bioRxiv preprint doi: https://doi.org/10.1101/2022.12.07.519040; this version posted December 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Redundant representations are required to disambiguate simultaneously presented complex stimuli

W. Jeffrey Johnston[1,2,3,*] and David J. Freedman[1,2]

[1]Graduate Program in Computational Neuroscience
[2]Department of Neurobiology
The University of Chicago, Chicago, IL, USA
[3]Center for Theoretical Neuroscience
Columbia University, New York, NY, USA

[*]Corresponding author: wjeffreyjohnston@gmail.com

December 7, 2022

Abstract

An individual observing a barking dog and purring cat together in a field has distinct pairs of representations of the two animals in their visual and auditory systems. Without prior knowledge, how does the observer infer that the dog barks and the cat purrs? This integration of distributed representations is called the assignment problem, and it must be solved to integrate distinct representations across but also within sensory modalities. Here, we identify and analyze a solution to the assignment problem: the representation of one or more common stimulus features in pairs of relevant brain regions – for example, estimates of the spatial position of both the cat and the dog represented in both the visual and auditory systems. We characterize how the reliability of this solution depends on different features of the stimulus set (e.g., the size of the set and the complexity of the stimuli) and the details of the split representations (e.g., the precision of each stimulus representation and the amount of overlapping information). Next, we implement this solution in a biologically plausible receptive field code and show how constraints on the number of neurons and spikes used by the code force the brain to navigate a tradeoff between local and catastrophic errors. We show that, when many spikes and neurons are available, representing stimuli from a single sensory modality can be done more reliably across multiple brain regions, despite the risk of assignment errors. Finally, we show that a feedforward neural network can learn the optimal solution to the assignment problem. We also discuss relevant results on assignment errors from the human working memory literature and show that several key predictions of our theory already have support.

Keywords: multisensory integration, representation assignment, distributed representations

1 Introduction

Coherent behavior in complex natural environments requires extensive and reliable integration of different forms of information about the world. For instance, a hawk bringing food to her hatchlings

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.07.519040; this version posted December 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

must attend to both the visual appearance of each hatchling and their vocalizations to infer which of them is most in need of food. While navigating these cluttered multisensory environments often appears effortless for hawks and other species, it requires two highly non-trivial computations: Object segmentation and representation assignment. For object segmentation, the continuous sensory world must be segmented into distinct objects – that is, each hatchling has to be recognized as a distinct entity, separated from both the other hatchlings and other background objects, such as the nest. The primitive rules that humans and other animals use to segment objects have been studied extensively as part of gestalt psychology[1] and related subfields[2, 3], as well as in machine learning[4, 5]. In the brain, this segmentation process is thought to be represented by border ownership cells, such as those observed in the primate visual system[6, 7]. However, even once this difficult object segmentation problem is solved, the resulting object representations are often still incomplete: An object representation in one brain region contains only a part of all the information about that object that is present in the entire brain. The integration of these distributed representations is referred to as representation assignment[8] – and is analogous to the extensively studied assignment problem from combinatorial optimization[9]. In the example above, the representations of the chicks from the visual system need to be integrated with the representations of the chicks from the auditory system, as both sensory systems provide information that is necessary to guide coherent behavior.

The integration of these distinct, parallel representations of the world has been previously studied in three principle ways. First, the integration of distinct features within the visual system has been studied in the context of feature integration theory[10, 11]. In feature integration theory, spatial attention is used to bind features together due to their spatial proximity. Spatial attention is deployed to different locations in sequence and only a single object is bound at a time. However, experimental work has shown that errors can occur in this integration process[12, 13], in which some of the features from two different locations are assigned to each other. We refer to these kinds of errors as an assignment errors, though they are also referred to as illusory conjunctions, swap errors, and misbinding errors. Assignment errors produce representations – such as a barking cat and meowing dog, see Figure 1a – that can be particularly catastrophic for coherent behavior. While feature integration theory provides a qualitative description of the assignment process using a common representation of spatial position to link different feature representations together, it does not provide a mechanistic explanation of how assignment errors depend on the qualities of the underlying neural representations.

Second, related to feature integration theory, the assignment problem has been studied as part of the literature on working memory capacity. Here, participants are asked to remember an array of multi-feature stimuli (e.g., oriented and colored bars at different spatial locations) across a delay period, before being cued by a single feature (e.g., position) to report one of the other features (e.g., color)[14–17]. In this setting, the cue must be assigned to the correct representation from the set of remembered stimuli. Here, the assignment problem is solved across two distinct time points rather than two distinct brain regions, but the underlying principles are the same. This literature shows that assignment errors increase with the number of stimuli that need to be remembered[15, 16] and that they become more likely when remembered stimuli are close together in the feature that is used to cue a stimulus for report (from above, spatial position)[16, 17]. The assignment error rate also depends on the precision of the representation of the cued feature[17]. Further, a neural model has been developed to explain these behavioral findings, which links the assignment error rate to neural population codes[16]. Our work builds on the theoretical work from the time-assignment setting by considering multiple common (i.e., cue) features[18] (but see [19]), as well as through its

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.07.519040; this version posted December 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

focus on assignment across (and within, as in feature integration theory) distinct brain regions at a single time point.

Third, representation assignment has also been studied in the context of multisensory integration[8, 20–23]. In this literature, both experiment and theory have primarily focused on the integration of multisensory representations of a single stimulus. This work has demonstrated that shared information is crucial to reliable integration. In particular, integration of auditory and visual information appears to rely on the shared representation of azimuthal position derived from both sensory systems[21]. That is, whether or not an auditory and visual stimulus are integrated depends on the mean squared-error (MSE) of the representations of azimuthal position in both sensory systems[21]. We refer to the MSE of these individual representations as the local MSE.

Here, we extend this analysis to sets of multiple objects and develop a general framework that demonstrates how the likelihood of assignment errors depends on the number of common features that are represented across multiple brain regions (such as azimuthal position) and the local MSE of those feature representations. We show that there is a tradeoff between highly redundant representations that produce low assignment error rates, but higher local MSE, and less redundant representations that risk more assignment errors but have generally lower local MSE. To make this tradeoff concrete, we analyze a receptive field model of neural representations, similar to those found in many sensory areas[24–28] and related to models explored in the working memory literature[16]. With this model and when keeping constant the total metabolic resources (i.e., number of neurons and spikes) used by the model, we show that the lowest total error (combined local and assignment errors) can sometimes be achieved by systems that split into distinct modules, despite the fact that such systems risk assignment errors. Thus, our framework provides a potential explanation for the extensive modularity and parallel processing thought to exist in the primate visual system[29–32] and other sensory systems[33–36]. Finally, we demonstrate that the assignment problem can be solved optimally by a feedforward neural network. We also discuss predictions for neural data, and link our results in more detail to behavioral results from the human working memory literature. Overall, this work demonstrates a general solution to the representation assignment problem, which arises whenever a neural system represents multiple stimuli – within single brain regions, across parallel brain regions, and across different time points.

2 Results

2.1 Redundant information is necessary to solve the assignment problem

Here, we investigate the assignment problem that arises when each of N stimuli have representations that are split across two brain regions, R_X and R_Y – in combinatorial optimization, this is referred to as the balanced assignment problem[9]. We want to correctly integrate the distributed representations of these N stimuli, which we formalize as finding the correct one-to-one mapping between the representations in R_X and the representations in R_Y. To give an example: a dog and a cat will be represented in at least two distinct regions of the brain, an auditory region R_X and a visual region R_Y (Figure 1a). The set of representations in R_X, referred to as $\hat{X}$, will consist of representations of each of the animals' vocalizations – similarly, $\hat{Y}$ in R_Y consists of representations of each animals' visual features (see *Definition of the representations* in *Methods* for more details). Many behaviors rely on integrated information about both the visual features and vocalizations of

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.07.519040; this version posted December 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

the animals (for instance, when deciding which animal to approach, it is useful to know whether they are barking or purring).

So, we want to choose the correct one-to-one map M between $\hat{X}$ and $\hat{Y}$ (Figure 1b). We show that if $\hat{X}$ and $\hat{Y}$ are independent, then all maps are equally likely (see *Definition of assignment errors* in *Methods*). In this case, assignment errors will occur with probability $1 - 1/N!$. Thus, redundancy between the representations in R_X and R_Y is necessary to reliably solve the assignment problem. More concretely, the case where $\hat{X}$ and $\hat{Y}$ are independent would be like the case where, in a working memory study similar to the ones discussed in the Introduction, we ask the participant to remember the color and orientation of a collection of bars, then cue them with a spatial frequency. Without additional instruction or extensive learning, the participant could do no better than chance – just as our system here could solve the assignment problem no more reliably than by guessing.

However, the required redundancy can take many forms. We will focus on the case where the redundancy manifests as a linear dependence between some of the features represented in R_X and some of the features represented in R_Y. In the working memory task example, this is similar to remembering both the color and orientation of each stimulus in an array. Then, being asked to report the color of the stimulus that had a particular orientation. Our analysis can also be extended to some kinds of nonlinear mappings between between the features in R_X and R_Y (see *Nonlinear mappings between* $\hat{X}$ *and* $\hat{Y}$ in *Supplement* and Figure S2).

One specific kind of nonlinear redundancy between the two regions is prior information about which stimuli are more or less likely. For example, someone familiar with cats and dogs could infer that a purring noise likely belongs to the cat they see in front of them, while the barking likely belongs to the dog. While such inferences are likely extremely important for behavior, they require the prior information to be learned over experience with these particular stimuli. We focus our analyses here on solving the assignment problem without this learning process (as for novel stimuli) and where such differences in prior likelihood are not available (e.g., assigning two different barks to two different novel dogs).

2.2 The assignment error rate depends on the precision of the common feature representation

Here, we study redundancy between R_X and R_Y that results from the representation of one or more of the same stimulus features (e.g., azimuthal position in both the visual and auditory systems, Figure 1c, left). These C common, or linearly related, features define a C-dimensional common feature space. Following the combinatorial assignment problem, we show that the most likely mapping is the one that minimizes the sum squared distance between integrated pairs in this common feature space (Figure 1c, right: top matrix is the the correct mapping, M_{C}, bottom matrix is the incorrect mapping M_{AE}, the sum of distances between integrated pairs in the two mappings are shown below). If the estimates of the common features in both regions are exactly correct, then this sum will be zero – an exact matching (Figure 1c). However, in general, we assume that each feature represented in R_X is decoded with Gaussian-distributed local mean squared error (MSE) D_X (Figure 1d). Further, we show that assignment errors occur precisely when the representation of one object crosses over the representation of another object in one region but not the other (Figure 1d, red lines). For two objects at distance δ from each other in the common feature space

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.07.519040; this version posted December 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

space, the probability of an assignment error between those two stimuli is approximated by

$$F(\delta) \approx Q\left(\frac{-\delta}{\sqrt{2D_X}}\right) + Q\left(\frac{-\delta}{\sqrt{2D_Y}}\right) \tag{1}$$

where $Q(x)$ is the cumulative density function for the standard normal distribution (Figure 1e, left) – for the full expression and derivation, see *Probability of assignment errors* in *Methods*.

The overall probability of an assignment error also relies on the probability that two objects are at distance δ from each other in the common space $P_C(\delta)$ (Figure 1e, right for $C = 1$) and on the number of stimuli N. Incorporating these, we show that the overall probability of an assignment error is upper bounded by,

$$AE_C \leq \binom{N}{2} \int d\delta \, P_C(\delta) \, F(\delta) \tag{2}$$

For one overlapping stimulus feature ($C = 1$) and for stimuli that are uniformly distributed in feature space, we show that,

$$AE_1 \approx \binom{N}{2} \frac{2\sqrt{D_X + D_Y}}{s\sqrt{\pi}} \tag{3}$$

where s is the size of the feature space for the commonly represented feature. This approximation (dashed lines, Figure 1f) closely matches the empirical assignment error rate (solid lines) across different numbers of stimuli (different colors; see *Assignment error rate approximation for* $C = 1$ in *Methods* for a full derivation of this expression). Now, using this formalization, we show that the assignment error rate decreases with additional commonly represented features $C > 1$ in the next section. We also show that the assignment error rate increases if the features are represented with asymmetric MSE across the two brain regions – i.e., if $D_X \neq D_Y$ (see *Asymmetric feature encoding* in *Supplement*).

