., Sahani, M., Mitra, P. P., and Andersen, R. A. (2002). Temporal structure in neuronal activity during working memory in macaque parietal cortex. *Nature neuroscience*, 5(8):805–811.

Ren, S.-Q., Li, Z., Lin, S., Bergami, M., and Shi, S.-H. (2019). Precise long-range microcircuit-to-microcircuit communication connects the frontal and sensory cortices in the mammalian brain. *Neuron*, 104(2):385–401.

Rigotti, M., Barak, O., Warden, M. R., Wang, X.-J., Daw, N. D., Miller, E. K., and Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*, 497(7451):585–590.

Roitman, J. D. and Shadlen, M. N. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. *Journal of neuroscience*, 22(21):9475–9489.

Saxena, S. and Cunningham, J. P. (2019). Towards the neural population doctrine. *Current opinion in neurobiology*, 55:103–111.

Seidemann, E. and Geisler, W. S. (2018). Linking v1 activity to behavior. *Annual review of vision science*, 4:287–310.

Shadlen, M. N., Britten, K. H., Newsome, W. T., and Movshon, J. A. (1996). A computational analysis of the relationship between neuronal and behavioral responses to visual motion. *Journal of Neuroscience*, 16(4):1486–1510.

Shadlen, M. N. and Newsome, W. T. (1996). Motion perception: seeing and deciding. *Proceedings of the national academy of sciences*, 93(2):628–633.

Shadlen, M. N. and Newsome, W. T. (1998). The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. *Journal of neuroscience*, 18(10):3870–3896.

Shahidi, N., Andrei, A. R., Hu, M., and Dragoi, V. (2019). High-order coordination of cortical spiking activity modulates perceptual accuracy. *Nature neuroscience*, 22(7):1148–1158.

Silvanto, J., Cowey, A., Lavie, N., and Walsh, V. (2005). Striate cortex (v1) activity gates awareness of motion. *Nature neuroscience*, 8(2):143–144.

Spoerer, C. J., Kietzmann, T. C., Mehrer, J., Charest, I., and Kriegeskorte, N. (2020). Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. *PLoS computational biology*, 16(10):e1008215.

Steinmetz, N. A., Zatka-Haas, P., Carandini, M., and Harris, K. D. (2019). Distributed coding of choice, action and engagement across the mouse brain. *Nature*, 576(7786):266–273.

Stringer, C., Pachitariu, M., Steinmetz, N., Carandini, M., and Harris, K. D. (2019a). High-dimensional geometry of population responses in visual cortex. *Nature*, 571(7765):361–365.

Stringer, C., Pachitariu, M., Steinmetz, N., Reddy, C. B., Carandini, M., and Harris, K. D. (2019b). Spontaneous behaviors drive multidimensional, brainwide activity. *Science*, 364(6437).

Uka, T. and DeAngelis, G. C. (2004). Contribution of area mt to stereoscopic depth perception: choice-related response modulations reflect task strategy. *Neuron*, 42(2):297–310.

Wang, X.-J. (2002). Probabilistic decision making by slow reverberation in cortical circuits. *Neuron*, 36(5):955–968.

Yang, H., Kwon, S. E., Severson, K. S., and O'connor, D. H. (2016). Origins of choice-related activity in mouse somatosensory cortex. *Nature neuroscience*, 19(1):127–134.

Zhao, Y., Yates, J. L., Levi, A. J., Huk, A. C., and Park, I. M. (2020). Stimulus-choice (mis) alignment in primate area mt. *PLoS computational biology*, 16(5):e1007614.

Zipser, K., Lamme, V. A., and Schiller, P. H. (1996). Contextual modulation in primary visual cortex. *Journal of Neuroscience*, 16(22):7376–7389.

Zylberberg, J., Pouget, A., Latham, P. E., and Shea-Brown, E. (2017). Robust information propagation through noisy neural circuits. *PLoS computational biology*, 13(4):e1005497.