What Representational Similarity Measures Imply about Decodable Information

Anonymous Author(s)
Affiliation
Address
email

Abstract

Neural responses encode information that is useful for a variety of downstream tasks. A common approach to understand these systems is to build regression models or "decoders" that reconstruct features of the stimulus from neural responses. Here, we investigate how to leverage this perspective to quantify the similarity of different neural systems. This is distinct from typical motivations behind neural network similarity measures like centered kernel alignment (CKA), canonical correlation analysis (CCA), and Procrustes shape distance, which highlight geometric intuition and invariances to orthogonal or affine transformations. We show that CKA, CCA, and other measures can be equivalently motivated from similarity in decoding patterns. Specifically, these measures quantify the average alignment between optimal linear readouts across a distribution of decoding tasks. We also show that the Procrustes shape distance upper bounds the distance between optimal linear readouts and that the converse holds for representations with low participation ratio. Overall, our work demonstrates a tight link between the geometry of neural representations and the ability to linearly decode information. This perspective suggests new ways of measuring similarity between neural systems and also provides novel, unifying interpretations of existing measures.

1 Introduction

The computational neuroscience and machine learning communities have developed a multitude of methods to quantify similarity in population-level activity patterns across neural systems. Indeed, a recent review by Klabunde et al. [16] catalogues over thirty approaches to quantifying representational similarity by various publications. Many of these measures quantify similarity in the *shape* or *representational geometry* of point clouds. For example, recent papers (e.g. [30, 8]) leverage the Procrustes distance and other concepts from *shape theory*—an established body of work that formalizes the notion of shape and ways to measure distance between shapes [14, 9]. Other measures of representational similarity are not quite this explicit but still emphasize desired *geometric invariance* properties. For example, work by Kornblith et al. [18] popularized centered kernel alignment (CKA) by emphasizing its invariance to isotropic scaling, translation, and orthogonal transformations. These are precisely the same invariances specified by classical shape theory [14, 9]. Earlier work by Raghu et al. [24] argued for a more flexible invariance to affine transformations, and they used canonical correlations analysis (CCA) for this purpose. Contemporary work in neuroscience is replete with similar geometric reasoning and quantitative frameworks [20].

Here we ask: what does the similarity between neural representations, potentially in terms of shape or geometry, imply about the similarity between *functions* performed by those neural systems? Potentially, not very much. Maheswaranathan et al. [22] showed that recurrent neural networks with different architectures performed the same task with similar dynamical algorithms, but with different

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

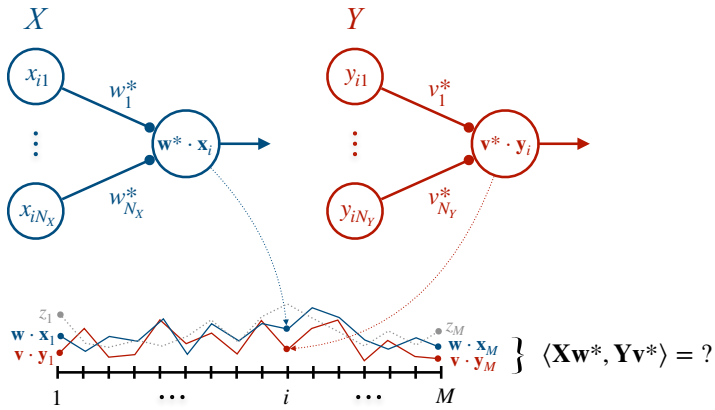

Figure 1: Neural circuit interpretation of decoding from neural representations. First, optimal linear decoding weights $\boldsymbol{w}^*$ and $\boldsymbol{v}^*$ are found by training over stimuli $i = 1 \dots M$ with targets $\boldsymbol{z}$. Then the similarity between the two systems could be measured by comparing the predictions of the optimal decoders. The unit activations x_{ij} and y_{ij}, with $i = 1 \dots M$ and $j = 1 \dots N_X$ or N_Y, can represent the elements of representation matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$, so this procedure can be succinctly summarized as computing the inner product $\langle \boldsymbol{X}\boldsymbol{w}^*, \boldsymbol{Y}\boldsymbol{v}^* \rangle$.

representational geometry. More recently, Lampinen et al. [21] showed that representational geometry can be biased by other nuisance variables, such as the order in which multiple tasks are learned.

On the other hand, several basic observations suggest that geometry and function ought to be related. Consider, for example, the common practice of using linear models to decode task-relevant variables from neural population activity [1, 19]. The premise behind these analyses is that anything linearly decodable from system A is readily accessible to layers or brain regions immediately downstream of A. Therefore, information that is linearly decodable from A may relate to functional role played by A in the complete system [4]. Notably, the invariances of typical representational similarity measures—translations, isotropic scalings, and orthogonal transformations—closely coincide with the set of transformations that leave decoding accuracy unchanged. For example, translations and rotations of neural population activity will not impact the accuracy of a linear decoder with an intercept term and an L_2 penalty on the weights. Thus, if we accept the premise that decoding accuracy is a proxy—even, perhaps, an imperfect proxy—for neural system function, then representational geometry may be a reasonable framework to capture something about this function.

We provide a unifying theoretical framework that connects several existing methods of measuring representational similarity to notions of similarity in terms of behavior on decoding tasks. We show that popular methods such as centered kernel alignment (CKA) [18] and similarity based on canonical correlations analysis (CCA) [24] can be interpreted as alignment scores between optimal linear decoders with particular weight regularizations. Additionally, we study how the *shape* of neural representations is related to decoding by deriving bounds that relate the expected squared difference between optimal decoder readouts and the Procrustes distance. We find that the Procrustes distance is a more strict notion of geometric dissimilarity than the expected distance between optimal decoder readouts, in that a small value of the Procrustes distance implies a small expected distance between optimal decoder readouts but the converse is not necessarily true. However, we also show that one can never mislead oneself by simultaneously measuring a small Procrustes distance between normalized representations and a large expected difference in decoded signals. This formalizes intuition (for example see [16]) regarding the relationship between functional and representational similarity measures for these specific choices of metrics.

2 Theoretical Framework for Linear Decoding

We use $\boldsymbol{X} \in \mathbb{R}^{M \times N_X}$ and $\boldsymbol{Y} \in \mathbb{R}^{M \times N_Y}$ to denote matrices holding responses to M stimulus conditions measured across neural populations consisting of N_X and N_Y neurons, respectively. Throughout this paper we will assume that the neural responses have been preprocessed so that the columns of $\boldsymbol{X}$ and $\boldsymbol{Y}$ have zero mean.

We first consider the problem of predicting (or "decoding") a target vector $\boldsymbol{z} \in \mathbb{R}^M$ from neural population responses by a linear function $\boldsymbol{X} \mapsto \boldsymbol{X}\boldsymbol{w}$ where $\boldsymbol{w} \in \mathbb{R}^{N_X}$. One can view this as a simplified neural circuit model (fig. 1), where each element of $\hat{\boldsymbol{z}} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon}$ is the firing rate readout unit over each of the M conditions. Here, we interpret $\boldsymbol{w}$ as a vector of synaptic weights and the random vector $\boldsymbol{\epsilon}$ represents potential noise corruptions; for example, $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \sigma^2)$ where $\sigma^2 > 0$ specifies the scale of noise. This neural circuit interpretation of decoding is fairly standard within the literature [31].

Within this setting, we formalize the problem of decoding $\boldsymbol{z}$ from the population activity $\boldsymbol{X}$ through the following class of optimization problems:

$$\underset{\boldsymbol{w}}{\text{maximize}} \quad \tfrac{1}{M}\boldsymbol{z}^\top \boldsymbol{X}\boldsymbol{w} - \tfrac{1}{2}\boldsymbol{w}^\top \boldsymbol{G}(\boldsymbol{X})\boldsymbol{w} \tag{1}$$

where $\boldsymbol{G}(\cdot)$ is a function mapping $\mathbb{R}^{M\times N}$ onto the set of symmetric positive definite $N \times N$ matrices. The objective is concave, and equating the gradient to zero yields a closed form solution:

$$\boldsymbol{w}^* = \tfrac{1}{M}\boldsymbol{G}(\boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{z} \tag{2}$$

To appreciate the relevance of eq. (1), note that if $\boldsymbol{G}(\boldsymbol{X}) = \boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I}$, then the solution $\boldsymbol{w}^*$ coincides with the optimum of the familiar ridge regression problem:

$$\underset{\boldsymbol{w}}{\text{minimize}} \quad \tfrac{1}{M}\|\boldsymbol{z} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{w}\|_2^2 \tag{3}$$

where $\lambda > 0$ specifies the strength of regularization on the coefficients. More generally, we can interpret eq. (1) as maximizing the inner product between the target, $\boldsymbol{z}$, and the linear readout, $\boldsymbol{X}\boldsymbol{w}$, plus a penalty term on the scale of $\boldsymbol{w}$. Indeed, $\boldsymbol{w} \mapsto \sqrt{\boldsymbol{w}^\top\boldsymbol{G}(\boldsymbol{X})\boldsymbol{w}}$ defines a norm on $\boldsymbol{w}$.