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.07.519040; this version posted December 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

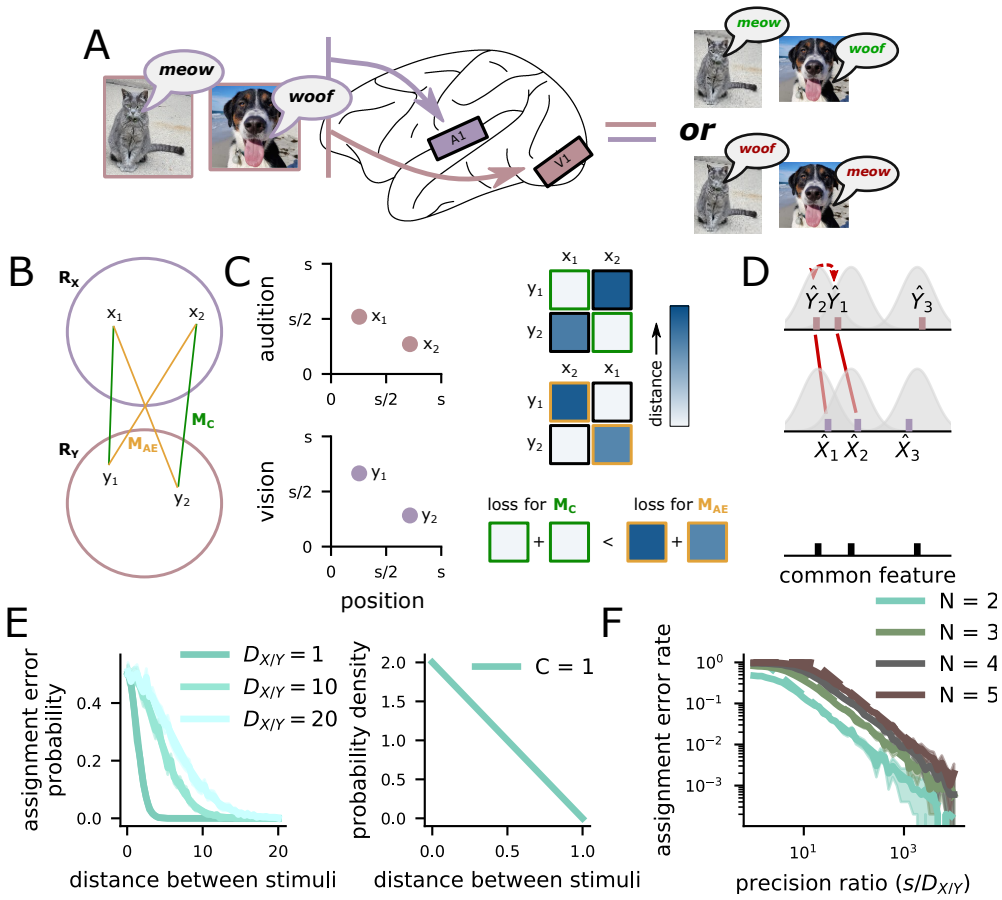

Figure 1: The assignment problem arises from distributed representations of multiple stimuli. **A** A schematic of the assignment problem. The brain receives both visual and auditory information about a dog and a cat, this information is initially separated in the brain. When combining the two sets of representations (auditory and visual), there are two possible integrations, one that correctly reconstructs a barking dog and meowing cat (left) and the other that incorrectly constructs a barking cat and meowing dog (right). The right is an example of an assignment error. **B** Schematic of the assignment process as selecting a mapping between two sets of representations. **C** (left) The assignment problem can be solved by the representation of a common stimulus feature in both brain regions. In our example, the auditory (top) and visual (bottom) representations can be integrated through a shared representation of azimuthal position (the shared x-axis). (right) The $N \times N$ cost matrix (here, each element is given by the distance in the shared feature space across representations from the two regions) for the correct (top) and incorrect (middle) mappings. The most likely assignment minimizes the trace of the matrix (bottom). **D** Assignment errors occur when the estimate of the common feature value for two stimuli cross over each other in one region, but not the other (middle, red arrow). The distribution around each point is the local MSE (D_X and D_Y). **E** (left) The probability that this crossing over occurs is high for nearby stimuli and low for distant stimuli, and increasing estimator variance makes assignment errors more likely at all distances. (right) For stimuli that are uniformly distributed in the full feature space, the distance between pairs of stimuli follows a triangular distribution with one commonly represented feature ($C = 1$). **F** The overall assignment error rate is the product of the two functions on the left – the assignment error probability at each distance weighted by the probability that there is a pair of stimuli at that distance. Dashed line: theory from Eq. 3; solid line: simulations.

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.07.519040; this version posted December 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

2.3 Additional commonly represented features decrease the assignment error rate

While distinct sensory systems tend to have fixed amounts of overlapping information, such as a common estimate of azimuthal position across the auditory and visual systems (which has been shown to be crucial for single stimulus integration[21]), within a single sensory system information is distributed across multiple brain regions, and those brain regions can represent variable amounts of overlapping information about the stimuli. In our framework, we show that increasing the number of commonly represented stimulus features across the two brain regions C decreases the assignment error rate.

In particular, increasing the number of commonly represented features C increases the average distance between pairs of points in the common feature space (Figure 2a). This change is captured by changes in the $p_C(\delta)$ term in Eq. 2. Increasing the number of commonly represented features moves the probability mass of this distribution away from zero (Figure 2b). The increase in the number of overlapping features from C to $C+1$ leads to an orders of magnitude decrease of the assignment error rate at the mean distance (Figure 2c). As a result of this shift to the right in $p_C(\delta)$ and the fact that Eq. 1 is strictly (and exponentially) decreasing in δ, the integral in Eq. 2 will also be strictly decreasing as C is increased. Thus, increasing C decreases the overall probability of assignment errors when holding local MSE constant (Figure 2d).

However, increasing C while holding local MSE constant comes with a significant cost. In particular, increasing the number of commonly represented stimulus features also increases the redundancy in information represented by the two brain regions. As discussed above, it is precisely this redundancy that allows the assignment problem to be solved (and solved more reliably for more overlapping features). Yet, in neural systems with limited metabolic resources, this redundancy reduces the efficiency of neural representations. We show that the redundancy between the two regions depends on both the number of overlapping features and the local MSE (Figure 2e),

$$\begin{aligned} R &= I(X;Y) = H(X) - H(X|Y) \\ &= \frac{C}{2} \log \frac{s^2}{D_X + D_Y} \end{aligned}$$

where $I(.;.)$ is the mutual information and $H(.)$ is the Shannon entropy, and s is the extent of the overlapping features (see *Calculating the redundancy between representations* in *Methods* for more details). Thus, as the assignment error rate is reduced by increased C, the redundancy of the representation is increased – we make the consequences of this explicit below.

Further, we show that a similar tradeoff holds for asymmetric feature representations (*Asymmetric feature encoding* in *Supplement* and Figure S1). On one hand, the common features can be represented with the same precision in both regions. This is the maximally redundant representation for a given C, and therefore leads to the lowest assignment error rate. On the other hand, the common feature can be represented asymmetrically (with high MSE in one region and low MSE in the other). This is a less redundant representation, and is associated with a higher assignment error rate.

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.07.519040; this version posted December 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

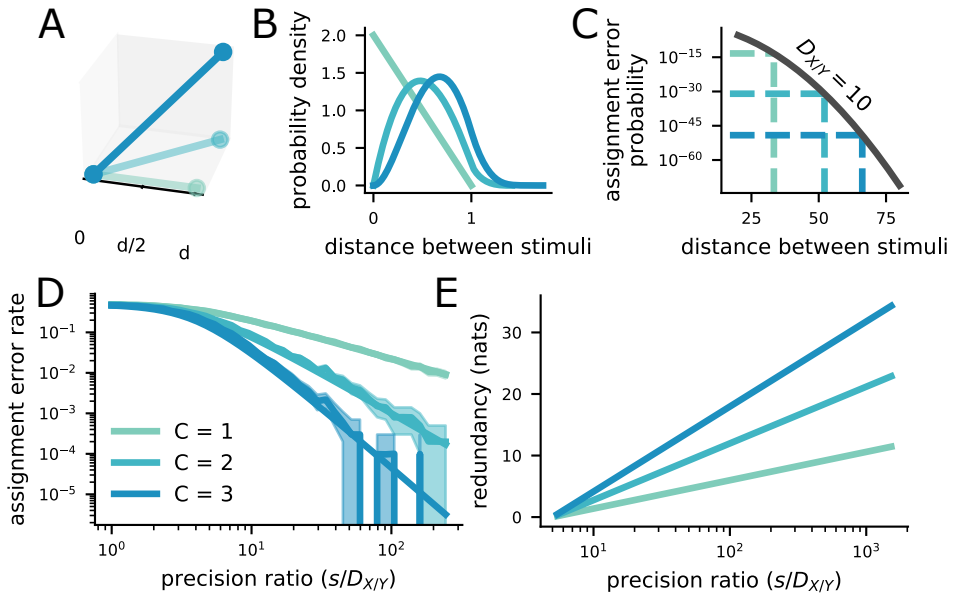

Figure 2: Increasing the number of commonly represented features decreases the assignment error rate, but increases the level of redundancy between the representations. The color legend is the same throughout the plot, and provided in **D**. **A** Schematic showing that the distance between two points increases with the dimensionality of the space that they are in. **B** The distribution of distances between two uniformly distributed points in a space of one, two, and three dimensions. The distribution shifts to the right (toward larger distances) as the dimensionality of the space increases. This is true for points with any distribution that has non-zero variance. **C** At the average distance between two points (dashed lines), the probability of an assignment error decreases by orders of magnitude as the dimensionality of the commonly represented space increases, without changing the estimator variance for the representations $D_{XY} = 10$. **D** The overall assignment error rate also decreases by orders of magnitude as the dimensionality of the commonly represented feature space increases, while holding D_{XY} constant. The difference becomes even larger as the precision ratio increases. **E** The redundancy between representations $\hat{X}$ and $\hat{Y}$ also increases as the dimensionality of the commonly represented feature space increases – again, the difference is increased at larger precision ratios. Thus, the assignment error rate is driven down at the cost of additional redundancy.

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.07.519040; this version posted December 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

2.4 Redundancy reduces the precision of the neural code

Now, we make our theory more concrete, by translating from dealing with local MSE, the assignment error rate, and redundancy to a biologically plausible neural code that has constrained metabolic energy, in the form of spiking energy and the number of neurons in the population. To do this, we analyze the errors made by a population of neurons with randomly positioned Gaussian receptive field (RF) responses (Figure 3a, top). This kind of neural code is thought to be used for a variety of different sensory features[24–28], including spatial position (as we have used as an example so far). Recent work has shown that RF codes make two kinds of errors[37]. The first kind of error is local, and is equivalent to the local MSE we have discussed so far. In this case, the decoder maps the noisy response to a nearby position on the stimulus representation manifold (Figure 3a, bottom, "local error"). This local MSE is captured by the Fisher information of the code[37–43]. The second kind of error is non-local, and is referred to as a threshold (or catastrophic) error[37, 44]. Catastrophic errors occur when the noisy response is mapped to a distant region on the stimulus manifold, due to large noise that points toward one of these regions of response space (Figure 3a, bottom, "threshold error"). In a noisy system, every response will give rise to some local MSE (Figure 3b, "local"); threshold errors only occur for a subset of responses, but their magnitude has the same order as the size of the stimulus space (Figure 3b, "threshold").

These two kinds of error are in tension with each other. To see this, we begin by deriving the average Fisher information of the random RF code,

$$\begin{aligned} \mathrm{FI} &\approx \frac{V}{\sigma_N^2 w^2} \frac{\frac{1}{2}\sqrt{\pi} - w}{\sqrt{\pi} - w} \\ &\approx \frac{V}{2\sigma_N^2 w^2} \\ &= \frac{\mathrm{SNR}^2}{2w^2} \end{aligned} \tag{4}$$

where V is the sum squared activity across the population (i.e., the metabolic cost due to spiking), σ_N is the standard deviation of the Gaussian noise added to the responses, $\mathrm{SNR}^2 = V/\sigma_N^2$ is the population signal-to-noise ratio, and w is the width of the receptive fields, which is assumed to be small (see *Random Gaussian receptive field codes* in *Methods* for the full derivation). We can see that the Fisher information increases when V increases or when w decreases (Figure 3c, local). However, the total MSE of the code is given by,

$$\mathrm{MSE} \approx (1 - p_{\mathrm{thr}}) \frac{1}{\mathrm{FI}} + p_{\mathrm{thr}} \frac{1}{6}$$

where p_{thr} is the probability of a threshold error, and this probability increases as w decreases (Figure 3c, threshold). All else held equal, decreasing the width of the RFs in the code will reduce the local MSE of the code (Figure 3c, local), while simultaneously increasing the probability of threshold errors (Figure 3c, threshold). In the following, we choose the RF width that minimizes the combined MSE of the code (Figure 3c, combined; see *Random Gaussian receptive field codes* in *Methods* for more details).

