Duality of Bures and Shape Distances with Implications for Comparing Neural Representations

Sarah E. Harvey[1,2] **Brett W. Larsen**[1,2,3] **Alex H. Williams**[1,2]
[1]New York University, Center for Neural Science, New York, NY, 10003
[2]Flatiron Institute, Center for Computational Neuroscience, New York, NY, 10010
[3]Flatiron Institute, Center for Computational Mathematics, New York, NY, 10010
{sharvey, brettlarsen, awilliams}@flatironinstitute.org

Editors: Marco Fumero, Emanuele Rodolà, Clementine Domine, Francesco Locatello, Gintare Karolina Dziugaite, Mathilde Caron

Abstract

A multitude of (dis)similarity measures between neural network representations have been proposed, resulting in a fragmented research landscape. Most of these measures fall into one of two categories. First, measures such as linear regression, canonical correlations analysis (CCA), and shape distances, all learn explicit mappings between neural units to quantify similarity while accounting for expected invariances. Second, measures such as representational similarity analysis (RSA), centered kernel alignment (CKA), and normalized Bures similarity (NBS) all quantify similarity in summary statistics, such as stimulus-by-stimulus kernel matrices, which are already invariant to expected symmetries. Here, we take steps towards unifying these two broad categories of methods by observing that the cosine of the Riemannian shape distance (from category 1) is equal to NBS (from category 2). We explore how this connection leads to new interpretations of shape distances and NBS, and draw contrasts of these measures with CKA, a popular similarity measure in the deep learning literature.

1 Introduction

Quantifying similarity between neural network representations is now a well-recognized topic in computational neuroscience and deep learning [22, 52]. In neuroscience, measures of representational similarity have been used to benchmark models of biological systems [21, 51], and to compare neural activity across different species [26]. In deep learning, they have been used to characterize learning dynamics [36], model robustness [18], and the effects of changing model architecture [31, 38].

Interest in this area has sparked a proliferation of measures to quantify representational (dis)similarity including: representational similarity analysis (RSA; [25]), centered kernel alignment (CKA; [23]), generalized shape distances [58], canonical correlations analysis (CCA; [44]), normalized Bures similarity (NBS; [54]), and the Riemannian covariance distance [50]. While all of these methods aim to quantify similar aspects of neural data, much more work is needed to formalize this intuition and to characterize the practical differences between these competing methods.

Here we develop a duality principle[1] that links shape distances [20, 58] to well-known quantities in optimal transport [32] and quantum information theory [40, 35, 57]. Although similar ideas

[1]Following Atiyah [2], we use the term *duality* to broadly mean a mathematical relationship that enables "two different points of view of looking at the same object."

Proceedings of the I edition of the Workshop on Unifying Representations in Neural Models (UniReps 2023).

were recently described in mathematical literature on infinite-dimensional covariance operators [34], these results appear thus far unnoticed within the computational neuroscience and machine learning communities. For example, we will see that two independently proposed (dis)similarity measures—the Riemannian shape distance and NBS—are essentially equivalent to each other, suggesting that this duality provides way to generalize the Riemannian shape distance to cases where the networks have differing dimensionality. Moreover, we point out CKA and NBS have been extensively compared by quantum information theorists as different measures of similarity, CKA corresponding to the normalized Hilbert-Schmidt inner product and NBS corresponding to the *fidelity* between positive semidefinite matrices with trace equal to 1 [30]. An important contribution of our work is to unify these disconnected literatures with a self-contained exposition, but we also aim to demonstrate how these connections can lead to novel theoretical analysis and insights in to representational (dis)similarity.

The rest of this manuscript is organized as follows. In section 2, we formalize the problem of comparing neural representations and summarize several relevant (dis)similarity measures. In our review of past work, we classify representational (dis)similarity scores into two main categories: those that explicitly align neural dimensions, and those that quantify similarity in stimulus-by-stimulus relationships. In section 3, we summarize our main theoretical result linking shape distances to NBS, explicitly connecting the two categories of (dis)similarity measures. In section 4, we explore the behavior of these distances when the number of stimuli (M) or the number of dimensions (N) goes to infinity. We show that the duality between NBS and shape distance can be leveraged to understand these asymptotic regimes. Finally, in section 5 we discuss how NBS and shape distances compare to CKA, a popular approach which does not enjoy the theoretical properties and interpretations discussed above. We show numerically and analytically that the relationship between these quantities is rather loose, and we therefore do not expect them to be interchangeable in practical applications.