Equation (1) is a useful generalization of eq. (3) because there are situations where the scale of the decoded signal matters. For instance, when the readout unit is corrupted with noise, $\hat{\boldsymbol{z}} = \boldsymbol{X}\boldsymbol{w} + \boldsymbol{\epsilon}$ where $\boldsymbol{\epsilon}_i \sim \mathcal{N}(0, \sigma^2)$, we can interpret eq. (1) as maximizing the signal-to-noise ratio subject to a cost on the magnitude of the weights $\boldsymbol{w}$.

In most cases, we will consider choices of $\boldsymbol{G}(\cdot)$ that can be written as:

$$\boldsymbol{G}(\boldsymbol{X}) = a\boldsymbol{X}^\top\boldsymbol{X} + b\boldsymbol{I} \tag{4}$$

for some $a \geq 0$ and $b > 0$. Under this choice, the penalty on $\boldsymbol{w}$ equals the sum of two terms:

$$\boldsymbol{w}^\top\boldsymbol{G}(\boldsymbol{X})\boldsymbol{w} = a\|\boldsymbol{X}\boldsymbol{w}\|_2^2 + b\|\boldsymbol{w}\|_2^2 \tag{5}$$

which can each be interpreted as constraints on the neural circuit model diagrammed in fig. 1. The first term, scaled by a, penalizes the activation level of the readout unit, which can be viewed as an energetic constraint. The second term, scaled by b, penalizes the strength of the synaptic weights, which can be viewed as a resource constraint for building and maintaining strong synapses. We note again that ridge regression is achieved as a special case when $a = 1$ and $b = \lambda$.

2.1 Quantifying differences between networks through the lens of decoding

We now turn our attention to the the problem of quantifying representational similarity between $\boldsymbol{X} \in \mathbb{R}^{M\times N_X}$ and $\boldsymbol{Y} \in \mathbb{R}^{M\times N_Y}$. As before, we specify a target vector $\boldsymbol{z} \in \mathbb{R}^M$ and optimize $\boldsymbol{w}$ as specified in eq. (1). We optimize $\boldsymbol{v}$ accordingly:

$$\underset{\boldsymbol{v}}{\text{maximize}} \quad \tfrac{1}{M}\boldsymbol{z}^\top \boldsymbol{Y}\boldsymbol{v} - \tfrac{1}{2}\boldsymbol{v}^\top \boldsymbol{G}(\boldsymbol{Y})\boldsymbol{v} \tag{6}$$

yielding the solution $\boldsymbol{v}^* = \frac{1}{M}\boldsymbol{G}(\boldsymbol{Y})^{-1}\boldsymbol{Y}^\top\boldsymbol{z}$. We are interested in quantifying similarity between $\boldsymbol{X}$ and $\boldsymbol{Y}$ by comparing the behavior of these optimal decoders. A straightforward way to quantify the alignment between the decoded signals is to compute their inner product:

$$\langle \boldsymbol{X}\boldsymbol{w}^*, \boldsymbol{Y}\boldsymbol{v}^*\rangle = \boldsymbol{w}^{*\top}\boldsymbol{X}^\top\boldsymbol{Y}\boldsymbol{v}^* = \boldsymbol{z}^\top\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^\top\boldsymbol{z} \tag{7}$$

where we have leveraged eq. (2) and the analogous closed form expression for $\boldsymbol{v}^*$ while introducing the normalized representations:

$$\tilde{\boldsymbol{X}} = \tfrac{1}{\sqrt{M}}\boldsymbol{X}\boldsymbol{G}(\boldsymbol{X})^{-1/2} \quad \text{and} \quad \tilde{\boldsymbol{Y}} = \tfrac{1}{\sqrt{M}}\boldsymbol{Y}\boldsymbol{G}(\boldsymbol{Y})^{-1/2} \tag{8}$$

Equation (7) is a reasonable quantification of similarity between $\boldsymbol{X}$ and $\boldsymbol{Y}$ with respect to a fixed decoding task specified by $\boldsymbol{z} \in \mathbb{R}^M$. This may be a fruitful approach for hypothesis-driven comparisons of neural systems, and we comment further on this possibility in our discussion.

On the other hand, many neural representations support a variety of downstream behavioral tasks. For example, features extracted in early stages of visual processing (edges and textures) can be used for a variety of downstream tasks such as object classification, segmentation, image compression, et cetera. How can eq. (7) be adapted to summarize the similarity between $\boldsymbol{X}$ and $\boldsymbol{Y}$ over a variety of decoding tasks? We explore three simple ideas.

Option 1 – best case. An optimistic measure of representational similarity between networks is to find the decoding task $\boldsymbol{z} \in \mathbb{R}^M$ that results in a maximal decoder alignment. Formally,

$$\max_{\|\boldsymbol{z}\|_2=1} \langle \boldsymbol{X}\boldsymbol{w}^*, \boldsymbol{Y}\boldsymbol{v}^* \rangle = \max_{\|\boldsymbol{z}\|_2=1} \boldsymbol{z}^\top \tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^\top \boldsymbol{z} \tag{9}$$

where the constraint that $\|\boldsymbol{z}\|_2 = 1$ is needed to keep the maximal value finite.

Option 2 – worst case. Another approach is to find the decoding task that minimizes alignment:

$$\min_{\|\boldsymbol{z}\|_2=1} \langle \boldsymbol{X}\boldsymbol{w}^*, \boldsymbol{Y}\boldsymbol{v}^* \rangle = \min_{\|\boldsymbol{z}\|_2=1} \boldsymbol{z}^\top \tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^\top \boldsymbol{z} \tag{10}$$

which is obviously the pessimistic counterpart to option 1 above.

Option 3 – average case. Instead of maximizing or minimizing over $\boldsymbol{z}$, we can approach the problem by averaging over a distribution of decoding tasks. Formally, we can quantify alignment by:

$$\mathbb{E}\langle \boldsymbol{X}\boldsymbol{w}^*, \boldsymbol{Y}\boldsymbol{v}^* \rangle \tag{11}$$

where the expectation is taken with respect to $\boldsymbol{z} \sim P_{\boldsymbol{z}}$.

In the next subsection, we show that the third option is closely related to several existing neural representational similarity measures. However, all three quantities are potentially of interest and they are easy to compute as we document below in propositions 1 and 2. Deriving these results is very straightforward, so we attach the proofs directly below.

Proposition 1. *The best case and worst case decoding alignment scores, eqs.* (9) *and* (10)*, are respectively given by the largest and smallest eigenvalues of:*

$$\frac{1}{2}\left(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^\top + \tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^\top \tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top\right) \tag{12}$$

Proposition 2. *The average case decoding alignment score, eq.* (11)*, is given by:*

$$\mathbb{E}\langle \boldsymbol{X}\boldsymbol{w}^*, \boldsymbol{Y}\boldsymbol{v}^* \rangle = \mathrm{Tr}[\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top] \tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^\top] \tag{13}$$

Proof of proposition 1. Let $\boldsymbol{A} \in \mathbb{R}^{M \times M}$ be any matrix, potentially non-symmetric. We can express $\boldsymbol{A}$ as a sum of a symmetric and skew symmetric matrix, $\boldsymbol{A} = \boldsymbol{B} + \boldsymbol{C}$, where:

$$\boldsymbol{B} = \frac{1}{2}\left(\boldsymbol{A} + \boldsymbol{A}^\top\right) \qquad \text{and} \qquad \boldsymbol{C} = \frac{1}{2}\left(\boldsymbol{A} - \boldsymbol{A}^\top\right) \tag{14}$$

For any $\boldsymbol{z} \in \mathbb{R}^M$, we have $\boldsymbol{z}^\top \boldsymbol{A}\boldsymbol{z} = \boldsymbol{z}^\top \boldsymbol{B}\boldsymbol{z} + \boldsymbol{z}^\top \boldsymbol{C}\boldsymbol{z}$. It is easy to verify that $\boldsymbol{z}^\top \boldsymbol{C}\boldsymbol{z}$ equals zero. Thus, $\boldsymbol{z}^\top \boldsymbol{A}\boldsymbol{z} = \boldsymbol{z}^\top \boldsymbol{B}\boldsymbol{z}$, and maximizing over $\boldsymbol{z}$ subject to $\|\boldsymbol{z}\|_2$ yields the largest eigenvalue of $\boldsymbol{B}$ (since it is symmetric). Likewise, minimizing yields the smallest eigenvalue of $\boldsymbol{B}$. Substituting $\boldsymbol{A} = \tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^\top$ into this argument proves the proposition. □

Proof of proposition 2. Using eq. (2) and the analogous closed form expression for $\boldsymbol{v}^*$, we have:

$$\mathbb{E}\langle \boldsymbol{X}\boldsymbol{w}^*, \boldsymbol{Y}\boldsymbol{v}^* \rangle = \mathbb{E}[\boldsymbol{z}^\top \tfrac{1}{M}\boldsymbol{X}\boldsymbol{G}(\boldsymbol{X})^{-1}\boldsymbol{X}^\top \tfrac{1}{M}\boldsymbol{Y}\boldsymbol{G}(\boldsymbol{Y})^{-1}\boldsymbol{Y}^\top \boldsymbol{z}] = \mathbb{E}[\boldsymbol{z}^\top \tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^\top \boldsymbol{z}]$$

where the final equality introduces the change of variables in eq. (8). Using the cyclic property of the trace operator and linearity of expectation to manipulate this further, we get:

$$\mathbb{E}[\boldsymbol{z}^\top \tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^\top \boldsymbol{z}] = \mathrm{Tr}[\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^\top \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top] \tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top]$$

Taking the transpose inside the trace yields eq. (13). □

3 Interpretations of CKA and related measures when $\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top] = \boldsymbol{I}$

Intuitively, proposition 2 says that the expected inner product between optimal linear readouts is equal to a weighted matrix inner product between normalized kernel matrices $\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top$ and $\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^\top$ (with respect to a positive semi-definite weighting matrix $\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top]$, which of course depends on the distribution of decoding targets $\boldsymbol{z}$). Such kernel matrices and inner products are important for several methods of quantifying representational similarity. We can therefore leverage this result to provide new interpretations of several distance measures. The most basic idea is to consider a distribution over $\boldsymbol{z}$ which satisfies $\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top] = \boldsymbol{I}$. This leads us to measure distances using standard Euclidean geometry, which we state as the following corollary to proposition 2.