Using this framework, we show how this total MSE scales with both population signal-to-noise ratio ($\mathrm{SNR} = \sqrt{V}/\sigma_N$). In particular, we show that, for high SNRs, the total MSE is explained primarily by the local MSE (Figure 3d, solid vs dotted line). This results from the fact that the threshold error rate decreases exponentially with SNR (*Random Gaussian receptive field codes* in

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.07.519040; this version posted December 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Methods). Second, we show that increasing the population size while holding the SNR constant leads to a smaller decrease in total MSE (Figure 3e). That is, while increasing the population size pushes smaller RFs to be optimal, thereby decreasing the local MSE, the tension between local and threshold errors causes the decrease in optimal RF size to be small. We also show that, as expected, requiring the code to represent $K = 2$ instead of $K = 1$ features causes higher error rates (Figure 3d, e, different lines).

Finally, to make the cost of redundancy explicit in terms of the local MSE of the code and connect this RF codes more directly to our theory of assignment errors, we analyze the case of $K = 4$ total features represented across two different brain regions, with $C = 1$, 2, and 3 overlapping features. We show that the increase in overlapping features leads to an increase in both the redundancy between the two representations and the local MSE of the code (Figure 3f, left, grey lines show constant SNR). Then, we show that, as anticipated, the increase in the number of overlapping features also leads to a decrease in the assignment error rate at these higher redundancy levels (Figure 3f, right). Thus, this modeling approach allows us to quantify the consequences of different levels of redundancy in a biologically plausible neural code.

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.07.519040; this version posted December 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

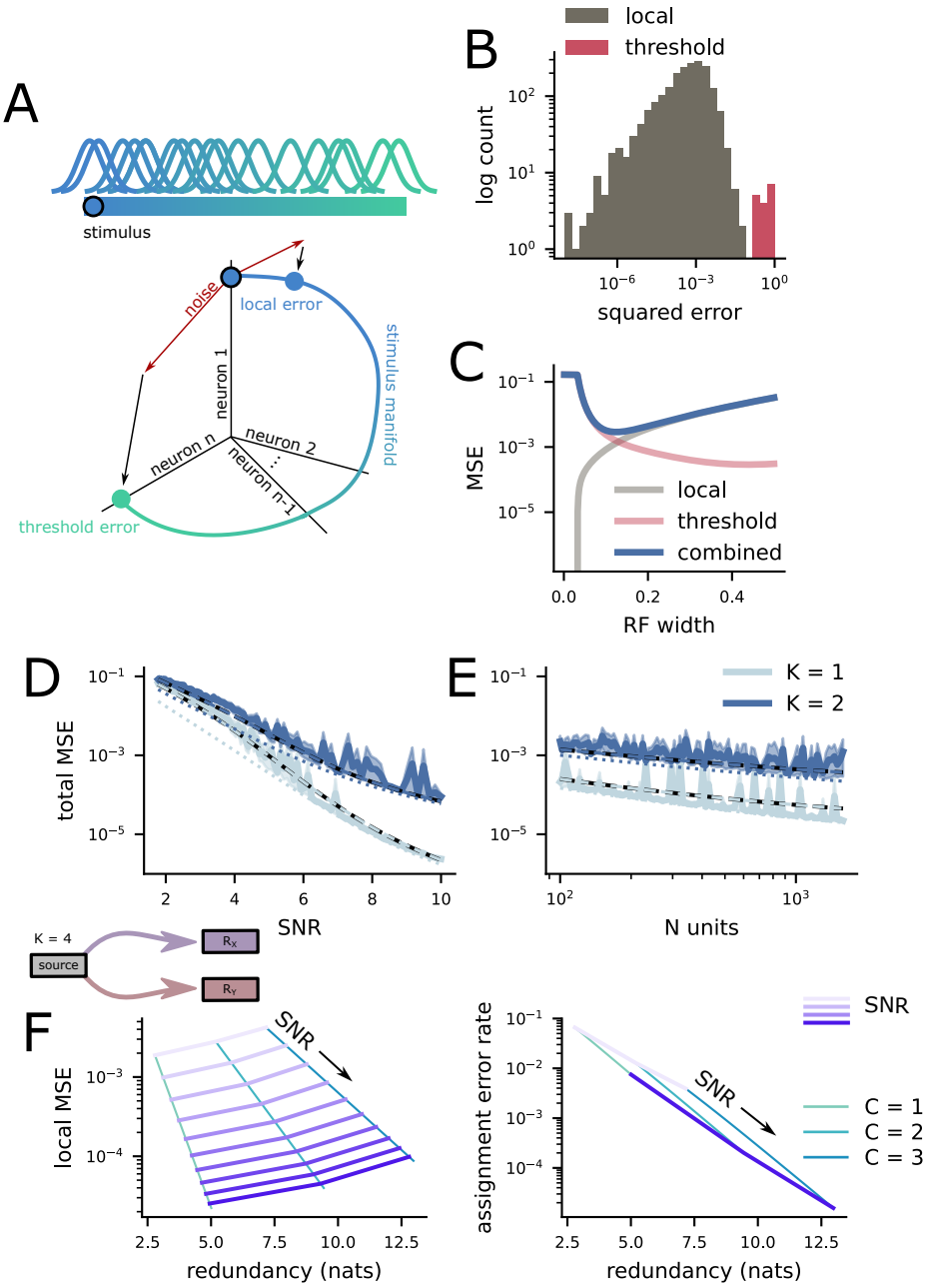

Figure 3: The efficiency of random receptive field codes. **A** A random Gaussian receptive field population. The width of the RFs are chosen to minimize the total MSE of the population. **B** A histogram of errors made by an example RF code. While the majority of errors are local (left, grey), a small proportion of errors are non-local threshold errors (right, red). **C** The two kinds of errors made by RF codes. Threshold errors dominate at small RF sizes, while local errors dominate elsewhere. We choose the RF size for our codes based on the combined error rate. **D** The dependence of total MSE on population SNR for fixed population size. The solid line is from simulations, the dashed line is the prediction of our theory, and the dotted line is local MSE only. **E** The dependence of total MSE on population size for fixed population SNR. The different line types are the same as **D**. **F** (above) Schematic of the situation being investigated. A $K = 4$ dimensional stimulus is represented across two distinct populations of neurons R_X and R_Y. Each population of neurons has $10,000$ units and an SNR varying between 5 and 10. (left) The average local MSE across the two codes and the redundancy across the two regions of the code schematized above, shown for different numbers of overlapping features C. The grey lines connect points with the same SNR across the different overlap levels. (right) The same as on the left, but for the assignment error rate.

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.07.519040; this version posted December 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

2.5 Modularity has benefits for encoding high-dimensional stimuli

So far, we have focused on the case of stimulus features that are encoded by separate brain regions as a consequence of being received by distinct sensory systems – such as the pitch of a sound and the spatial frequency of an image. In this case, modularity is unavoidable and the assignment problem must be solved to integrate the disparate percepts. However, there is extensive evidence for modularity in the brain even within single sensory systems (e.g., [30], though these modules are at least partially redundant[45]). In our framework, this comes with a clear drawback: The assignment problem must be solved when stimulus features are represented in distinct brain regions, and thus the multi-region representations require redundancy between the different regions.

Here, we contrast receptive field (RF) codes for relatively high-dimensional stimuli ($K = 8$) that are distributed across one, two, or three brain regions (Figure 4a). For example, a two-region code for the $K = 8$-dimensional stimulus could encode $K_{X,Y} = 4$ features in each region. However, such a code would make assignment errors at chance levels when presented with multiple stimuli. Thus, we consider codes which represent at least one redundant feature for every two brain regions (e.g., one region could represent four features and the other five, as in Figure 4a, middle). In every case where we compare multi-region codes to each other, the total code (that is, summing across brain regions) is constrained to have the same number of neurons and the same total population power.

Within each region, we use the random RF codes described in the previous section. First, we restrict our analyses to codes with negligible threshold error rates (i.e., $p_{\text{thr}} < .01$). Then, we use the local MSE of each region code (Figure 4b, top) to compute the assignment error rate when integrating across regions (Figure 4b, middle), as developed earlier in the paper. Together, these lead to a total MSE that accounts for local MSE and catastrophic assignment errors (Figure 4b, bottom).

We show that representing the K-dimensional stimuli in multiple brain regions can produce representations with lower total MSE, even when accounting for assignment errors. As a result, this work provides a justification for the modularity observed in the brain from the perspective of reliable and efficient neural codes. For stimuli with $K = 8$ features and smaller neural population sizes or population SNRs, then the lowest total MSE is achieved by a single region code (Figure 4c, bottom and left). In this case, avoiding catastrophic assignment errors is crucial. However, for larger neural population sizes and population SNRs, codes that distribute the representation over two or even three regions can provide lower total MSE than the single region code (Figure 4f, top right).

This results from a combination of two factors: First, the multi-region codes must represent redundant information to reliably solve the assignment problem. Thus, they have a disadvantage compared to the single region code, which need not represent any redundant information. However, second, lower-dimensional RF codes tend to produce smaller errors than higher-dimensional codes, especially very high-dimensional codes. In particular, the local MSE of an RF code depends on RF width, and not directly on dimensionality (Eq. 4). Thus, if low- and high-dimensional codes have the same RF width, then they could, in principle, have the same local MSE. In practice, though, the high-dimensional code will require exponentially more neurons than the low-dimensional code to achieve full coverage of the stimulus space at that width (and to control threshold errors). In our setting, the single region, higher-dimensional codes do have more neurons, but only by a factor of two or three – not enough to achieve equivalent local MSE to the lower-dimensional RF codes. Thus, multi-region codes begin to outperform single-region codes when the gains from their inherent efficiency become larger than the losses from their required redundancy.

We also observe that the difference in total MSE between these different coding strategies is rela-

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.07.519040; this version posted December 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

tively small (Figure 4B, bottom). While our current theory does not provide a guarantee of this in all cases, it points to an interesting potential invariance: The increase in redundancy required to represent a set of stimuli across multiple regions appears to be largely compensated for by the increase in efficiency of the constituent codes. Thus, many different coding strategies appear to provide similar performance.

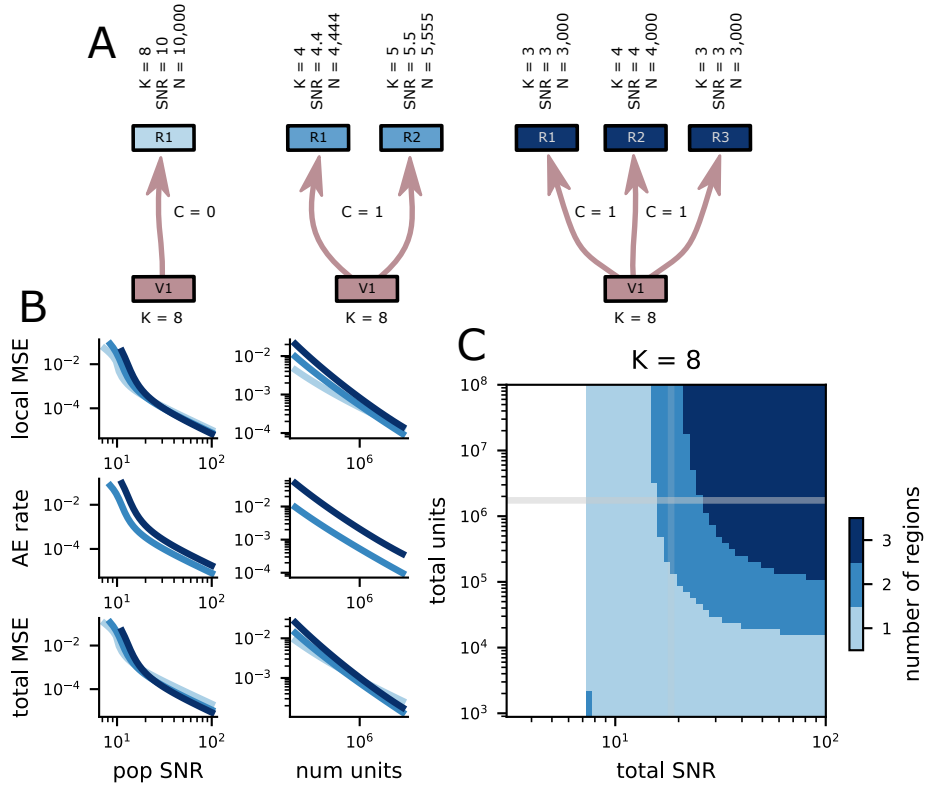

Figure 4: Multiple-region representations minimize error in many conditions. **A** Schematic of three different possibilities for representing a $K = 8$-dimensional stimulus. The features can all be represented in a single population (left); they can be distributed across two populations (middle); they can be distributed across three populations (right). The two and three population codes must represent some features redundantly to avoid assignment errors. **B** (top) The scaling of local MSE for the three codes above with population SNR (left) and the total number of units (right). The colors are the same as in **A** and **C**. (middle) The same as top but showing the scaling of the assignment error rate. (bottom) The same as top but showing the scaling of the total MSE. **C** The number of regions that minimizes the total distortion of the representation of $K = 8$ features for different numbers of neural units and population SNR. The white parts of the plot are when no representational scheme achieved total MSE $< .1$.