2 Review of Representational (Dis)similarity Measures

Let $f_x : \mathcal{Z} \mapsto \mathsf{R}^{N_x}$ and $f_y : \mathcal{Z} \mapsto \mathsf{R}^{N_y}$ be two neural networks that map inputs over some domain $\mathcal{Z}$ to neural activation vectors (e.g. the vector of activations produced at a hidden layer of a deep network). Here, N_x and N_y respectively denote the number of neurons in each network. This mapping from inputs to neural activations is typically considered to be deterministic (but see [11] for the stochastic setting).

How similar is neural network $f_x(\cdot)$ to neural network $f_y(\cdot)$? That is, how similar are the functions $f_x(\cdot)$ and $f_y(\cdot)$ over a representative collection of inputs $z_1, \ldots, z_M \in \mathcal{Z}$? We proceed by measuring neural responses $f_x(z_1) \ldots f_x(z_M)$ and stacking them row-wise into a matrix $\boldsymbol{X} \in \mathsf{R}^{M \quad N_x}$. Likewise, we form a matrix $\boldsymbol{Y} \in \mathsf{R}^{M \quad N_y}$ from the second network's responses, $f_y(z_1) \ldots f_y(z_M)$. Intuitively, one can view these matrices as approximations to each network's input-output mapping over a discrete set of M points.

In general, $N_x \neq N_y$, but even if $N_x = N_y$, there is no reason we expect the raw Euclidean distance, $\|\boldsymbol{X} - \boldsymbol{Y}\|_F$ to be meaningful since the neurons (columns of $\boldsymbol{X}$ and $\boldsymbol{Y}$) are often labelled in arbitrary order. Thus, we are interested in measuring a distance *that is invariant to a specified set of nuisance transformations* in the representations. For example, if we would like to ignore orthogonal transformations (including permutations of the neuron indices), we ought to develop distance functions for which $d(\boldsymbol{X}, \boldsymbol{Y}) = d(\boldsymbol{X}, \boldsymbol{Y}\boldsymbol{Q})$ and also $d(\boldsymbol{X}, \boldsymbol{X}\boldsymbol{Q}) = 0$ for any orthogonal matrix $\boldsymbol{Q}$. This can be formalized by defining an equivalence relation between neural responses and defining a metric over the corresponding equivalence classes [58].

Existing representational (dis)similarity measures between $\boldsymbol{X}$ and $\boldsymbol{Y}$ roughly fall into two broad camps: methods that learn explicit mappings to align neural dimensions, and methods that utilize stimulus-by-stimulus similarity matrices to compare across networks (fig. 1). The main purpose of our paper is to provide a bridge between these approaches, and so we summarize a few primary examples below. A comprehensive review of these methods is far beyond the scope of this paper, but we point the reader to Klabunde et al. [22] and Sucholutsky et al. [52] as useful papers to cross-reference. Indeed, Klabunde et al. [22] enumerate over 30 methods to quantify representational (dis)similarity, showing the strong need for organizing theoretical principles to relate and unify these approaches.

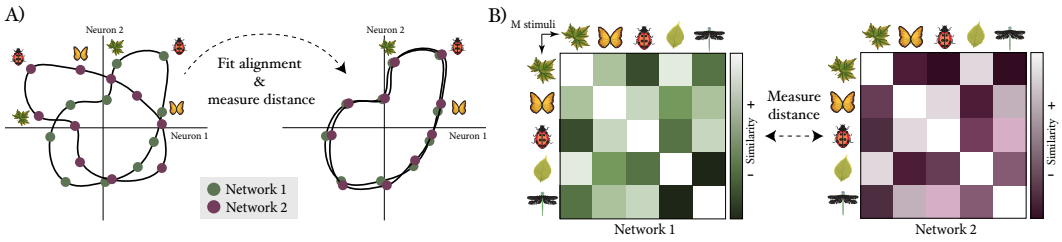

Figure 1: We consider methods of measuring representational similarity from two broad categories: (A) alignment-based measures, which fit a mapping that aligns neural dimensions, and (B) methods that compare stimulus-by-stimulus representational similarity matrices.