Corollary 1. *For any distribution over decoding targets satisfying $\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top] = \boldsymbol{I}$, the Frobenius inner product between normalized kernel matrices $\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top$ and $\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^\top$ is equal to the expected alignment between decoded signals:*

$$\mathbb{E}\langle \boldsymbol{X}\boldsymbol{w}^*, \boldsymbol{Y}\boldsymbol{v}^* \rangle = \mathrm{Tr}[\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top \tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^\top] \tag{15}$$

Further, when $\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top] = \boldsymbol{I}$, the squared Euclidean distance between $\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top$ and $\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^\top$ is equal to the expected squared Euclidean distance between decoded signals:

$$\mathbb{E}\|\boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{Y}\boldsymbol{v}^*\|_2^2 = \tfrac{1}{M^2}\|\boldsymbol{X}\boldsymbol{G}(\boldsymbol{X})^{-1}\boldsymbol{X}^\top - \boldsymbol{Y}\boldsymbol{G}(\boldsymbol{Y})^{-1}\boldsymbol{Y}^\top\|_F^2 = \|\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top - \tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^\top\|_F^2 \tag{16}$$

where the equality on the left hand side follows from the identity $\mathbb{E}\|\boldsymbol{A}\boldsymbol{z}\|_2^2 = \|\boldsymbol{A}\|_F^2$ for some matrix $\boldsymbol{A} \in \mathbb{R}^{M\times M}$ and $\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top] = \boldsymbol{I}$, and the equality on the right hand side follows from the definitions in eq. (8). This corollary can be used to unify several neural representational similarity measures, each corresponding to a different choice of the penalty function $\boldsymbol{G}(\boldsymbol{X})$. Further, it yields an interpretation of each measure in terms of the expected overlap or difference in decoding readouts between networks.

First, for any $b > 0$, the following choice of $\boldsymbol{G}(\boldsymbol{X})$ yields the **linear CKA** score [6, 18]:

$$\boldsymbol{G}(\boldsymbol{X}) = b\boldsymbol{I} \quad \Rightarrow \quad \frac{\mathbb{E}\langle \boldsymbol{X}\boldsymbol{w}^*, \boldsymbol{Y}\boldsymbol{v}^* \rangle}{\sqrt{\mathbb{E}\langle \boldsymbol{X}\boldsymbol{w}^*, \boldsymbol{X}\boldsymbol{w}^* \rangle \mathbb{E}\langle \boldsymbol{Y}\boldsymbol{v}^*, \boldsymbol{Y}\boldsymbol{v}^* \rangle}} = \frac{\mathrm{Tr}[\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{Y}\boldsymbol{Y}^\top]}{\|\boldsymbol{X}\boldsymbol{X}^\top\|_F\|\boldsymbol{Y}\boldsymbol{Y}^\top\|_F} = \mathrm{CKA}(\boldsymbol{X}, \boldsymbol{Y})$$

where we have normalized the expected inner product to obtain a similarity score ranging between 0 and 1. Note that the centering step of CKA is taken care of by our assumption, stated at the beginning of section 2, that the columns of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are preprocessed to have mean zero. Thus, linear CKA be interpreted as a normalized alignment score over optimal linear decoders with quadratic weight regularization.

The CKA score is closely related to Euclidean distance, so it is not surprising that we can extend our analysis to other methods that utilize this distance. For example, the sample **GULP distance**, proposed by [3], is the equal to the Euclidean distance between normalized kernel matrices with a regularization parameter $\lambda > 0$. In particular,

$$\boldsymbol{G}(\boldsymbol{X}) = \boldsymbol{X}^\top\boldsymbol{X} + \lambda\boldsymbol{I} \quad \Rightarrow \quad \mathbb{E}\|\boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{Y}\boldsymbol{v}^*\|_2^2 = \|\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top - \tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^\top\|_F^2 = \hat{\mathrm{GULP}}^2(\boldsymbol{X}, \boldsymbol{Y})$$

yields the plug-in estimate of the squared GULP distance (see sec 3 in [3]). We therefore see that the Euclidean distance between normalized kernel matrices can be interpreted as equal to the expected squared Euclidean distance between decoded signals, or also as an approximation of the population formulation of the GULP distance established in [3] which, contrary to our framework, takes an expectation over the distribution of inputs and a supremum over all regression functions.

Canonical correlation analysis (CCA) is another method that has been used to compare neural representations that fits nicely into our framework [24, 30]. When the two networks have the same dimension, i.e. $N = N_X = N_Y$, the output of CCA is a sequence of N canonical correlation coefficients $1 \geq \rho_1 \geq \cdots \geq \rho_N \geq 0$. The average squared canonical coefficients is sometimes used as a measure of similarity,[1] and Kornblith et al. [18] showed this to be related to the linear CKA score

[1] In this statistics literature, this summary statistic is sometimes called *Yanai's generalized generalized coefficient of determination* [25, 32].

on whitened representations. In particular, assuming $\boldsymbol{X}^\top \boldsymbol{X}$ and $\boldsymbol{Y}^\top \boldsymbol{Y}$ are invertible, we have:

$$\boldsymbol{G}(\boldsymbol{X}) = \boldsymbol{X}^\top \boldsymbol{X} \quad \Rightarrow \quad \frac{\mathbb{E}\langle \boldsymbol{X}\boldsymbol{w}^*, \boldsymbol{Y}\boldsymbol{v}^* \rangle}{\sqrt{\mathbb{E}\langle \boldsymbol{X}\boldsymbol{w}^*, \boldsymbol{X}\boldsymbol{w}^* \rangle \mathbb{E}\langle \boldsymbol{Y}\boldsymbol{v}^*, \boldsymbol{Y}\boldsymbol{v}^* \rangle}} = \frac{1}{N}\sum_{i=1}^{N} \rho_i^2 \tag{17}$$

In practice, it is often useful to incorporate regularization into CCA because the gram matrices $\boldsymbol{X}^\top \boldsymbol{X}$ and $\boldsymbol{Y}^\top \boldsymbol{Y}$ may be ill-conditioned. As mentioned above, the regularization used in GULP distance corresponds to choosing $\boldsymbol{G}(\boldsymbol{X}) = \boldsymbol{X}^\top \boldsymbol{X} + \lambda \boldsymbol{I}$. A similar approach proposed in [30] is to use $\boldsymbol{G}(\boldsymbol{X}) = (1-\alpha)\boldsymbol{X}^\top \boldsymbol{X} + \alpha \boldsymbol{I}$ for a hyperparameter $0 \leq \alpha \leq 1$. This effectively allows for a continuous interpolation between a CCA-based similarity score and linear CKA.