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.07.519040; this version posted December 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

2.6 Feedforward neural networks can solve the assignment problem

In the previous section, we analyzed the best-case assignment error rate for populations of neurons with randomly positioned RFs. However, there is no guarantee that a readout from the neural population will be able to achieve this best-case error rate without extensive and potentially metabolically expensive computations. In particular, solutions to the assignment problem used in combinatorial optimization are often framed as linear programming problems and require operations that have no clear neural analog[9].

Here, we focus on the case with two stimuli that are described by three total features ($K = 3$). The stimuli are initially represented by two distinct populations R_X and R_Y – each of which encode two features (Figure 5A, left). For both populations, one feature is unique to that population and the other is common across both populations ($C = 1$). We train a feedforward neural network model (Figure 5B) to take noisy input representations (Figure 5A, right) from R_X and R_Y and produce a representation of the two unique features from R_X and R_Y in an output population, which is also structured as a random RF code (Figure 5A, right). To do this reliably, the network must solve the assignment problem.

First, we train networks to perform this integration (of the two inputs) and marginalization (of the common feature) directly, without a hidden layer (Figure 5B, only R_X, R_Y, and "output"). Then, we quantify whether the activity in the output network is more similar to the correctly or incorrectly assigned representation (Figure 5C). The network without any hidden layers does not learn to assign the stimuli correctly (Figure 5D, end-to-end learning with no integration layer).

This implies that representations of features from multiple sensory modalities (e.g., the expression of a speaker and their words) may need to be derived from higher-dimensional representations that also incorporate the linking, common feature (e.g., estimates of spatial positions in the previous case). However, this does not imply that all of the features of an object must be represented at once: only those that need to combined, along with however many common features are required to achieve reliable assignment. This two-stage computation (a dimensionality expansion followed by a dimensionality reduction) has been observed in many neural structures[] and has been shown to useful for other computations, such as representation frame-of-reference recoding[20].

Second, we show that networks trained to first reconstruct the three dimensional stimulus representations (Figure 5C) and then marginalize out the common feature achieve the minimal assignment error rate, while networks trained directly to construct the marginalized representations achieve low, but not minimal assignment error rates (Figure 5E, end-to-end learning compared to integration learning). While full understanding of this performance difference would require extensive investigation of the representations in the integration layer that are outside of the scope of this paper, this result emphasizes the importance of the integration step. Further, it demonstrates that random RF codes can be used to solve the assignment problem in an optimally reliable way and that they outperform representations learned through backpropagation.

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.07.519040; this version posted December 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

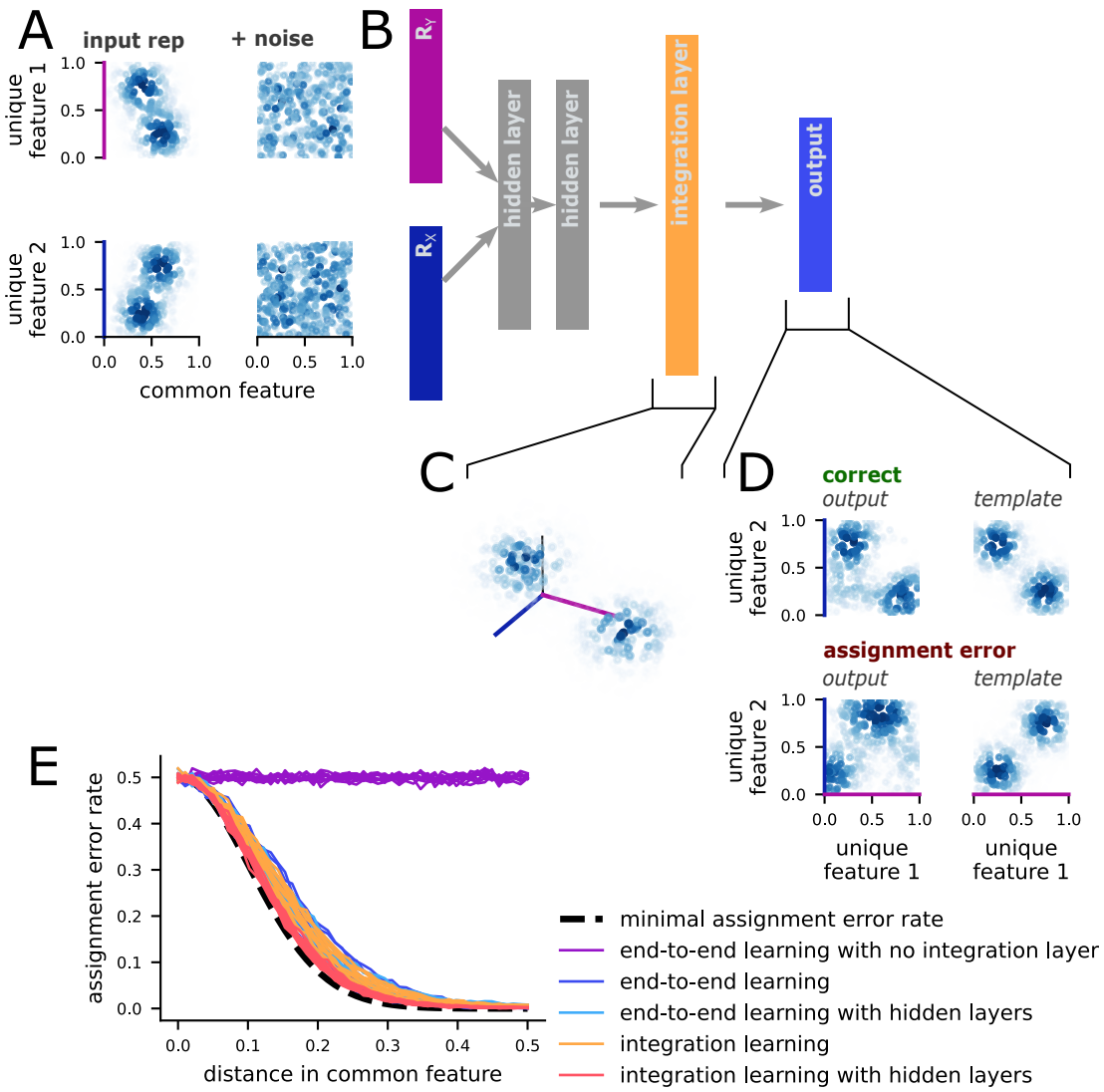

Figure 5: The assignment problem can be solved with one feedforward layer. **A** Example of the input, first without (left) and then with (right) additive noise. The visualization shows the activity level of units in the random RF population. Each point is one unit, and the point is positioned at the unit's RF location. The color shows the level of activity. The two inputs share one common feature (x-axis, $C = 1$) and each have one unique feature (y-axis). **B** Schematic of the feedforward integration, then marginalization network. In the full network, the input layers are followed by two hidden layers, then an integration layer, then the output layer. The output layer constructs a representation of only the two originally unique features of the stimuli, marginalizing out the common feature. **C** Example of target activity in the integration layer. The three features of the stimuli are represented together. The visualization is the same as **A**. **D** Example of a correct output (top) and an output with an assignment error (bottom). The output produced by an example network is shown on the left; the right shows the target output (top) or the prototypical assignment error (bottom). **E** The assignment error rates for stimuli with different distances in the common feature, using several variations on the network described in **B**. The end-to-end learning models are trained to reconstruct the marginalized representation from the inputs, either with or without intervening hidden and integration layers. The integration learning models are trained to construct both the integrated representation and the output target, with or without intervening hidden layers.

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.07.519040; this version posted December 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

3 Discussion

We have described a neural instantiation of the assignment problem, which must be solved when the brain integrates distributed representations of multiple stimuli. We showed that assignment errors depend on the distance between stimuli in the shared representation space, and that increasing the dimensionality of that shared space will – on average – decrease their probability. However, this shared space comes at the cost of redundancy between the distributed representations. To make this cost more concrete, we studied assignment errors in random receptive field codes. In these codes, we showed an advantage for modular representations of high-dimensional stimuli in some conditions, which is consistent with modularity observed in cortex[29–36] (e.g., in the primate ventral and dorsal visual streams). Further, we showed that a simple neural network can reliably (and optimally) solve the assignment problem.

Our work produces several predictions for experimental data. First, efficient coding dictates that the redundancy in representations of the sensory world should be reduced as much as possible[46, 47]. However, our work shows that redundancy between the stimulus representations present in different brain regions is necessary to disambiguating multiple stimuli – other potentialy benefits of redundancy across and within brain regions are robustness to noise[44], damage, and interference[48]. As a result, we predict that brain regions that commonly represent multiple stimuli (e.g., sensory regions) will all have some redundancy with other brain regions. Our framework can be used to predict assignment error rates based on the level of redundancy between two brain regions. Thus, future work could take advantage of large-scale, multi-region recordings to directly test our framework for assignment errors.

Second, our work predicts that explicit, high-dimensional integrated representations will be present when the assignment problem is solved. In particular, if auditory and visual representations are being integrated to solve a multi-sensory classification task, then we predict that there will be an explicit representation of the relevant auditory and visual features (along with the common feature used in assignment, likely spatial position) before a representation of the decision variable can be computed. This integrated representation may be found in multi-sensory associative regions, which have also been identified as potentially important for representation recoding[20]. Future work could identify whether these recoding and assignment operations can be performed simultaneously in the same integrated representation.

Third, our work makes several predictions for what kinds of assignment errors will be more or less likely, many of which already have behavioral support in the time-assignment context from working memory experiments in humans. In particular, our framework predicts that the assignment error rate will increase for stimuli that are nearby in the common feature space, which is validated in [16, 17] (and see the related theoretical work in [16]). It also predicts that assignment error rates will decrease with the inclusion of additional features in the common feature space, which is validated here [18] (but see [19]). Finally, our framework predicts that the assignment error rate will increase with the number of stimuli, which is validated here [15, 16]. While this support is qualitative rather than quantitative, we believe that it lends strength to our predictions for neural data. Further, we believe that a strength of our framework is that it unifies several different types of assignment, including integration across distinct time points – as in the behavioral work here – and integration across distinct brain regions. However, these forms of integration are likely to differ to some degree in practice, even if only due to the kinds of redundant information available. Future work will be necessary to characterize what these potential differences mean for patterns of errors in different

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.07.519040; this version posted December 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

assignment contexts.

The simultaneous representation of multiple stimuli in neural activity is not fully understood. In particular, while we have provided a solution to the assignment problem, which must be solved when integrating across multiple brain regions, the assignment problem is only one of several difficulties that arise in this context. Additional problems include the segmentation, or clustering, of sensory information into a discrete set of causes (i.e., stimuli or objects). While there is extensive work on this process in psychology[1–3] and in machine learning[4, 5], the neural mechanisms are not fully understood (but see [6, 49]). A second additional problem is the binding of stimulus features within a single population of neurons once these features are derived by segmentation. This is essential for producing the unambiguous within-region representations that we have assumed here. The primary method of binding discussed in the literature is through nonlinear mixed selectivity for distinct stimulus features[16, 50, 51] (but see [52, 53]). However, further investigation into each of these constituent problems (such as how multiple representations in a single neural population interact with each other, see [54]) as well as broader efforts to unify all of these systems will be necessary to develop a full understanding of the rich ways that simultaneously representing multiple stimuli constrains the neural code.

In summary, we have developed a general framework for understanding how the brain makes sense of distributed representations of multiple objects – that is, how the brain solves the representation assignment problem. This framework also points to several new directions in the study of neural codes: directions that explicitly contextualize neural representations in a distributed network of brain regions as well as move toward understanding how neural codes can be made suitable for encoding heterogeneous, multi-object environments.

Acknowledgments: We are thankful to Stephanie Palmer, Ji Xia, and Matteo Alleman for their comments on an earlier version of this work. We are also thankful to Allison Ong, Mahham Fayyaz, and Stephanie Hoker for administrative support. This work was supported by NIH F31EY029155 (WJJ), NSF 1707398 (WJJ), Simons Foundation 542983SPI (WJJ), Gatsby Charitable Foundation GAT3708 (WJJ), NIH R01EY019041 (DJF), CRCNS NIH R01MH115555 (DJF), NSF NCS 1631571 (DJF), and a DOD Vannevar Bush Fellowship (DJF).

Author contributions: WJJ conceived of the project. DJF supervised the project development. WJJ created the model and performed the calculations, model simulations, and data analysis. WJJ and DJF wrote the paper.

Competing interests: The authors declare no competing interests.

References

1. Köhler, W. Gestalt psychology. *Psychologische Forschung* **31,** XVIII–XXX (1967).
2. Spelke, E. S. Principles of object perception. *Cognitive science* **14,** 29–56 (1990).
3. Wagemans, J. *et al.* A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure–ground organization. *Psychological bulletin* **138,** 1172 (2012).
4. Cheng, B. *et al. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation* in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020), 12475–12485.
5. Chen, H. *et al.* Unsupervised Segmentation in Real-World Images via Spelke Object Inference. *arXiv preprint arXiv:2205.08515* (2022).

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.07.519040; this version posted December 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

6. Zhou, H., Friedman, H. S. & Von Der Heydt, R. Coding of border ownership in monkey visual cortex. *Journal of Neuroscience* **20,** 6594–6611 (2000).
7. Lamme, V. A., Rodriguez-Rodriguez, V. & Spekreijse, H. Separate processing dynamics for texture elements, boundaries and surfaces in primary visual cortex of the macaque monkey. *Cerebral cortex* **9,** 406–413 (1999).
8. Pouget, A., Deneve, S. & Duhamel, J.-R. A computational perspective on the neural basis of multisensory spatial representations. *Nature Reviews Neuroscience* **3,** 741–747 (2002).
9. Burkard, R., Dell'Amico, M. & Martello, S. Assignment Problems. *SIAM* (2009).
10. Treisman, A. M. & Gelade, G. A feature-integration theory of attention. *Cognitive psychology* **12,** 97–136 (1980).
11. Quinlan, P. T. Visual feature integration theory: Past, present, and future. *Psychological bulletin* **129,** 643 (2003).
12. Treisman, A., Schmidt, H., *et al.* Illusory conjunctions in the perception of objects. *1982* **14,** 107–141 (1982).
13. Ivry, R. B. & Prinzmetal, W. Effect of feature similarity on illusory conjunctions. *Perception & psychophysics* **49,** 105–116 (1991).
14. Zhang, W. & Luck, S. J. Discrete fixed-resolution representations in visual working memory. *Nature* **453,** 233–235 (2008).
15. Bays, P. M., Catalao, R. F. & Husain, M. The precision of visual working memory is set by allocation of a shared resource. *Journal of vision* **9,** 7–7 (2009).
16. Schneegans, S. & Bays, P. M. Neural architecture for feature binding in visual working memory. *Journal of Neuroscience* **37,** 3913–3925 (2017).
17. McMaster, J. M., Tomić, I., Schneegans, S. & Bays, P. M. Swap errors in visual working memory are fully explained by cue-feature variability. *Cognitive psychology* **137,** 101493 (2022).
18. Rajsic, J., Swan, G., Wilson, D. E. & Pratt, J. Accessibility limits recall from visual working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **43,** 1415 (2017).
19. Arnicane, A. & Souza, A. S. Assessing the robustness of feature-based selection in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance* **47,** 731 (2021).
20. Deneve, S., Latham, P. E. & Pouget, A. Efficient computation and cue integration with noisy population codes. *Nature Neuroscience* **4,** 826–831. ISSN: 1097-6256. `http://www.ncbi.nlm.nih.gov/pubmed/11477429` (2001).
21. Körding, K. P. *et al.* Causal inference in multisensory perception. *PLoS one* **2,** e943 (2007).
22. Dokka, K., DeAngelis, G. C. & Angelaki, D. E. Multisensory integration of visual and vestibular signals improves heading discrimination in the presence of a moving object. *Journal of Neuroscience* **35,** 13599–13607 (2015).
23. Dokka, K., Park, H., Jansen, M., DeAngelis, G. C. & Angelaki, D. E. Causal inference accounts for heading perception in the presence of object motion. *Proceedings of the National Academy of Sciences* **116,** 9060–9065 (2019).
24. Olshausen, B. A. & Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381,** 607–609 (1996).
25. Vinje, W. E. & Gallant, J. L. Sparse coding and decorrelation in primary visual cortex during natural vision. *Science* **287,** 1273–1276 (2000).
26. Lewicki, M. S. Efficient coding of natural sounds. *Nature Neuroscience* **5** (2002).
27. Smith, E. C. & Lewicki, M. S. Efficient auditory coding. *Nature* **439,** 978–982 (2006).

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.07.519040; this version posted December 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

28. Perez-Orive, J. *et al.* Oscillations and Sparsening of Odor Representations in the Mushroom Body. *Science* **297,** 359–365 (2002).
29. Ungerleider, L. G. & Mishkin, M. *Two cortical visual systems* 1982.
30. Mishkin, M., Ungerleider, L. G. & Macko, K. A. Object vision and spatial vision: two cortical pathways. *Trends in Neurosciences* **6,** 414–417 (1983).
31. Goodale, M. A. & Westwood, D. A. An evolving view of duplex vision: Separate but interacting cortical pathways for perception and action. *Current Opinion in Neurobiology* **14,** 203–211 (2004).
32. Nassi, J. J. & Callaway, E. M. Parallel processing strategies of the primate visual system. *Nature Reviews Neuroscience* **10,** 360–372. ISSN: 1471-003X (2009).
33. Kaas, J. H. & Hackett, T. A. 'What' and 'where' processing in auditory cortex. *Nature neuroscience* **2,** 1045 (1999).
34. Savic, I., Gulyas, B., Larsson, M. & Roland, P. Olfactory functions are mediated by parallel and hierarchical processing. *Neuron* **26,** 735–745 (2000).
35. Tian, B., Reser, D., Durham, A., Kustov, A. & Rauschecker, J. P. Functional specialization in rhesus monkey auditory cortex. *Science* **292,** 290–293 (2001).
36. Dijkerman, H. C. & De Haan, E. H. Somatosensory processing subserving perception and action: Dissociations, interactions, and integration. *Behavioral and brain sciences* **30,** 224–230 (2007).
37. Kim, J. H., Fiete, I. & Schwab, D. J. Superlinear Precision and Memory in Simple Population Codes. *arXiv preprint arXiv:2008.00629* (2020).
38. Seung, H. S. & Sompolinsky, H. Simple models for reading neuronal population codes. *Proceedings of the national academy of sciences* **90,** 10749–10753 (1993).
39. Zhang, K. & Sejnowski, T. J. Neuronal tuning: To sharpen or broaden? *Neural computation* **11,** 75–84. arXiv: `arXiv:1011.1669v3`. `http://www.ncbi.nlm.nih.gov/pubmed/9950722` (1999).
40. Abbott, L. F. & Dayan, P. The effect of correlated variability on the accuracy of a population code. *Neural computation* **11,** 91–101. ISSN: 0899-7667 (1999).
41. Bethge, M., Rotermund, D. & Pawelzik, K. Optimal short-term population coding: when Fisher information fails. *Neural computation* **14,** 2317–2351 (2002).
42. Berens, P., Ecker, A. S., Gerwinn, S., Tolias, A. S. & Bethge, M. Reassessing optimal neural population codes with neurometric functions. *Proceedings of the National Academy of Sciences* **108,** 4423–4428 (2011).
43. Finkelstein, A., Ulanovsky, N., Tsodyks, M. & Aljadeff, J. Optimal dynamic coding by mixed-dimensionality neurons in the head-direction system of bats. *Nature communications* **9,** 3590 (2018).
44. Johnston, W. J., Palmer, S. E. & Freedman, D. J. Nonlinear mixed selectivity supports reliable neural computation. *PLoS computational biology* **16,** e1007544 (2020).
45. Hong, H., Yamins, D. L. K., Majaj, N. J. & DiCarlo, J. J. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature Neuroscience* **19,** 613–622. ISSN: 1097-6256. `http://www.nature.com/doifinder/10.1038/nn.4247%7B%5C%%7D5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/26900926` (2016).
46. Barlow, H. B. Possible principles underlying the transformations of sensory messages. *Sensory Communication,* 217–234 (1961).
47. Barlow, H. B. Redundancy reduction revisited. *Network* **12,** 241–253 (2001).

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.07.519040; this version posted December 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

48. Li, N., Daie, K., Svoboda, K. & Druckmann, S. Robust neuronal dynamics in premotor cortex during motor planning. *Nature* **532,** 459–64. ISSN: 1476-4687. http://dx.doi.org/10.1038/nature17643%7B%5C%%7D5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/27074502 (2016).
49. Roelfsema, P. R. *et al.* Cortical algorithms for perceptual grouping. *Annual review of neuroscience* **29,** 203–227 (2006).
50. Matthey, L., Bays, P. M. & Dayan, P. A probabilistic palimpsest model of visual short-term memory. *PLoS computational biology* **11,** e1004003 (2015).
51. Fine, J. M., Johnston, W. J., Yoo, S. B. M., Ebitz, R. B. & Hayden, B. Y. Subspace orthogonalization as a mechanism for binding values to space. *arXiv preprint arXiv:2205.06769* (2022).
52. Gray, C. M., König, P., Engel, A. K. & Singer, W. Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature* **338,** 334–337 (1989).
53. Martin, A. B. & von der Heydt, R. Spike Synchrony Reveals Emergence of Proto-Objects in Visual Cortex. *Journal of Neuroscience* **35,** 6860–6870. ISSN: 0270-6474. http://www.jneurosci.org/content/35/17/6860.short (2015).
54. Orhan, A. E. & Ma, W. J. Neural population coding of multiple stimuli. *Journal of Neuroscience* **35,** 3825–3841 (2015).

4 Methods

4.1 Definition of the objects

An object set S of size N is N independent samples from the multidimensional probability distribution $p(s)$ – thus, $p(S) = \prod_i^N p(s_i)$. Each s_i is a vector of length K, so $p(s)$ is a probability distribution over a K-dimensional space. Each of these dimensions represents a feature of the object, such as color, spatial frequency, pitch, or orientation. In the majority of the paper, we assume that the objects are uniformly distributed in the space – that is, each point in the K-dimensional volume is equally likely to occur.

4.2 Definition of the representations

We focus on the representation of our K-dimensional objects in two brain regions, R_X and R_Y. We assume that both regions have some common and some unique information. That is, neither region encodes all K object features. This is guaranteed to be the case when R_X and R_Y represent early sensory areas from different sensory modalities – and there is evidence that a balance of unique and common information is preserved across hierarchies of sensory brain regions even within single modalities, such as in the primate visual system.

Formally, R_X encodes a subset of the K total object features, denoted as F_X – and similarly for R_Y. Each of the features can be identified by their index from 1 to K, and region R_X is said to encode feature $i \in \{1, 2, ..., K\}$ with local distortion D_X^i, which is the variance of an optimal estimator for the value of feature i from the neural activity in region R_X – and, again, similarly for R_Y.

Thus, the subset of features represented in R_X that are not represented in R_Y are the unique information from R_X – that is, $F_X \setminus F_Y$. When R_X is an auditory region and R_Y is a visual region, then these unique features might include representations of pitch and timbre. Further, the subset of features represented in both R_X and R_Y are the common information, which is essential

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.07.519040; this version posted December 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

for solutions to the assignment problem. The size of this intersection $|F_X \cap F_Y| = C_{XY} = C$ has important consequences for the assignment error rate, and the achievable local distortion when the representation capacity is constrained.

Here, we study the reliability of inferences about the original set of N objects S, each of which are described by K feature dimensions, from the neural activity in two distinct regions, R_X and R_Y, which both encode some common and some unique information about the objects. Inferring the original object set from two distinct representations requires the two sets of representations to be combined with each other, which we refer to as assignment.

4.3 Definition of assignment errors

As mentioned above, to make inferences about the whole object set from two distinct sources of information (i.e., R_X and R_Y), the two sets of representations must be integrated with each other. When there is only one object ($N = 1$), then this integration is trivial as there is only one possible one-to-one mapping between the two sets of representations, and this mapping is correct. However, when there is more than one object ($N > 1$), then assignment errors become possible. In particular, for N objects, there are $N!$ possible one-to-one mappings (i.e., assignments) between the two sets of representations, $\hat{X}$ and $\hat{Y}$. Thus, if there is no information about which mapping to select, assignment errors have the probability $1 - 1/N!$, which is near 1 for even relatively small N.

Formally, we can frame the mapping selection problem as an inference about which of the possible maps M is most likely to account for the observed representations $\hat{X}$ and $\hat{Y}$ in R_X and R_Y, respectively. This can be written as,

$$\begin{aligned} p(M|\hat{X},\hat{Y}) &= \frac{p(\hat{X},\hat{Y}|M)p(M)}{p(\hat{X},\hat{Y})} \\ &\propto p(\hat{X}|\hat{Y},M)p(\hat{Y}|M) \\ &\propto p(\hat{X}|\hat{Y},M) \end{aligned} \tag{M1}$$

where we proceed by first assuming that the prior probability of each map is the same (i.e., $p(M)$ is uniform) and that the representations in one region do not depend on the map (i.e., $p(\hat{Y}|M) = p(\hat{Y})$, and similarly for $\hat{X}$). Thus, we are left with a single term, $p(\hat{X}|\hat{Y},M)$, that gives the likelihood of the representations observed in one region (here, $\hat{X}$, but X and Y are interchangeable) conditioned on a particular map M and the representations observed in the other region ($\hat{Y}$). If $\hat{X}$ and $\hat{Y}$ are independent of each other, then $p(\hat{X}|\hat{Y},M) = p(\hat{X})$, as above, and all maps are equally likely. So, it is dependence between the representations in R_X and R_Y that enables the correct assignment to be selected at a rate better than chance. Thus, dependence between those representations is necessary for a reliable solution to the assignment problem. In general, this observation already indicates to us that we should expect pairs of brain regions to encode some common information – so that the assignment problem can be solved – and some unique information – due to both distinct sensory systems, but also due to considerations related to efficient coding, which we make explicit below.