Finally, we note a connection to recent work by Giaffar et al. [12], who studied the following quantity, which they call the **Effective Number of Shared Dimensions (ENSD)**:

$$\text{ENSD}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{\text{Tr}[\boldsymbol{X}\boldsymbol{X}^\top]\,\text{Tr}[\boldsymbol{Y}\boldsymbol{Y}^\top]\,\text{Tr}[\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{Y}\boldsymbol{Y}^\top]}{\text{Tr}[\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{X}^\top]\,\text{Tr}[\boldsymbol{Y}\boldsymbol{Y}^\top\boldsymbol{Y}\boldsymbol{Y}^\top]} \tag{18}$$

This is simply a rescaling of the inner product, $\text{Tr}[\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{Y}\boldsymbol{Y}^\top]$ and therefore easily fits within our framework using the following choice for $\boldsymbol{G}(\boldsymbol{X})$:

$$\boldsymbol{G}(\boldsymbol{X}) = \frac{\text{Tr}[\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{X}^\top]}{M\,\text{Tr}[\boldsymbol{X}\boldsymbol{X}^\top]}\boldsymbol{I} \quad \Rightarrow \quad \mathbb{E}\langle \boldsymbol{X}\boldsymbol{w}^*, \boldsymbol{Y}\boldsymbol{v}^* \rangle = \text{ENSD}(\boldsymbol{X}, \boldsymbol{Y}) \tag{19}$$

One of the most interesting features of ENSD is it's connection to the **participation ratio**, $\mathcal{R}$:

$$\mathcal{R}(\boldsymbol{X}\boldsymbol{X}^\top) = \frac{\left(\text{Tr}[\boldsymbol{X}\boldsymbol{X}^\top]\right)^2}{\text{Tr}[\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{X}^\top]} = \text{ENSD}(\boldsymbol{X}, \boldsymbol{X}) \tag{20}$$

which is used within physics and computational neuroscience (e.g. [11, 7, 26]) as a continuous analogue to the rank of a matrix, and is a measure of the effective dimensionality of the data matrix $\boldsymbol{X}$. In particular, one can show that $1 \leq \mathcal{R}(\boldsymbol{X}\boldsymbol{X}^\top) \leq \text{rank}(\boldsymbol{X}\boldsymbol{X}^\top)$, and that these two inequalities respectively saturate when $\boldsymbol{X}\boldsymbol{X}^\top$ has only one nonzero eigenvalue and when all nonzero eigenvalues are equal. It is also interesting to note that the *inverse* participation ratio can be captured by computing the expected norm of the optimal linear readout under a very simple choice for $\boldsymbol{G}(\boldsymbol{X})$:

$$\boldsymbol{G}(\boldsymbol{X}) = \frac{\text{Tr}[\boldsymbol{X}\boldsymbol{X}^\top]}{M}\boldsymbol{I} \quad \Rightarrow \quad \mathbb{E}\langle \boldsymbol{X}\boldsymbol{w}^*, \boldsymbol{X}\boldsymbol{w}^* \rangle = \mathbb{E}\|\boldsymbol{X}\boldsymbol{w}^*\|^2 = \frac{1}{\mathcal{R}(\boldsymbol{X}\boldsymbol{X}^\top)} \tag{21}$$

We reiterate that all of the results enumerated above depend on choosing a distribution over decoding targets which satisfies $\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top] = \boldsymbol{I}$. This enforces the decoding targets $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_M$ to be uncorrelated random variables with unit variance. The unit variance constraint is analogous to the constraint that $\|\boldsymbol{z}\|_2 = 1$ that we imposed when maximizing or minimizing over $\boldsymbol{z}$ in eqs. (9) and (10). Indeed, scaling the variance of the $\boldsymbol{z}_i$'s simply scales the magnitude of $\mathbb{E}\langle \boldsymbol{X}\boldsymbol{w}^*, \boldsymbol{Y}\boldsymbol{v}^* \rangle$. Thus, setting the variance of each sample to one (or some other constant) is reasonable. The constraint that $\mathbb{E}[\boldsymbol{z}_i\boldsymbol{z}_j] = 0$ for all $i \neq j$ may be relaxed, but we leave a detailed investigation of this possibility to future work.

4 Relating the Shape of Neural Representations to Decoding

In corollary 1 we saw that the Euclidean distance and Frobenius inner product between $\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top$ and $\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^\top$ enjoys a nice interpretation in terms of (average) decoder similarity. This does not provide a completely satisfactory answer of our original question: What does the *geometry* of a neural representation imply about its function (i.e. decoder performance)? As discussed in section 1, comparing neural representations through linear kernel matrices, $\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top$ and $\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^\top$, is motivated from geometric principles—namely, their invariance to orthogonal transformations. Indeed, for any orthogonal matrix $\boldsymbol{R}$, one can verify that:

$$\boldsymbol{G}(\boldsymbol{X}\boldsymbol{R})^{-1} = \boldsymbol{R}^\top \boldsymbol{G}(\boldsymbol{X})^{-1}\boldsymbol{R} \tag{22}$$

whenever $\boldsymbol{G}(\boldsymbol{X})$ is parameterized as in eq. (4). Therefore, then the transformation $\boldsymbol{X} \mapsto \boldsymbol{X}\boldsymbol{R}$ leaves the whitened linear kernel matrix unchanged. That is, if $\boldsymbol{X}' = \boldsymbol{X}\boldsymbol{R}$, we have:

$$\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top = \tfrac{1}{M}\boldsymbol{X}\boldsymbol{G}(\boldsymbol{X})^{-1}\boldsymbol{X}^\top = \tfrac{1}{M}\boldsymbol{X}\boldsymbol{R}\boldsymbol{G}(\boldsymbol{X}\boldsymbol{R})^{-1}\boldsymbol{R}^\top\boldsymbol{X}^\top = \tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}'^\top \tag{23}$$

The **Procrustes shape distance**, $\mathcal{P}$, is another popular method which is invariant to orthogonal transformations [30, 8]. This approach connects to a broader and decades-old literature on *shape analysis* [9, 14] whose applications include anatomical comparisons across biological species [10], and analysis of planar curves such as handwriting data [27]. If the two neural responses are mean-centered and have the same dimensionality, then the Procrustes distance is simply the minimal Euclidean distance obtained by rotating and reflecting one set of responses onto the other. That is, when $\boldsymbol{X} \in \mathbb{R}^{M\times N_X}$ and $\boldsymbol{Y} \in \mathbb{R}^{M\times N_Y}$ and we have $N = N_X = N_Y$, the Procrustes distance is:

$$\mathcal{P}(\boldsymbol{X}, \boldsymbol{Y}) = \min_{\boldsymbol{R}\in\mathcal{O}(N)} \|\boldsymbol{X} - \boldsymbol{Y}\boldsymbol{R}\|_F \tag{24}$$

When $N_X \neq N_Y$, a straightforward generalization of the Procrustes distance is to zero pad column-wise so that the matrix dimensions match. Geometrically, this can be interpreted as embedding the lower-dimensional point cloud into the higher-dimensional space and optimizing $\boldsymbol{R}$ within this space to align the point clouds.

The fact that Procrustes distance is defined in terms of an optimal alignment transformation is appealing both intuitively and because it suggests avenues to generalize the method, such as using permutations (instead of orthogonal transformations) to align representations [15]. Additionally, some empirical results have highlighted cases where Procrustes performs "better" than linear CKA [8, 5]. Rigorously defining and benchmarking what it means to be "better" in this context remains an open discussion [17], but it is nonetheless of great interest to understand what Procrustes distance and associated definitions of *shape* imply about neural decoding.

In appendix A.1 we prove the following result, which states that the Procrustes distance between appropriately normalized representations constrains the expected Euclidean distance between decoded signals, or equivalently, that the expected Euclidean distance between decoded signals bounds the Procrustes distance from above and below.

Proposition 3. *Defining normalization constants* $\alpha = \operatorname{Tr}\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top$ *and* $\beta = \operatorname{Tr}\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^\top$*, we have an upper and lower bound on the Procrustes distance:*

$$(\alpha+\beta) - \sqrt{(\alpha+\beta)^2 - \mathcal{R}_\Delta \mathbb{E}\|\boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{Y}\boldsymbol{v}^*\|_2^2} \leq \mathcal{P}(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}})^2 \leq \sqrt{\mathcal{R}_\Delta \mathbb{E}\|\boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{Y}\boldsymbol{v}^*\|_2^2} \tag{25}$$

where $\mathcal{R}_\Delta = \mathcal{R}(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top - \tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^\top)$ *is the participation ratio of the matrix difference* $\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top - \tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^\top$*. Equivalently, we may rewrite eq.* (25) *to find*

$$\frac{1}{\mathcal{R}_\Delta}\mathcal{P}(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}})^4 \leq \mathbb{E}\|\boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{Y}\boldsymbol{v}^*\|_2^2 \leq \frac{2(\alpha+\beta)\mathcal{P}(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}})^2 - \mathcal{P}(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}})^4}{\mathcal{R}_\Delta} \tag{26}$$

We derive these bounds by leveraging the equivalence between the Procrustes distance between representations $\mathcal{P}(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}})$ and a distance metric on positive semi-definite kernel matrices, the *Bures distance* [13].

We recall that the choice of function $\boldsymbol{G}(\cdot)$ in the decoding problem eq. (1) (which determines optimal readout weights $\boldsymbol{w}^*$ and $\boldsymbol{v}^*$) will set the normalization transformation of the representations $\tilde{\boldsymbol{X}} = \frac{1}{\sqrt{M}}\boldsymbol{X}\boldsymbol{G}(\boldsymbol{X})^{-1/2}$ and $\tilde{\boldsymbol{Y}} = \frac{1}{\sqrt{M}}\boldsymbol{Y}\boldsymbol{G}(\boldsymbol{Y})^{-1/2}$. When $\boldsymbol{G}(\cdot)$ is chosen such that $\operatorname{Tr}\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top = \operatorname{Tr}\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^\top = 1$[2], we can plot these bounds with $\alpha = \beta = 1$ for various values of participation ratio $\mathcal{R}_\Delta$ (fig. 2).

Both upper and lower bounds are saturated (appendix A.1), and reveal an allowed region of both Procrustes and expected Euclidean distance as a function of the participation ratio $\mathcal{R}_\Delta$. Under the

[2]For example $\boldsymbol{G}(\boldsymbol{X}) = \frac{N_X}{M}\boldsymbol{X}^\top\boldsymbol{X}$ or the CKA choice from above with $\boldsymbol{G}(\boldsymbol{X}) = (M^{-1}\operatorname{Tr}\boldsymbol{X}^\top\boldsymbol{X})\boldsymbol{I}$.