4.3.1 Probability of assignment errors

Given the above inference process, we can now characterize the likelihood that an assignment error occurs given different levels of common information shared between $\hat{X}$ and $\hat{Y}$. From our formalization of the representations above, we know that each object feature i is estimated from the

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.07.519040; this version posted December 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

activity of neurons in region R_X with variance D_X^i, and similarly for R_Y. Here, we assume that these estimates are unbiased and Gaussian distributed, with mean equal to the true value of the object feature and variance as given. The Gaussian distribution is the maximum entropy distribution for fixed mean and variance, which means that these estimates will contain less information about the true value of the object feature than any other distribution with the same mean and variance – thus, this assumption represents an upper bound on the difficulty of the integration task.

Using this formalization, we can write an explicit form for Eq. M1,

$$p(M|\hat{X},\hat{Y}) \propto \prod_i^N \prod_j^C \exp\left(-\frac{(\hat{x}_i^j - M_i\hat{y}^j)^2}{2(D_X+D_Y)}\right)$$
$$\log(p(M|\hat{X},\hat{Y})) \propto -\sum_i^N \sum_j^C (\hat{x}_i^j - M_i\hat{y}^j)^2 \tag{M2}$$

where $\hat{x}_i^j$ and $\hat{y}_i^j$ are the values of feature j for object i from the $N \times C$ estimation matrix of the C commonly represented features for each of the N objects. The common features are ordered consistently across $\hat{x}$ and $\hat{y}$, while the N different objects are not. The map M is an $N \times N$ permutation matrix. Thus, in finding the most likely map, we would search over the $\binom{N}{2}$ possible permutation matrices to find the one that maximizes Eq. M2. This is equivalent to finding the permutation matrix that minimizes the sum squared distance between integrated object representation pairs in the C-dimensional shared feature space.

From the above, it follows that an assignment error occurs precisely when the representation of two objects cross over each other in one of the two brain regions, but not in both – as schematized in Figure 1c, red arrow and assignment lines. For the commonly represented feature values of two objects, x_1^c and x_2^c, the probability that this crossover happens in R_X depends on the distribution of the distances between their estimates,

$$\hat{x}_2^c - \hat{x}_1^c \sim \mathcal{N}(\delta, 2D_X)$$

where δ is the distance between the true values of x_1^c and x_2^c (i.e., $\delta = x_2^c - x_1^c$) and we assume, without loss of generality, that $x_1^c < x_2^c$. Since δ is still Gaussian distributed, we write the probability that the estimate of the first object becomes greater than the estimate of the second object (i.e., that $\hat{x}_1^c > \hat{x}_2^c$) as

$$P(\text{cross in } R_X) = Q\left(\frac{-\delta}{\sqrt{2D_X}}\right)$$

where $Q(.)$ is the cumulative distribution function for the standard Gaussian distribution.

Following this, the full probability of an assignment error incorporates the probability that the cross occurs in R_X or R_Y as well as that it occurs in both (which would not result in an assignment error). We write this probability as,

$$F(\delta) = Q\left(\frac{-\delta}{\sqrt{2D_X}}\right) + Q\left(\frac{-\delta}{\sqrt{2D_Y}}\right) - Q\left(\frac{-\delta}{\sqrt{2D_X}}\right) Q\left(\frac{-\delta}{\sqrt{2D_Y}}\right)$$

where the final term is the probability that a cross occurs in both regions. While we have discussed a single common feature here, this expression is general, and applies for any value of C, so long as

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.07.519040; this version posted December 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

the local distortion D_X and D_Y for all of the common features is the same within each region, which we assume in the majority of the text. This expression already gives us insight into how assignment errors depend on D_X and D_Y for stimuli at some distance δ in a common feature space. However, in general, assignment errors also depend on how likely it is that two stimuli at a particular distance will be observed – that is, on $p_C(\delta)$. In the main text, we develop this dependence, as well as a dependence on the number of objects, which results in Eq. 2.

4.3.2 Assignment error rate approximation for $C = 1$

For one overlapping object feature ($C = 1$), we derive an approximate closed form for Eq. 2. The derivation is as follows,

$$\begin{aligned}
AE_1 &\leq \binom{N}{2} \int_0^s d\delta \, p_C(\delta) \, F(\delta) \\
&= \binom{N}{2} \int_0^s d\delta \, \frac{2(s-\delta)}{s^2} \left(Q\left(\frac{-\delta}{\sqrt{2D_X}}\right) + Q\left(\frac{-\delta}{\sqrt{2D_Y}}\right) - Q\left(\frac{-\delta}{\sqrt{2D_X}}\right) Q\left(\frac{-\delta}{\sqrt{2D_Y}}\right) \right) \\
&\approx \binom{N}{2} \int_0^s d\delta \, \frac{2(s-\delta)}{s^2} \left(Q\left(\frac{-\delta}{\sqrt{2D_X}}\right) + Q\left(\frac{-\delta}{\sqrt{2D_Y}}\right) \right) \\
&= \binom{N}{2} \frac{2}{s^2} \int_0^s d\delta \, (s-\delta) Q\left(\frac{-\delta}{\sqrt{2D_X}}\right) + (s-\delta) Q\left(\frac{-\delta}{\sqrt{2D_Y}}\right)
\end{aligned}$$

since the two terms in this sum are analogous to each other, we deal with them separately before combining.

So, for the first part of each term,

$$\begin{aligned}
\int_0^s d\delta \, s \, Q\left(\frac{-\delta}{\sqrt{2D_i}}\right) &= -s\sqrt{2D_i} \left(-\frac{\delta}{\sqrt{2D_i}} Q\left(-\frac{\delta}{\sqrt{2D_X}}\right) + \phi\left(-\frac{\delta}{\sqrt{2D_X}}\right) \right) \Big|_0^s \\
&= -s\sqrt{2D_i} \left(\phi\left(-\frac{s}{\sqrt{2D_i}}\right) - \frac{s}{\sqrt{2D_i}} Q\left(-\frac{s}{\sqrt{2D_i}}\right) \right) + s\sqrt{\frac{D_i}{\pi}} \\
&\approx s\sqrt{\frac{D_i}{\pi}}
\end{aligned}$$

where ϕ is the standard normal density function and the approximation in the last line holds when $s \gg D_i$, which is the regime we focus on for the main text.

Now, for the second part of each term,

$$\begin{aligned}
-\int_0^s d\delta \, \delta \, Q\left(\frac{-\delta}{\sqrt{2D_X}}\right) &= -4D_i \left(\left(\frac{\delta^2}{2D_i} - 1 \right) Q\left(-\frac{\delta}{\sqrt{2D_i}}\right) - \frac{\delta}{\sqrt{2D_i}} \phi\left(-\frac{\delta}{\sqrt{2D_i}}\right) \right) \Big|_0^s \\
&= -4D_i \left(\left(\frac{s^2}{2D_i} - 1 \right) Q\left(-\frac{s}{\sqrt{2D_i}}\right) - \frac{s}{\sqrt{2D_i}} \phi\left(-\frac{s}{\sqrt{2D_i}}\right) \right) - 2D_i \\
&\approx -2D_i
\end{aligned}$$

where, again, the approximation in the last line holds when $s \gg D_i$.

Then, combining the two expressions above,

$$\frac{2}{s^2} \int_0^s d\delta \, (s-\delta) \, Q\left(-\frac{\delta}{\sqrt{2D_i}}\right) \approx \frac{2}{s}\sqrt{\frac{D_i}{\pi}} - \frac{4}{s^2} D_i$$

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.07.519040; this version posted December 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Before, finally, combining the terms corresponding to D_X and D_Y gives,

$$\begin{aligned} AE_1 &\approx \binom{N}{2}\left[\frac{2\sqrt{D_X}}{s\sqrt{\pi}} - \frac{4D_X}{s^2} + \frac{2\sqrt{D_Y}}{s\sqrt{\pi}} - \frac{4D_Y}{s^2}\right] \\ &= \binom{N}{2}\left[\frac{2\sqrt{D_X + D_Y}}{s\sqrt{\pi}} - \frac{4(D_X + D_Y)}{s^2}\right] \\ &\approx \binom{N}{2}\frac{2\sqrt{D_X + D_Y}}{s\sqrt{\pi}} \end{aligned}$$

This final form is given as Eq. 3 in the main text.

4.4 Random Gaussian receptive field codes

We consider codes for a uniformly distributed K-dimensional stimulus. Without loss of generality, we take the stimulus to be between 0 and 1. The code consists of N units with Gaussian receptive fields. We assume that all units have the same field widths w and height P, but random centers μ^i. In particular, the field of a neuron i is given by

$$r_i(x) = P\exp\left[-\sum_j^K \frac{(x_j - \mu_j^i)^2}{2w_j^2}\right] + \nu$$

where $\nu \sim \mathcal{N}(0, \sigma_N)$ and, here, we assume that $w = w_j \forall j$.

We model the response to multiple stimuli as the sum of responses to each individual stimulus, so that

$$R_i(X) = \sum_{x\in X} r_i(x) + \nu$$

where X is a set of stimuli. Note that the noise ν is only added once.

4.4.1 The spiking energy of the code

First, we compute the average squared L_2-norm of the code response across all stimuli, which we will refer to as V and which we use as a measure of the spiking metabolic energy used by the code,

$$\begin{aligned} V &= \mathbb{E}_x\left[\sum_i^N r_i^2(x)\right] = \sum_i^N \int dx_1 ... \int_0^1 dx_K p(x) r_i(x)^2 \\ &= \sum_i^N \int_0^1 dx_1 ... \int_0^1 dx_K p(x) P\exp\left[-\sum_j^K \frac{(x_j - \mu_j)^2}{w^2}\right] \\ &= P^2 \sum_i^N \prod_j^K \int_0^1 dy_j p(y_j)\exp\left[-\frac{y_j^2}{w}\right] \\ &= P^2 \sum_i^N \prod_j^K \int_0^1 dy_j (2 - 2y_j)\exp\left[-\frac{y_j^2}{w}\right] \end{aligned} \tag{M3}$$

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.07.519040; this version posted December 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

where Eq. M3 changes $y_j^2 = (x_j - \mu_j)^2$. Because both x_j and μ_j are uniformly distributed random variables, we know that $|x_j - \mu_j| = y_j \sim T(0, 1, 0)$ where $T(a, b, c)$ is a triangular distribution.

We then split the above into two terms. For the first,

$$\begin{aligned}\int_0^1 dx_j 2 \exp\left[-\frac{y_i^2}{w^2}\right] &= \sqrt{\pi} w \text{erf}\left(\frac{y}{b}\right)|_0^1 \\ &= \sqrt{\pi} w \text{erf}\left(\frac{1}{w}\right)\end{aligned}$$

and the second,

$$\begin{aligned}\int_0^1 dx_j 2 y_i \exp\left[-\frac{y_i^2}{w^2}\right] &= -w^2 \exp\left[-\frac{x^2}{w^2}\right]|_0^1 \\ &= w^2\left(1 - \exp\left[-\frac{1}{w^2}\right]\right)\end{aligned}$$

So, together,

$$\begin{aligned}V &= P^2 \sum_i^N \prod_j^K \sqrt{\pi} w \text{erf}\left(\frac{1}{w}\right) - w^2\left(1 - \exp\left[-\frac{1}{w^2}\right]\right) \\ &= NP^2 \prod_j^K \sqrt{\pi} w \text{erf}\left(\frac{1}{w}\right) - w^2\left(1 - \exp\left[-\frac{1}{w^2}\right]\right) \\ &= NP^2 \left[\sqrt{\pi} w \text{erf}\left(\frac{1}{w}\right) - w^2\left(1 - \exp\left[-\frac{1}{w^2}\right]\right)\right]^K\end{aligned}$$

and, further, we will tend to deal with $w < .5$, so,

$$V \approx NP^2 \left[\sqrt{\pi} w - w^2\right]^K$$

and

$$P = \sqrt{\frac{V}{N\left[\sqrt{\pi} w - w^2\right]^K}}$$

4.5 The Fisher information of the code

We will use the Fisher information as a measure of the magnitude of local errors via the Cramer-Rao bound. In our framework, the Fisher information along a particular dimension j is given by

$$\begin{aligned}I_j(x) &= \mathbb{E}_r\left[\left(\frac{\partial}{\partial x_j} \log p(r|x)\right)^2\right] \\ &= \mathbb{E}_r\left[\left(-\frac{\partial}{\partial x_j} \frac{(r - r(\bar{x}))^2}{2\sigma_N^2}\right)^2\right]\end{aligned}$$

We want to use the Fisher information to understand the average code performance across the whole stimulus space, rather than only for single points in the stimulus space. There is a complication

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.07.519040; this version posted December 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

here that the Fisher information is related to the MSE of an optimal, unbiased estimator by the Cramer-Rao bound, which states:

$$\text{MSE}(x_j) \geq \frac{1}{I_j(x)}$$

so, we would really like to evaluate

$$\mathbb{E}_x\left[\text{MSE}(x_j)\right] \geq \mathbb{E}_x\left[\frac{1}{I_j(x)}\right]$$

However, this cannot be evaluated analytically. Instead, we evaluate

$$\mathbb{E}_x\left[\text{MSE}(x_j)\right] \geq \frac{1}{\mathbb{E}_x\left[I_j(x)\right]}$$

which is a good approximation of what we want so long as the variance of $I_j(x)$ over x is small relative to the mean.