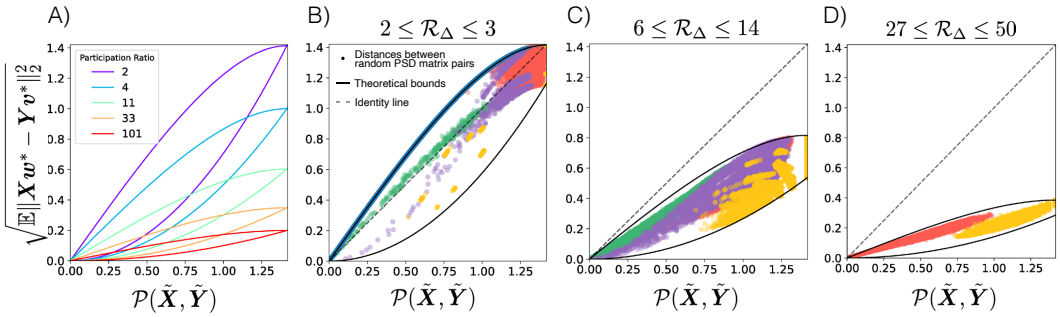

Figure 2: Bounds in eq. (25) plotted (solid lines) for varying participation ratio $\mathcal{R}_\Delta$ with $\alpha = \beta = 1$. A) Allowed regions (between the solid curves of like color) of Procrustes distance and expected Euclidean distance between decoded signals for different values of participation ratio $\mathcal{R}_\Delta$. B–D) Allowed regions for particular $\mathcal{R}_\Delta$ intervals (black solid lines) populated with calculated distances between pairs of randomly sampled positive semi-definite (PSD) matrices of size 50×50, subsampled to those that have $\mathcal{R}_\Delta$ in the particular interval (colored points). Different colors represent different random ensembles of positive semi-definite matrix pairs, which are described in appendix A.2.

$\alpha = \beta = 1$ normalization, the Procrustes distance may range between 0 and $\sqrt{2}$, and the expected Euclidean distance between decoded signals may take values between 0 and $\frac{2}{\sqrt{\mathcal{R}_\Delta}}$, which is reflected in fig. 2.

The bounds in eq. (25) and eq. (26) reveal an imperfect connection between the Procrustes distance and the expected Euclidean distance between decoded signals, such as the GULP distance. We recall that every choice of readout weight normalization $\boldsymbol{G}(\boldsymbol{X})$ leads to a measure of representational distance $\mathbb{E}\|\boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{Y}\boldsymbol{v}^*\|_2^2$, and that there is an associated Procrustes distance on the normalized representations $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Y}}$, $\mathcal{P}(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}})$. These bounds tell us that the expected Euclidean distance $\mathbb{E}\|\boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{Y}\boldsymbol{v}^*\|_2^2$ between decoded signals and this Procrustes distance constrain each other in a very particular way. Namely, when the Procrustes distance between normalized representations is small, then the expected Euclidean distance between decoded signals must also be small.[3] However, the converse is only necessarily true when the participation ratio of the difference of the kernel matrices $\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top - \tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^\top$ is also small (e.g. fig. 2 B). Indeed, as the dimensionality of $\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top - \tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^\top$ increases, we see that the Procrustes distance between normalized representations can be large, but the expected differences in decoding can be relatively small, with the effect becoming more and more exaggerated as $\mathcal{R}_\Delta$ becomes large (fig. 2 D). In appendix A.1 we show that for $\alpha = \beta = 1$, the minimum possible participation ratio $\mathcal{R}_\Delta$ is 2. Therefore, fig. 2 also illustrates another interesting phenomenon, namely that the upper left-hand corner of these plots is impossible to populate. That is, in the case of large expected difference in decoded signals, one can never simultaneously measure a small Procrustes distance between the normalized representations. Thus we observe quantitatively, as may be intuitive, that the Procrustes distance offers a more strict notion of geometric dissimilarity than the expected Euclidean distance between decoded signals.

5 Discussion

Understanding meaningful ways to measure similarities between neural representations is clearly a very complex problem. The literature demonstrates a proliferation of different techniques [16, 28], but an underdeveloped understanding of how these methods relate to each other. Less still is known mathematically about how representational similarity measures interact with functional similarity or the amount and type of information that is decodable from the representation. This paper presents a theoretical framework centered around the linear decoding of information from representations, which allows us to understand some existing popular methods for measuring representational similarity as alignment scores or distances between optimal linear decoders with different choices of weight

[3] The scale of "small" and "large" here are determined by α, β, and $\mathcal{R}_\Delta$, as the range of the Procrustes distance is between 0 and $\sqrt{\alpha + \beta}$, while range of the expected Euclidean distance between decoded signals is between 0 and $\frac{\alpha+\beta}{\sqrt{\mathcal{R}_\Delta}}$.

regularization. These connections relied on averaging the expected alignment between decoded signals over a distribution of decoding targets $\boldsymbol{z}$ with $\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top] = \boldsymbol{I}$. In the future it could be interesting to explore modifying these assumptions. For instance, instead of maximizing, minimizing, or taking the expectation over decoding targets, are there potentially interesting sets of fixed decoding targets that make sense when comparing networks in particular contexts? Furthermore, here we focused on linear regression as a decoding method; a potentially interesting line of future work could be to extend this framework to linear classifiers (e.g. support vector machines or multi-class logistic regression). Lastly, in this paper we considered quantifying representational similarity across a finite set of M stimulus conditions. However, these finite-dimensional approaches can sometimes be framed as approximations or estimators for a population version of the problem. For example, the framework in [23] for the Procrustes distance, or [3] for the GULP distance (the plugin estimator of which is a special case of our framework as we saw above). Developing the general framework of this paper in the $M \to \infty$ regime, and studying the behavior of estimators for similarity scores in the limited sample regime could be an important topic of future work.

References

[1] Guillaume Alain and Yoshua Bengio. *Understanding intermediate layers using linear classifier probes*. 2017. URL: https://openreview.net/forum?id=HJ4-rAVtl.

[2] Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. "On the Bures–Wasserstein distance between positive definite matrices". *Expositiones Mathematicae* 37.2 (2019), pp. 165–191. URL: https://www.sciencedirect.com/science/article/pii/S0723086918300021.

[3] Enric Boix-Adsera, Hannah Lawrence, George Stepaniants, and Philippe Rigollet. "GULP: a prediction-based metric between representations". *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh. Vol. 35. Curran Associates, Inc., 2022, pp. 7115–7127.

[4] Rosa Cao. "Putting representations to use". *Synthese* 200.2 (2022), p. 151.

[5] Nathan Cloos, Moufan Li, Markus Siegel, Scott L Brincat, Earl K Miller, Guangyu Robert Yang, and Christopher J Cueva. "Differentiable optimization of similarity scores between models and brains". *arXiv preprint arXiv:2407.07059* (2024).

[6] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. "Algorithms for learning kernels based on centered alignment". *The Journal of Machine Learning Research* 13 (2012), pp. 795–828.

[7] Paulina Dąbrowska, Nicole Voges, Michael von Papen, Junji Ito, David Dahmen, Alexa Riehle, Thomas Brochier, and Sonja Grün. "On the Complexity of Resting State Spiking Activity in Monkey Motor Cortex". *Cerebral Cortex Communications* 2 (2021).

[8] Frances Ding, Jean-Stanislas Denain, and Jacob Steinhardt. "Grounding Representation Similarity Through Statistical Testing". *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan. Vol. 34. Curran Associates, Inc., 2021, pp. 1556–1568.

[9] Ian L Dryden and Kanti V Mardia. *Statistical shape analysis: with applications in R*. John Wiley & Sons, 2016.

[10] Ashraf M T Elewa, ed. *Morphometrics*. 2004th ed. Berlin, Germany: Springer, 2004.

[11] Peiran Gao and Surya Ganguli. "On simplicity and complexity in the brave new world of large-scale neuroscience". *Current Opinion in Neurobiology* 32 (2015). Large-Scale Recording Technology (32), pp. 148–155. URL: https://www.sciencedirect.com/science/article/pii/S0959438815000768.

[12] Hamza Giaffar, Camille Rullán Buxó, and Mikio Aoi. "The Effective Number of Shared Dimensions Between Paired Datasets". *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*. Ed. by Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li. Vol. 238. Proceedings of Machine Learning Research. PMLR, 2024, pp. 4249–4257. URL: https://proceedings.mlr.press/v238/giaffar24a.html.

[13] Sarah E. Harvey, Brett W. Larsen, and Alex H. Williams. "Duality of Bures and Shape Distances with Implications for Comparing Neural Representations". *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models*. Ed. by Marco Fumero, Emanuele Rodolá, Clementine Domine, Francesco Locatello, Karolina Dziugaite, and Caron Mathilde. Vol. 243. Proceedings of Machine Learning Research. PMLR, 2024, pp. 11–26.

[14] David George Kendall, Dennis Barden, Thomas K Carne, and Huiling Le. *Shape and shape theory*. John Wiley & Sons, 2009.

[15] Meenakshi Khosla and Alex H Williams. "Soft Matching Distance: A metric on neural representations that captures single-neuron tuning". *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models*. Ed. by Marco Fumero, Emanuele Rodolá, Clementine Domine, Francesco Locatello, Karolina Dziugaite, and Caron Mathilde. Vol. 243. Proceedings of Machine Learning Research. PMLR, 2024, pp. 326–341.

[16] Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. "Similarity of neural network models: A survey of functional and representational measures". *arXiv preprint arXiv:2305.06329* (2023).

[17] Max Klabunde, Tassilo Wald, Tobias Schumacher, Klaus Maier-Hein, Markus Strohmaier, and Florian Lemmerich. "ReSi: A Comprehensive Benchmark for Representational Similarity Measures". *arXiv preprint arXiv:2408.00531* (2024).

[18] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. "Similarity of Neural Network Representations Revisited". *International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 3519–3529.

[19] Nikolaus Kriegeskorte and Pamela K Douglas. "Interpreting encoding and decoding models". *Current Opinion in Neurobiology* 55 (2019). Machine Learning, Big Data, and Neuroscience, pp. 167–179. URL: `https://www.sciencedirect.com/science/article/pii/S0959438818301004`.

[20] Nikolaus Kriegeskorte and Xue-Xin Wei. "Neural tuning and representational geometry". *Nature Reviews Neuroscience* 22.11 (2021), pp. 703–718.

[21] Andrew Kyle Lampinen, Stephanie CY Chan, and Katherine Hermann. "Learned feature representations are biased by complexity, learning order, position, and more". *arXiv preprint arXiv:2405.05847* (2024).

[22] Niru Maheswaranathan, Alex Williams, Matthew Golub, Surya Ganguli, and David Sussillo. "Universality and individuality in neural dynamics across large populations of recurrent networks". *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 2019.

[23] Dean A Pospisil, Brett W. Larsen, Sarah E Harvey, and Alex H Williams. "Estimating Shape Distances on Neural Representations with Limited Samples". *International Conference on Learning Representations*. 2024. URL: `https://openreview.net/forum?id=kvByNnMERu`.

[24] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. "Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability". *Advances in neural information processing systems* 30 (2017).

[25] James O Ramsay, Jos ten Berge, and George PH Styan. "Matrix correlation". *Psychometrika* 49.3 (1984), pp. 403–423.

[26] Stefano Recanatesi, Serena Bradde, Vijay Balasubramanian, Nicholas A. Steinmetz, and Eric Shea-Brown. "A scale-dependent measure of system dimensionality". *Patterns* 3.8 (2022), p. 100555. URL: `https://www.sciencedirect.com/science/article/pii/S266638992200160X`.

[27] Anuj Srivastava and Eric P Klassen. *Functional and shape data analysis*. 1st ed. Springer Series in Statistics. New York, NY: Springer, 2016.

[28] Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C Love, Erin Grant, Jascha Achterberg, Joshua B Tenenbaum, et al. "Getting aligned on representational alignment". *arXiv preprint:2310.13018* (2023).

[29] John Watrous. *The theory of quantum information*. Cambridge University Press, 2018.

[30] Alex H Williams, Erin Kunz, Simon Kornblith, and Scott Linderman. "Generalized shape metrics on neural representations". *Advances in Neural Information Processing Systems* 34 (2021), pp. 4738–4750.

[31] Daniel LK Yamins and James J DiCarlo. "Using goal-driven deep learning models to understand sensory cortex". *Nature neuroscience* 19.3 (2016), pp. 356–365.

[32] Haruo Yanai. “Unification of various techniques of multivariate analysis by means of generalized coefficient of determination”. *Kodo Keiryogaku* 1.1 (1974), pp. 46–54.

A Appendix

A.1 Bounds on the Procrustes distance in terms of the Euclidean distance

Here we derive upper and lower bounds on the Procrustes distance between neural representations $\boldsymbol{X} \in \mathbb{R}^{M \times N_X}$ and $\boldsymbol{Y} \in \mathbb{R}^{M \times N_Y}$ in terms of the Euclidean distance between linear kernel matrices $\boldsymbol{K}_X = \boldsymbol{X}\boldsymbol{X}^\top$ and $\boldsymbol{K}_Y = \boldsymbol{Y}\boldsymbol{Y}^\top$. This result is applied in the main text to relate the expected Euclidean distance between decoded signals $\mathbb{E}\|\boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{Y}\boldsymbol{v}^*\|_2^2 = \|\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top - \tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^\top\|_F^2$ to the Procrustes distance $\mathcal{P}(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}})$.

We will make use of the equivalence between the Procrustes distance between representation matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ and a notion of distance on the space of positive semi-definite kernel matrices $\boldsymbol{K}_X = \boldsymbol{X}\boldsymbol{X}^\top$ and $\boldsymbol{K}_Y = \boldsymbol{Y}\boldsymbol{Y}^\top$ called the Bures distance, defined as

$$d_B(\boldsymbol{K}_X, \boldsymbol{K}_Y) = \sqrt{\mathrm{Tr}[\boldsymbol{K}_X] + \mathrm{Tr}[\boldsymbol{K}_Y] - 2\,\mathrm{Tr}\left[\left(\boldsymbol{K}_X^{1/2}\boldsymbol{K}_Y\boldsymbol{K}_X^{1/2}\right)^{1/2}\right]}. \tag{27}$$

The third trace term on the right hand side of eq. (27) is often called the *fidelity*, defined as

$$\mathcal{F}(\boldsymbol{K}_X, \boldsymbol{K}_Y) = \mathrm{Tr}\left[\left(\boldsymbol{K}_X^{1/2}\boldsymbol{K}_Y\boldsymbol{K}_X^{1/2}\right)^{1/2}\right]. \tag{28}$$

A.1.1 Lower bound

For the lower bound, we are inspired by the Fuchs-van-de-Graaf inequalites that are used in quantum information theory to assert an approximate equivalence between two measures of quantum state similarity, the *trace distance* and the *fidelity*. This approach is relevant from a mathematical perspective for our purposes, since quantum states are represented by positive semi-definite matrices normalized to have trace 1. Here, we use a similar approach as that often used to derive the Fuchs-van de Graaf inequalites to instead relate the Euclidean distance and the Bures distance on positive semidefinite matrices, while also relaxing the trace 1 normalization.

We will make use of the following identity.

Lemma 1.

$$\|\alpha uu^\dagger - \beta vv^\dagger\|_* = \sqrt{(\alpha+\beta)^2 - 4\alpha\beta|\langle u, v\rangle|^2} \tag{29}$$

for all unit vectors u, v and all non-negative real numbers α and β.

Proof. First we recognize that the nuclear norm of a matrix can be calculated by summing the singular values of that matrix. Furthermore, if that matrix is Hermitian, the singular values are the absolute values of the eigenvalues. Define $\boldsymbol{M} = \alpha uu^\dagger - \beta vv^\dagger$. Our matrix M is Hermitian and has at most two eigenvalues, so we will look for expressions for these in terms of α and β. Note that if u and v are the same unit vector, then $\boldsymbol{M}$ has rank 1, so we expect the eigenvalues will depend also on the inner product $\langle u, v\rangle$.

The eigenvectors of $\boldsymbol{M}$ will be of the form $\psi = cu + dv$ for some scalar c and d. By direct calculation:

$$\begin{aligned} \boldsymbol{M}\psi =& \lambda\psi \\ \boldsymbol{M}(cu+dv) =& (\alpha c + \alpha d\langle u, v\rangle)u - (\beta c\langle v, u\rangle + \beta d)v \\ =& \lambda(cu+dv) \end{aligned} \tag{30}$$

Therefore, we see that the eigenvalues must satisfy both

$$\lambda = \frac{\alpha c + \alpha d\langle u, v\rangle}{c} \quad \text{and} \quad \lambda = -\frac{\beta c\langle v, u\rangle + \beta d}{d} \tag{31}$$

Setting these equal to each other and solving the resulting quadratic for c gives:

$$c = \frac{-d(\alpha+\beta) \pm d\sqrt{(\alpha+\beta)^2 - 4\alpha\beta|\langle v, u\rangle|^2}}{2\beta\langle v, u\rangle}. \tag{32}$$

So the two eigenvectors are proportional to

$$\psi_\pm = \left(\frac{(\alpha+\beta)}{2} \mp \frac{1}{2}\sqrt{(\alpha+\beta)^2 - 4\alpha\beta|\langle v, u\rangle|^2}\right)u - \beta\langle v, u\rangle v. \tag{33}$$

To find the eigenvalues, we combine eq. (31) and eq. (32), arriving at

$$\lambda_\pm = \frac{(\beta-\alpha)}{2} \pm \frac{1}{2}\sqrt{(\alpha+\beta)^2 - 4\alpha\beta|\langle v, u\rangle|^2}. \tag{34}$$