So,

$$\begin{aligned}
I_j =&\mathbb{E}_x\left[I_j(x)\right] = \frac{P}{\sigma_N^2}\sum_i^N\prod_k^K\int_0^1 dy_k p(y_k)\exp\left[-\frac{y_k^2}{w^2}\right]\int_0^1 dy_j p(y_j)\frac{y_j^2}{w^2}\exp\left[-\frac{y_j^2}{w^2}\right] \\
=&N\frac{P^2}{\sigma_N^2 w^4}\left[\sqrt{\pi}w\text{erf}\left(\frac{1}{w}\right) - w^2\left(1-\exp\left[-\frac{1}{w^2}\right]\right)\right]^{K-1} \\
&\left[\frac{1}{2}w^2\left(\sqrt{\pi}w\text{erf}\left(\frac{1}{w}\right) - 2\exp\left[-\frac{1}{w^2}\right]\right) - w^2\left(w^2-(w^2+1)\exp\left[-\frac{1}{w^2}\right]\right)\right]
\end{aligned}$$

using a similar set of integral tricks as above. For small relatively small s (roughly, $w < .5$), we can approximate this expression as

$$\begin{aligned}
I &\approx \frac{NP^2}{\sigma_N^2 w^4}\left[\sqrt{\pi}w - w^2\right]^{K-1}\left[\frac{1}{2}w^3\sqrt{\pi} - w^4\right] \\
&= \frac{NP^2}{\sigma_N^2}w^{K-2}\left[\sqrt{\pi} - w\right]^{K-1}\left[\frac{1}{2}\sqrt{\pi} - w\right] \\
&= \frac{V}{\sigma_N^2}\frac{w^{K-2}}{\left[\sqrt{\pi}w - w^2\right]^K}\left[\sqrt{\pi} - w\right]^{K-1}\left[\frac{1}{2}\sqrt{\pi} - w\right] \\
&= \frac{V}{\sigma_N^2 w^2}\frac{\left[\sqrt{\pi} - w\right]^{K-1}}{\left[\sqrt{\pi} - w\right]^K}\left[\frac{1}{2}\sqrt{\pi} - w\right] \\
&= \frac{V}{\sigma_N^2 w^2}\frac{\frac{1}{2}\sqrt{\pi} - w}{\sqrt{\pi} - w}
\end{aligned}$$

4.5.1 The threshold error rate of the code

We compute the rate of threshold errors by following the derivations from [1, 2]. In particular, threshold errors occur when the response in the neural population is closer to a non-stimulus representation than the current stimulus representation. To proceed, we discretize the population response, such that there are $1/S(w, K)$ distinct subpopulations, only one of which is active for a

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.07.519040; this version posted December 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

given stimulus, where $S(w, K)$ is the volume of a K-dimensional sphere with radius $2w$. Importantly, this ratio increases both as K increases and w decreases. Since we know the L_2-norm of the code is $\sqrt{V}$, we know that the distance between the stimulus response group and a non-stimulus response group is $\sqrt{2V}$. So, we need to find the probability that the noise crosses half that distance, to become closer to the non-stimulus response group. This is given by

$$\begin{aligned} p_{\text{switch}} &= Q\left(-\frac{\sqrt{2V}}{2\sigma_N}\right) \\ &= \frac{1}{2}\text{erfc}\left(-\frac{\sqrt{V}}{2\sigma_N}\right) \\ &\approx \frac{1}{2}\exp\left[-\frac{V}{4\sigma_N^2}\right]\frac{2\sigma_N}{\sqrt{V\pi}} \\ &= \frac{\sigma_N}{\sqrt{V\pi}}\exp\left[-\frac{V}{4\sigma_N^2}\right] \end{aligned}$$

where the third line follows from an approximation of the error function for large argument.

Then, by a union bound, we find that,

$$p_{\text{thr}} \approx \frac{1}{S(w,K)}\frac{\sigma_N}{\sqrt{V\pi}}\exp\left[-\frac{V}{4\sigma_N^2}\right]$$

However, we find in simulations that this underestimates the probability of threshold errors. The reason is simple: $\sqrt{V}$ represents the average L_2-norm. However, in the random RF code, there is significant variance around this average for different stimuli. As a consequence, stimuli with low L_2-norms will be exponentially more likely to have threshold errors, but this is not accounted for by our equation. To incorporate this, we replace V with V_λ, where,

$$V_\lambda = \max\left(V - \lambda\sigma_V(w), 0\right)$$

where $\sigma_\lambda(w)$ is the standard deviation of V across stimuli for a random RF code, and it depends on w. This quantity is approximated analytically in the supplement (see *Variance of the spiking energy V of the random RF codes* in *Supplement*).

Given this correction and setting $\lambda = 2$, we find good agreement between the resulting total predicted MSE and the empircal MSE (Figure 3d, e).

4.5.2 Total MSE of the code

Finally, we can write the total approximate MSE for small w as follows,

$$\text{MSE} \approx \frac{2w^2\sigma_N^2}{V} + p_{\text{thr}}(w, V, \sigma_V)\frac{1}{6}$$

where the $\frac{1}{6}$ is the average squared size of a threshold error, calculated from the triangle distribution. We find that this expression has good agreement with the MSE estimated from simulations (Figure 3d, e).

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.07.519040; this version posted December 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

4.6 Calculating the redundancy between representations

We quantify the redundancy between R_X and R_Y as the mutual information between $\hat{X}$ and $\hat{Y}$,

$$\begin{aligned} R &= I(X;Y) = H(X) - H(X|Y) \\ &= K_X \frac{1}{2}\log 2\pi s^2 - (K_X - C)\frac{1}{2}\log 2\pi s^2 - C\frac{1}{2}\log 2\pi(D_X + D_Y) \\ &= C\frac{1}{2}\log 2\pi s^2 - C\frac{1}{2}\log 2\pi(D_X + D_Y) \\ &= \frac{C}{2}\log\frac{s^2}{D_X + D_Y} \\ &= \frac{C}{2}\log\frac{s^2(1-\Delta D^2)}{4D_S^2} \end{aligned}$$

where R is the redundancy between the representations in region X and Y (in nats). As discussed throughout this work, this redundancy is crucial for solving the assignment problem. The redundancy is proportional to the number of commonly represented features C – so, increasing C will produce relatively large increases in redundancy, and, as we have seen, can be expected to effectively reduce assignment errors. Further, increasing the asymmetry of feature representations ΔD reduces the level of redundancy between the R_X and R_Y – so, as anticipated, increasing this asymmetry will increase the assignment error rate.

This redundancy represents the cost of our solution to the assignment problem. Thus, we desire a solution to the assignment problem that achieves a particular assignment error rate and local distortion magnitude while using as few bits – and, in particular, as few redundant bits – as possible.

4.7 The random RF integration model

The stimuli for the integration model are described by $K = 3$ dimensions. On each "trial," we sample two uniformly distributed, random stimuli. Two features are represented in each of the two input populations. The two populations each represent $C = 1$ overlapping and one unique feature, with $N = 400$ units. Each population is noisy and has SNR $= \sqrt{20}$.

When present, the integration layer has $N = 2000$ units. The training targets are generated by obtaining representations from a $K = 3$-dimensional random RF code. When present, the hidden layers are two layers of 500 units, with no specific training target. The output population is $N = 400$ units, and the training target is given by samples from a random RF code for the two unique features in the input representation.

4.7.1 End-to-end learning

The network is trained to produce the output population directly. The different network architectures (integration or no integration, hidden layers or no hidden layers) indicate how many and what kind of intervening layers are present; the training objective does not change.

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.07.519040; this version posted December 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

4.7.2 Integration learning

The network is trained to produce the activity from a random RF code in the integration layer (i.e., the full $K = 3$-dimensional representation). Then, the output is learned by Ridge regression and thresholding to the output layer.

4.7.3 Training procedure

The nonlinear layers are learned by backpropagation with stochastic gradient descent, using the Keras Tensorflow[3] environment and the Adam optimizer. The batch size is set to 200 and the network is trained for a maximum of 200 epochs on 50000 input samples with early stopping (if performance on a validation set stops improving for 2 epochs). In practice, the early stopping is often triggered after around 20 epochs.

References

1. Johnston, W. J., Palmer, S. E. & Freedman, D. J. Nonlinear mixed selectivity supports reliable neural computation. *PLoS computational biology* **16,** e1007544 (2020).
2. Kim, J. H., Fiete, I. & Schwab, D. J. Superlinear Precision and Memory in Simple Population Codes. *arXiv preprint arXiv:2008.00629* (2020).
3. Abadi, M. *et al. Tensorflow: A system for large-scale machine learning* in *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)* (2016), 265–283.

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.07.519040; this version posted December 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

S1 Supplement

S1 Asymmetric feature encoding

In general, the representations of the same object feature may not be encoded with the same fidelity in each region. For instance, estimates of azimuthal position from the visual system are thought to be much more accurate than those from the auditory system[1]. Further, the relative accuracy of estimates of the azimuthal position and elevation of a particular sound shifts across the sensory field[2]. Here, we show how these asymmetric feature representations affect the assignment error rate. In all cases, we assume that the two sources of stimulus information are optimally combined by the downstream brain region, such that the estimator variance of a commonly represented feature D_S is given by the following relationship,

$$D_S = \frac{D_X D_Y}{D_X + D_Y}$$

where D_X and D_Y are, as before, the estimator variance of the representation of that feature in regions R_X and R_Y, respectively (Figure S1a). Using this equation for the integrated estimator variance, we can rewrite D_X and D_Y in terms of the integrated estimator variance D_S and the representation asymmetry ΔD,

$$D_X = \frac{2D_S}{1 - \Delta D}$$
$$D_Y = \frac{2D_S}{1 + \Delta D}$$

where, without loss of generality, we have assumed $D_Y \geq D_X$ (Figure S1b) and the representation asymmetry ΔD is constrained to be on the interval $[0, 1)$. When $\Delta D = 0$, the feature representations are perfectly symmetric, as before, and $D_X = D_Y = 2D_S$; when $\Delta D \to 1$, then $D_X \to D_S$ and $D_Y \to \infty$. A fully asymmetric representation provides no redundancy and thus cannot be used to solve the assignment problem, effectively reducing the number of commonly represented features C.

Further, we show that increasing representation asymmetry increases the assignment error rate, both for all individual distances between stimuli δ (Figure S1c) and overall (Figure S1d, and see *Representation asymmetry increases the assignment error rate* in *Methods*). That is, increasing the level of asymmetry while keeping other features of the representations constant increases the assignment error rate. Next, we evaluate the redundancy between $\hat{X}$ and $\hat{Y}$ as a function of precision ratio and representation asymmetry. As expected, increasing representation asymmetry also decreases the level of redundancy between the two sets of representations (Figure S1e), as it makes the set of representations observed in one region less informative about the representations observed in the other.

In some cases, asymmetric feature representations are imposed by the physical limits of different sensory systems, such as with audition and vision; in others, however, asymmetric feature representations can be designed, such as is hypothesized to be the case across the canonical dorsal and ventral visual processing streams in the primate. In that case, representations of position are thought to be encoded with high fidelity in the dorsal visual stream, but with relatively low fidelity in the ventral visual stream. Yet, the assignment problem must also be solved when integrating across those two sets of regions, so an asymmetric representation of spatial position across them may seem non-optimal. Next, we resolve this apparent contradiction by showing that, in some cases,

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.07.519040; this version posted December 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

asymmetric feature representations increase the efficiency of solutions to the assignment problem by reducing the redundancy between the two regions.