Finally, to calculate $\|\boldsymbol{M}\|_*$ we simply sum the magnitudes of these eigenvalues. By inspecting the term under the square root, we see that λ_- is always negative. Therefore, $|\lambda_+| + |\lambda_-| = \lambda_+ - \lambda_-$, and we find

$$\begin{aligned} \|\boldsymbol{M}\|_1 =& |\lambda_+| + |\lambda_-| \\ \implies \|\alpha uu^\dagger - \beta vv^\dagger\|_1 =& \sqrt{(\alpha+\beta)^2 - 4\alpha\beta|\langle v, u\rangle|^2} \end{aligned} \tag{35}$$

□

We now find a lower bound on the Bures distance between kernel matrices $\boldsymbol{K}_X$ and $\boldsymbol{K}_Y$ in terms of the Euclidean distance between the same matrices, $\|\boldsymbol{K}_X - \boldsymbol{K}_Y\|_F^2$. There are many choices of $\boldsymbol{X}$ and $\boldsymbol{Y}$ multiply to the same positive semi-definite kernel matrices via $\boldsymbol{K}_X = \boldsymbol{X}\boldsymbol{X}^\top$ and $\boldsymbol{K}_Y = \boldsymbol{Y}\boldsymbol{Y}^\top$. Define real positive scalars $\alpha = \operatorname{Tr}\boldsymbol{X}\boldsymbol{X}^\top$ and $\beta = \operatorname{Tr}\boldsymbol{Y}\boldsymbol{Y}^\top$, and the unit vectors

$$\hat{x} = \frac{\text{vec}(\boldsymbol{X})}{\|\text{vec}(\boldsymbol{X})\|_2} \quad \text{and} \quad \hat{y} = \frac{\text{vec}(\boldsymbol{Y})}{\|\text{vec}(\boldsymbol{Y})\|_2} \tag{36}$$

where $\text{vec}(\boldsymbol{X})$ is the vectorization linear transformation converting the $M \times N_X$-dimensional matrix $\boldsymbol{X}$ into an MN_X-dimensional vector. We will assume that the representations have been zero-padded such that $N_X = N_Y$.

By Uhlmann's theorem [29], there exists some choice of $\boldsymbol{X}$ and $\boldsymbol{Y}$, such that

$$|\langle\sqrt{\alpha}\hat{x}, \sqrt{\beta}\hat{y}\rangle| = \mathcal{F}(\boldsymbol{K}_X, \boldsymbol{K}_Y) \tag{37}$$

One can see this by writing the fidelity as a nuclear norm of $\boldsymbol{X}^\top\boldsymbol{Y}$ and considering the variational form of the nuclear norm as a maximization of the Hilbert-Schmidt inner product over unitary transformations (see [13] for a more explicit discussion of this).

Now use our identity from earlier with $u = \hat{x}$ and $v = \hat{y}$.

$$\begin{aligned} \|\alpha\hat{x}\hat{x}^\top - \beta\hat{y}\hat{y}^\top\|_1 =& \sqrt{(\alpha+\beta)^2 - 4\alpha\beta|\langle\hat{x}, \hat{y}\rangle|^2} \\ =& \sqrt{(\alpha+\beta)^2 - 4\mathcal{F}(\boldsymbol{K}_X, \boldsymbol{K}_Y)^2} \end{aligned} \tag{38}$$

The operation that takes $\alpha\hat{x}\hat{x}^\top$ and $\beta\hat{y}\hat{y}^\top$ to $\boldsymbol{K}_X$ and $\boldsymbol{K}_Y$, respectively, is called the partial trace. By the monotonicity of the nuclear norm under partial tracing [29], we have:

$$\|\boldsymbol{K}_X - \boldsymbol{K}_Y\|_* \leq \|\alpha\hat{x}\hat{x}^\top - \beta\hat{y}\hat{y}^\top\|_* = \sqrt{(\alpha+\beta)^2 - 4\mathcal{F}(\boldsymbol{K}_X, \boldsymbol{K}_Y)^2}. \tag{39}$$

With equality when $\boldsymbol{K}_X$ and $\boldsymbol{K}_Y$ are rank-1, that is $\boldsymbol{K}_X = \alpha\hat{x}\hat{x}^\top$ and $\boldsymbol{K}_Y = \beta\hat{y}\hat{y}^\top$.

Lastly, we rewrite the nuclear norm $\|\boldsymbol{K}_X - \boldsymbol{K}_Y\|_*$ in terms of the Euclidean distance and the *participation ratio* of the matrix $\boldsymbol{K}_X - \boldsymbol{K}_Y$. The participation ratio of a matrix $\boldsymbol{A} \in \mathbb{R}^{M \times N}$ is defined as

$$\mathcal{R}(\boldsymbol{A}) = \frac{\|\boldsymbol{A}\|_1^2}{\|\boldsymbol{A}\|_F^2} = \frac{(\sum_i \sigma_i)^2}{\sum_i \sigma_i^2}, \tag{40}$$

so we have

$$\|\boldsymbol{K}_X - \boldsymbol{K}_Y\|_* = \sqrt{\mathcal{R}_\Delta}\|\boldsymbol{K}_X - \boldsymbol{K}_Y\|_F \tag{41}$$

where $\mathcal{R}_\Delta = \mathcal{R}(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top - \tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^\top)$. We can also use the definition of the Bures distance eq. (27) to replace $\mathcal{F}(\boldsymbol{K}_X, \boldsymbol{K}_Y) = \frac{1}{2}(\alpha + \beta - d_B(\boldsymbol{K}_X, \boldsymbol{K}_Y)^2)$. Making these substitutions into eq. (39), and solving for $d_B(\boldsymbol{K}_X, \boldsymbol{K}_Y)^2$, we find:

$$d_B(\boldsymbol{K}_X, \boldsymbol{K}_Y)^2 \geq (\alpha + \beta) - \sqrt{(\alpha + \beta)^2 - \mathcal{R}_\Delta \|\boldsymbol{K}_X - \boldsymbol{K}_Y\|_F^2}. \tag{42}$$

This bound is saturated when the inequalty in eq. (39) is an equality, or when $\boldsymbol{K}_X$ and $\boldsymbol{K}_Y$ are both rank 1.

For a bound that is independent of the participation ratio, we could replace $\mathcal{R}_\Delta$ with its ostensible minimal value of 1. When $\boldsymbol{K}_X$ and $\boldsymbol{K}_Y$ are normalized to have trace 1, we would then have

$$d_B(\boldsymbol{K}_X, \boldsymbol{K}_Y)^2 \geq 2 - \sqrt{4 - \|\boldsymbol{K}_X - \boldsymbol{K}_Y\|_F^2} \approx \frac{1}{4}\|\boldsymbol{K}_X - \boldsymbol{K}_Y\|_F^2. \tag{43}$$

However, in this case that $\operatorname{Tr} \boldsymbol{K}_X = \operatorname{Tr} \boldsymbol{K}_Y = 1$, we can actually arrive at a tighter lower bound than eq. (43), because we can say something interesting about the minimal participation ratio of matrices of the form $\boldsymbol{K}_X - \boldsymbol{K}_Y$. It turns out that under this normalization, the participation ratio is lower bounded by 2, with the minimum of 2 achieved when $\boldsymbol{K}_X$ and $\boldsymbol{K}_Y$ are rank 1 (also in this case the bound in eq. (42) will be saturated).

Lemma 2. *For positive semidefinite matrices $\boldsymbol{K}_X$ and $\boldsymbol{K}_Y$ with $\operatorname{Tr} \boldsymbol{K}_X = \operatorname{Tr} \boldsymbol{K}_Y$,*

$$\mathcal{R}(\boldsymbol{K}_X - \boldsymbol{K}_Y) \geq 2. \tag{44}$$

Proof. By noting that $\boldsymbol{K}_X - \boldsymbol{K}_Y$ is Hermitian, we rewrite eq. (40) as

$$\mathcal{R}(\boldsymbol{K}_X - \boldsymbol{K}_Y) = \frac{(\sum_i |\lambda_i|)^2}{\sum_i |\lambda_i|^2} \tag{45}$$

where λ_i is the ith eigenvalue of the matrix difference $\boldsymbol{K}_X - \boldsymbol{K}_Y$. Next, we recognize that if $\operatorname{Tr} \boldsymbol{K}_X = \operatorname{Tr} \boldsymbol{K}_Y$, then $\operatorname{Tr}(\boldsymbol{K}_X - \boldsymbol{K}_Y) = 0$ and therefore the eigenvalues of $\boldsymbol{K}_X - \boldsymbol{K}_Y$ sum to 0. The set of eigenvalues $\{\lambda_1, \lambda_2, ..., \lambda_M\}$ can then always be partitioned into a set of positive eigenvalues $\{\lambda_1^+, \lambda_2^+, ..., \lambda_p^+\}$, and a set of negative eigenvalues $\{\lambda_1^-, \lambda_2^-, ..., \lambda_q^-\}$, with $p + q = M$. The negative eigenvalues must balance the positive eigenvalues in magnitude, so we must have

$$\sum_{i=1}^{p} \lambda_i^+ = \sum_{j=1}^{q} |\lambda_j^-|. \tag{46}$$

We can rewrite eq. (45) in terms of sums over the positive and negative eigenvalue sets:

$$
\begin{aligned}
\mathcal{R}(\boldsymbol{K}_X - \boldsymbol{K}_Y) &= \frac{4(\sum_{i=1}^p \lambda_i^+)^2}{\sum_{i=1}^p (\lambda_i^+)^2 + \sum_{j=1}^q |\lambda_j^-|^2} \\
&= \frac{4\big[\sum_{i=1}^p (\lambda_i^+)^2 + \sum_{j\neq k}^p \lambda_j^+ \lambda_k^+\big]}{(\sum_{i=1}^p \lambda_i^+)^2 - \sum_{j\neq k}^p \lambda_j^+ \lambda_k^+ + (\sum_{i=1}^q |\lambda_i^-|)^2 - \sum_{j\neq k}^q |\lambda_j^-||\lambda_k^-|} \\
&= \frac{4\big[\sum_{i=1}^p (\lambda_i^+)^2 + \sum_{j\neq k}^p \lambda_j^+ \lambda_k^+\big]}{2[(\sum_{i=1}^p (\lambda_i^+)^2) + \frac{1}{2}(\sum_{j\neq k}^p \lambda_j^+ \lambda_k^+ - \sum_{j\neq k}^q |\lambda_j^-||\lambda_k^-|)]} \\
&= 2\left[\frac{\sum_{i=1}^p (\lambda_i^+)^2 + \sum_{j\neq k}^p \lambda_j^+ \lambda_k^+}{\sum_{i=1}^p (\lambda_i^+)^2 + \frac{1}{2}(\sum_{j\neq k}^p \lambda_j^+ \lambda_k^+ - \sum_{j\neq k}^q |\lambda_j^-||\lambda_k^-|)}\right]
\end{aligned}
\tag{47}
$$

By inspection, the term in brackets is ≥ 1 (this can be seen by considering the signs and relative magnitudes of each of the summed terms). So we have

$$
\mathcal{R}(\boldsymbol{K}_X - \boldsymbol{K}_Y) \geq 2. \tag{48}
$$

□

The inequality eq. (44) is saturated when $\boldsymbol{K}_X - \boldsymbol{K}_Y$ is rank 2 and thus only has two equal magnitude and opposite signed eigenvalues, as can be seen from the first line of eq. (47). We can also use eq. (34) to see that when $\boldsymbol{K}_X$ and $\boldsymbol{K}_Y$ are rank 1 and trace 1, we have

$$
\mathcal{R}(\boldsymbol{K}_X - \boldsymbol{K}_Y) = \frac{\|\boldsymbol{K}_X - \boldsymbol{K}_Y\|_1^2}{\|\boldsymbol{K}_X - \boldsymbol{K}_Y\|_F^2} = \frac{(\alpha+\beta)^2 - 4\alpha\beta|\langle v, u\rangle|^2}{\frac{1}{2}[(\alpha+\beta)^2 - 4\alpha\beta|\langle v, u\rangle|^2]} = 2. \tag{49}
$$

For $\boldsymbol{K}_X$ and $\boldsymbol{K}_Y$ normalized to have trace 1, we can then state a tighter bound than eq. (43) using the minimal participation ratio in eq. (44). We have

$$
d_B(\boldsymbol{K}_X, \boldsymbol{K}_Y)^2 \geq 2 - \sqrt{4 - 2\|\boldsymbol{K}_X - \boldsymbol{K}_Y\|_F^2}. \tag{50}
$$

A.1.2 Upper bound

We can bound the Bures distance using its variational form [2],

$$
d_B^2(\boldsymbol{K}_X, \boldsymbol{K}_Y) = \min_{\boldsymbol{Q}\in\mathcal{O}(M)} \|\boldsymbol{K}_X^{1/2} - \boldsymbol{K}_Y^{1/2}\boldsymbol{Q}\|_F^2 \leq \|\boldsymbol{K}_X^{1/2} - \boldsymbol{K}_Y^{1/2}\|_F^2 \tag{51}
$$

The Powers-Størmer inequality implies that $\|\boldsymbol{K}_X^{1/2} - \boldsymbol{K}_Y^{1/2}\|_F^2 \leq \|\boldsymbol{K}_X - \boldsymbol{K}_Y\|_*$, so

$$
d_B^2(\boldsymbol{K}_X, \boldsymbol{K}_Y) \leq \|\boldsymbol{K}_X - \boldsymbol{K}_Y\|_*. \tag{52}
$$

Expressing the nuclear norm in terms of the participation ratio of $\boldsymbol{K}_X - \boldsymbol{K}_Y$ using eq. (41), we have an upper bound on the Bures distance:

$$
d_B^2(\boldsymbol{K}_X, \boldsymbol{K}_Y) \leq \sqrt{\mathcal{R}_\Delta}\|\boldsymbol{K}_X - \boldsymbol{K}_Y\|_F. \tag{53}
$$

The inequality in eq. (51) is saturated when the optimal orthogonal transformation that aligns $\boldsymbol{K}_X^{1/2}$ and $\boldsymbol{K}_Y^{1/2}$ is the identity. This occurs when $\boldsymbol{K}_X$ and $\boldsymbol{K}_Y$ are simultaneously diagonalizable. The Powers-Størmer inequality in eq. (52) is saturated when $\boldsymbol{K}_X$ and $\boldsymbol{K}_Y$ have eigenvalues that only take values in $\{0, 1\}$.

A.2 Figure 2 details

Figure 2 panels (B-D) show the allowed regions of Procrustes distance and expected Euclidean distance between decoded signals for representations normalized to have $\operatorname{Tr} \tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top = \operatorname{Tr} \tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^\top = 1$, with varying participation ratio $\mathcal{R}_\Delta$. We have populated the plots with points representing the respective distances between randomly sampled pairs of positive semi-definite matrices. These positive semi-definite matrices represent normalized kernel matrices $\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top$ and $\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^\top$. We have seen in the main text that the squared Euclidean distance between these kernel matrices is equivalent to the expected squared Euclidean distance between decoded signals, $\mathbb{E}\|\boldsymbol{X}\boldsymbol{w}^* - \boldsymbol{Y}\boldsymbol{v}^*\|_2^2 = \|\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top - \tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^\top\|_F^2$, when the distribution of decoding targets satisfies $\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top] = \boldsymbol{I}$. On the other hand, [13] demonstrated how the Procrustes distance $\mathcal{P}(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}})$ is equivalent to the Bures distance eq. (27), $d_B(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top, \tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^\top)$. Therefore, both distances on these plots can be calculated from samples of positive semi-definite matrix pairs.

Each color of points in fig. 2 (B-D) represents a different random ensemble of $M \times M$ PSD matrices with $M = 50$, which are then binned into one of the three plots if the participation ratio $\mathcal{R}_\Delta$ lies in the respective range indicated above each plot panel.

- Pink points are generated by sampling eigenvalues for the two PSD matrices from a Dirichlet distribution with concentration parameters logarithmically spaced between 10^{-3} and 10^3. Matrices of eigenvectors are then randomly sampled from the set of orthogonal matrices.
- Green points were generated by sampling a random matrix $\boldsymbol{A}$ of size $M \times r$ with each element drawn from the uniform distribution $\mathcal{U}[0, 1]$. A uniform random integer r was selected between 1 and 24. A PSD matrix was generated by multiplying $\boldsymbol{A}\boldsymbol{A}^\top$. Another PSD matrix was then generated by adding a randomly weighted matrix of the same dimension with standard normal distributed entries $\boldsymbol{N}$ to $\boldsymbol{A}$, so $\boldsymbol{B} = \boldsymbol{A} + \epsilon\boldsymbol{N}$, where $\epsilon \sim \mathcal{U}[0, 1]$. The distance between $\frac{1}{\operatorname{Tr}\boldsymbol{A}\boldsymbol{A}^\top}\boldsymbol{A}\boldsymbol{A}^\top$ and $\frac{1}{\operatorname{Tr}\boldsymbol{B}\boldsymbol{B}^\top}\boldsymbol{B}\boldsymbol{B}^\top$ is then computed using these two distance metrics.
- Blue points are seen to only occupy fig. 2 (B), as these correspond to distances between matrices of rank 1 and trace 1. As we saw earlier in this appendix, in this case the participation ratio of the difference between these two PSD matrices is always 2. The rank 1 matrices were sampled using the same method as the green points above, but setting $r = 1$.
- Purple points were generated by sampling one matrix from a Wishart distribution $W_M(r, \boldsymbol{I})$, with degrees of freedom r chosen as a uniformly random integer between 1 and 50. This matrix was normalized to have trace 1. The second matrix was generated by adding to the first matrix another matrix drawn from the Wishart distribution $W_M(n, \boldsymbol{I})$, with n a randomly selected integer between 1 and 10. Both matrices are normalized to have trace 1.
- Yellow points represent distances between PSD matrices that are simultaneously diagonalizable. These were generated by sampling M values uniformly between 0 and 1 for each matrix, but and setting these values to 0 if they fall below the threshold 0.6. The resultant values were then normalized to sum to 1, which became the eigenvalues for the two PSD matrices. A matrix of eigenvectors was then randomly selected from the set of orthogonal matrices.