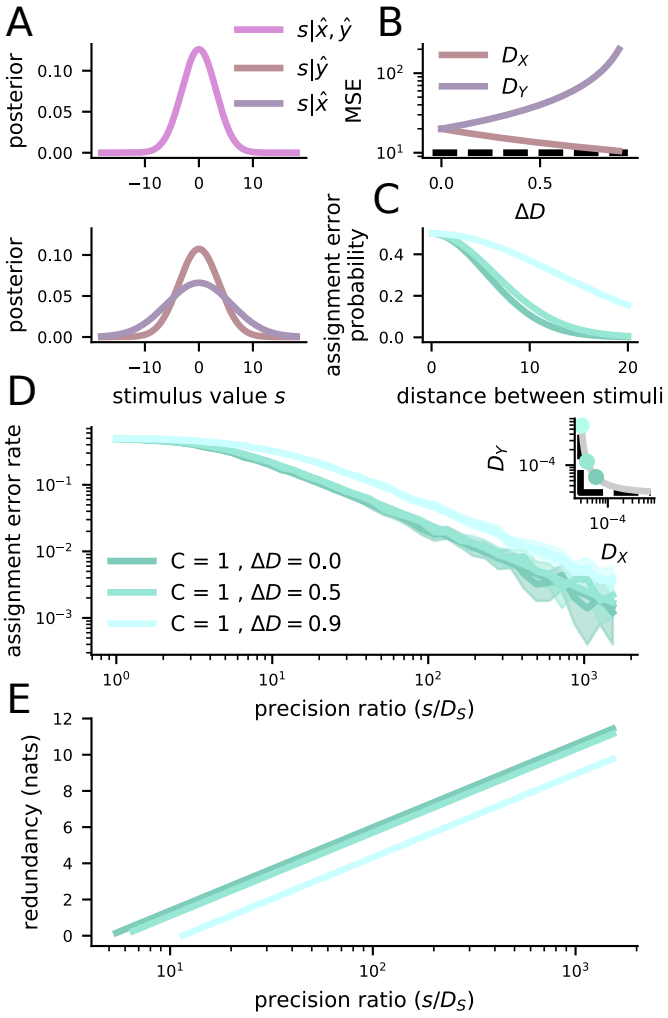

Figure S1: Asymmetric feature representations increase the assignment error rate, but decrease redundancy. **A** Schematic of representation integration. The stimulus estimates from R_X and R_Y are optimally combined to give rise to a stimulus estimate with variance $D_S = D_X D_Y/(D_X + D_Y)$. **B** We can keep D_S constant (dashed line) and define D_X and D_Y in terms of D_S and the representation asymmetry ΔD, which is 0 when $D_X = D_Y$ and goes to 1 when $D_X \to D_S$ and $D_Y \to \infty$, or vice versa. **C** Increasing representation asymmetry ΔD increases the assignment error rate for pairs of stimuli at all distances. **D** The overall assignment error rate increases with increases in representation asymmetry across all precision ratios. **E** The redundancy between $\hat{X}$ and $\hat{Y}$ decreases with increases in representation asymmetry across all precision ratios.

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.07.519040; this version posted December 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

S1.1 Representation asymmetry increases the assignment error rate

For asymmetric local distortion, Eq. 1 can be rewritten as

$$F(\delta, \Delta D) \approx Q\left(\frac{-\delta\sqrt{1-\Delta D}}{\sqrt{4D_S}}\right) + Q\left(\frac{-\delta\sqrt{1+\Delta D}}{\sqrt{4D_S}}\right)$$

which reveals that one term of the sum will decrease with increased asymmetry and the other term will increase. Yet, it is not clear how this should affect the assignment error rate at a particular δ. To resolve this, we study the derivative of $F(\delta, \Delta D)$ with respect to ΔD,

$$\frac{\partial}{\partial \Delta D} F(\delta, \Delta D) \approx \frac{\delta}{4\sqrt{D_S}}\left(f\left(\frac{-\delta\sqrt{1-\Delta D}}{\sqrt{4D_S}}\right) - f\left(\frac{-\delta\sqrt{1+\Delta D}}{\sqrt{4D_S}}\right)\right)$$

where f is the probability density function of the standard Gaussian. From this, we can see that the first term will always be larger than or the same size (when $\Delta D = 0$) as the second term. Thus, the assignment error rate will increase as ΔD increases.

S2 Nonlinear mappings between $\hat{X}$ and $\hat{Y}$

In the main text, we consider representations in R_X and R_Y that directly share one or more features (i.e., the features are identical, Figure S2, left). Alternatively, some of the features encoded by R_Y could have a linear relationship with some of the features encoded by R_X, such that:

$$\hat{Z}_C = MY + b + \epsilon$$

where Y is the true values of all the features encoded in R_Y, $\hat{Z}_C$ is the estimate of the common features encoded by R_X, M is a linear transform, b is an intercept, and $\epsilon \sim \mathcal{N}(0, D_Y)$. So, we can then perform assignment by comparing the common features decoded from R_X, $\hat{X}_C$ and the common features decoded from Y into R_X's frame of reference, $\hat{Z}_C$. In this case, we are comparing quantities with D_X and D_Y as in the main text, though these two quantities need not be the same.

However, in principle, the dependence between $\hat{X}$ and $\hat{Y}$ could be nonlinear (Figure S2, right). There are two main cases here. First,

$$\hat{Z}_C = f\left(Y\right) + \epsilon$$

where the function $f(.)$ is a nonlinear decoder for the common features from R_Y. In this case, the noise is assumed to be uniform across different values of Y. Here, the equations that we develop apply directly, since we can account for different noise levels between $\hat{X}_C$ and $\hat{Z}_C$ by differing D_X and D_Y.

In contrast, the second case would be that in which,

$$\hat{Z}_C = f\left(Y\right) + \epsilon(Y)$$

where the noise also depends on the value of the common feature. We can still approach this case using our framework, but the integrals that we are required to evaluate are unlikely to be tractable or easily approximable except for very simple noise dependencies.

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.07.519040; this version posted December 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

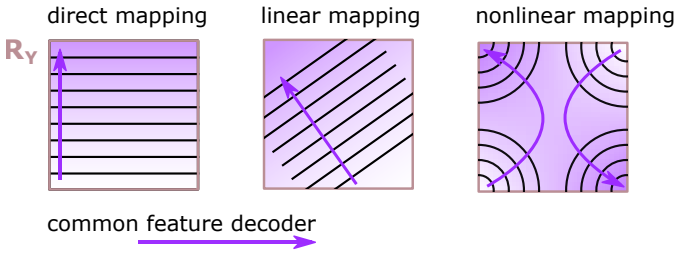

Figure S2: Different mappings between the features in R_X and R_Y. (left) A direct mapping where the two regions encode the same features. (center) A linear mapping, where the two regions encode slightly different features, but where the features can be related by a linear map. (right) A nonlinear mapping, where the two regions encode different features, but that have a nonlinear dependence with each other.

S3 Variance of the spiking energy V of the random RF codes

Due to the random placement of the receptive fields, there will tend to be significant variance in V across different stimulus values. This variance is important for quantifying threshold errors. So,

$$\mathrm{Var}_x(V) = \mathbb{E}_x\left[V^2\right] - \mathbb{E}_x[V]^2$$

where the second term is already given above. For the first term,

$$\begin{aligned}\mathbb{E}_x\left[V^2\right] &= \mathbb{E}_x\left[\left(\sum_i^N r_i^2(x)\right)^2\right] \\ &= \sum_i^N \int_0^1 dx_1 ... \int_0^1 dx_D p(x) r_i^4(x) + \sum_{i\neq j}^{N(N-1)} \int_0^1 dx_1 ... \int_0^1 dx_D p(x) r_i^2(x) r_j^2(x)\end{aligned}$$

Here, the first sum follows along the same lines as above, where for a particular i,

$$\begin{aligned}\int_0^1 dx_1 ... \int_0^1 dx_D p(x) r_i^4(x) &= P^4 \prod_j^D \int_0^1 dx_j p(y_j) \exp\left[-\frac{2y_j^2}{w^2}\right] \\ &= 2P^4 \prod_j^D \int_0^1 dx_j (1-y_j) \exp\left[-\frac{2y_j^2}{w^2}\right] \\ &= P^4 \prod_j^D \sqrt{\frac{\pi}{2}} w \mathrm{erf}\left(\frac{\sqrt{2}}{w}\right) - \frac{w^2}{2}\left(1 - \exp\left[-\frac{2}{w^2}\right]\right)\end{aligned}$$

where the integral is given as above when changing w for $w/\sqrt{2}$.

The second term requires some additional trickery. We proceed for particular units i and h along dimension j, where we notice that $p(y_{ij})$ and $p(y_{hj})$ are not independent. Intuitively, if y_{ij} is large, then this must mean that x_j is near the edge of the stimulus space (i.e., near 0 or 1) and that implies that y_{hj} can take on large values too. To evaluate the integral, we need to decouple these two probabilities. So, we (re-)introduce x_j such that $p(y_{hj}|x_j)p(x_j) = p(y_{hj})$. So, we need to evaluate the triple integral (for particular units h and i, along dimension j),

$$P^4 \int_0^1 dx_j \int_{-x_j}^{1-x_j} dy_i \exp\left[-\frac{2y_i^2}{w^2}\right] \int_{-x_j}^{1-x_j} dy_h \exp\left[-\frac{2y_h^2}{w^2}\right]$$

bioRxiv preprint doi: https://doi.org/10.1101/2022.12.07.519040; this version posted December 7, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

where now the integrals are over a uniform distribution. So, we obtain,

$$P^4 \int_0^1 dx_j p(x_j) \left[\frac{1}{2}\sqrt{\pi} w \left(\text{erf}\left(\frac{1-x_j}{w}\right) + \text{erf}\left(\frac{x_j}{w}\right)\right)\right]^2$$

We can expand this into three separate terms. All of the terms can be integrated except one, and that one can be approximated for moderately small w.

First,

$$\begin{aligned}\int_0^1 dx_j \text{erf}^2\left(\frac{1-x_j}{w}\right) &= \int_0^1 dx_j \text{erf}^2\left(\frac{x_j}{w}\right) \\ &= \frac{2}{\sqrt{\pi}} w \exp\left[-1/w^2\right] \text{erf}\left(1/w\right) + \text{erf}^2\left(1/w\right) - \sqrt{\frac{2}{\pi}} w \text{erf}\left(\sqrt{2}/w\right)\end{aligned}$$

and, second,

$$\begin{aligned}\int_0^1 dx_j \text{erf}\left(\frac{1-x_j}{w}\right) \text{erf}\left(\frac{x_j}{w}\right) &\approx \int_0^{.5} dx_j \text{erf}\left(\frac{x_j}{w}\right) + \int_{.5}^1 dx_j \text{erf}\left(\frac{1-x_j}{w}\right) \\ &= 2\int_0^{.5} dx_j \text{erf}\left(\frac{x_j}{w}\right) \\ &= \text{erf}\left(\frac{1}{2w}\right) - \frac{2w}{\sqrt{\pi}}\left[1 - \exp\left[-\frac{1}{4w^2}\right]\right]\end{aligned}$$

So,

$$\begin{aligned}\mathbb{E}_x\left[V^2\right] =& N P^4 \prod_j^D \left[\sqrt{\frac{\pi}{2}} w \text{erf}\left(\frac{\sqrt{2}}{w}\right) - \frac{w^2}{2}\left(1 - \exp\left[-\frac{2}{w^2}\right]\right)\right] \\ &+ N(N-1) P^4 \prod_j^D \frac{1}{2}\pi w^2 Z\end{aligned}$$

where

$$\begin{aligned}Z =& \frac{2}{\sqrt{\pi}} w \exp\left[-1/w^2\right] \text{erf}\left(1/w\right) + \text{erf}^2\left(1/w\right) - \sqrt{\frac{2}{\pi}} w \text{erf}\left(\sqrt{2}/w\right) \\ &+ \text{erf}\left(\frac{1}{2w}\right) - \frac{2w}{\sqrt{\pi}}\left[1 - \exp\left[-\frac{1}{4w^2}\right]\right]\end{aligned}$$

which can be further simplified in the small w case to find

$$\sigma_V^2 \approx \frac{V^2}{N w^D \left[2\pi\right]^{D/2}} + \frac{V^2 \frac{\pi}{2}^D}{\left[\sqrt{\pi} - w\right]^{2D}} \left[2 - \frac{\sqrt{2}+2}{\sqrt{\pi}} w\right]^D - V^2$$

This expression has two parts: a decay with population size (with N, first term) and a stable value for arbitrarily large population size (the second two terms).

References

1. Körding, K. P. *et al.* Causal inference in multisensory perception. *PLoS one* **2,** e943 (2007).
2. Makous, J. C. & Middlebrooks, J. C. Two-dimensional sound localization by human listeners. *The journal of the Acoustical Society of America* **87,** 2188–2200 (1990).