2.1 (Dis)similarity measures that transform or align neural dimensions

Here we review the first major category of representational (dis)similarity measures. Recall that the main challenge is that raw distances, such as $\|\boldsymbol{X} - \boldsymbol{Y}\|_F$, are typically meaningless due to nuisance transformations. To overcome this, one option is to optimize alignment functions $g_x : \mathsf{R}^{N_x} \mapsto \mathcal{S}$ and $g_y : \mathsf{R}^{N_y} \mapsto \mathcal{S}$ which respectively map the rows of $\boldsymbol{X}$ and $\boldsymbol{Y}$ into a common space $\mathcal{S}$. Then, given a dissimilarity measure $d : \mathcal{S} \times \mathcal{S} \mapsto \mathsf{R}$ we can optimize the alignment:

$$\underset{g_x, g_y}{\text{minimize}} \;\; d(g_x(\boldsymbol{X}), g_y(\boldsymbol{Y})) \quad \text{subject to} \;\; g_x \in \mathcal{G}_x \;,\; g_y \in \mathcal{G}_y \tag{1}$$

where $\mathcal{G}_x$ and $\mathcal{G}_y$ represent the class of permitted alignment functions for $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively. Note that we are using $g_x(\boldsymbol{X})$ and $g_y(\boldsymbol{Y})$ to denote the row-wise application of $g_x(\cdot)$ and $g_y(\cdot)$ to matrix-valued inputs.

Intuitively, the minimal value attained in eq. (1) will be invariant to nuisance transformations that are contained within $\mathcal{G}_x$ or $\mathcal{G}_y$. For example, **linear regression** is a popular method to quantify similarity between artificial and biological neural networks [48], and can be viewed as a special case of eq. (1). Specifically, to predict $\boldsymbol{Y}$ from $\boldsymbol{X}$, we would choose $\mathcal{S} = \mathsf{R}^{N_y}$, constrain $g_y(\cdot)$ to be the identity mapping, and minimize $g_x(\cdot)$ over the set of affine mappings from $\mathsf{R}^{N_x} \mapsto \mathsf{R}^{N_y}$. Importantly, linear regression does not produce a symmetric notion of (dis)similarity: predicting $\boldsymbol{Y}$ from $\boldsymbol{X}$ will produce a different result than predicting $\boldsymbol{X}$ from $\boldsymbol{Y}$. Variants such as **canonical correlations analysis (CCA)** provide symmetric, affine-invariant measures of representational similarity, and have been used in deep learning and in neuroscience [53, 44, 14]. To see that CCA is a special case of eq. (1), we choose $\mathcal{S} = \mathsf{R}^n$ where $n = \min(N_x, N_y)$ and minimize $-1 \cdot \mathrm{Tr}[g_x(\boldsymbol{X})^\top g_y(\boldsymbol{Y})]$ subject to $g_x(\boldsymbol{X})^\top g_x(\boldsymbol{X}) = g_y(\boldsymbol{Y})^\top g_y(\boldsymbol{Y}) = \boldsymbol{I}$. Intuitively, CCA finds mappings of $\boldsymbol{X}$ and $\boldsymbol{Y}$ into a common space R^n which maximize correlations along orthogonal dimensions. Other methods like **permutation matching** of individual neurons [29], can also be viewed as special cases of eq. (1).

In summary, the examples above show that eq. (1) captures a broad variety of existing approaches. We now review another example, **shape distances** [20], which feature prominently into our main narrative. Kendall [19] defined shape as the structure left by a set of M *landmark* points in R^N after rotations, translations, and isotropic scalings are ignored. These M landmark points in our context become the collection of M inputs to each network over which the representational similarity will be measured. Unlike traditional shape theory, we will additionally consider reflections as in R^N as nuisance transformations, since a permutation of neuron labels (which we argued above is typically arbitrary) can require a reflection.

Assuming that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are $M \times N$ matrices, we can define their angular distance:

$$\theta(\boldsymbol{X}, \boldsymbol{Y}) = \cos^{-1}\left(\frac{\mathrm{Tr}[\boldsymbol{X}^\top \boldsymbol{Y}]}{\sqrt{\mathrm{Tr}[\boldsymbol{X}^\top \boldsymbol{X}]\,\mathrm{Tr}[\boldsymbol{Y}^\top \boldsymbol{Y}]}}\right). \tag{2}$$

which generalizes the elementary formula for the angle between two vectors. One can then define the **Riemannian shape distance** as the length of the shortest geodesic path between two shapes, and show that this is given by [20]:

$$\theta\;(\boldsymbol{X}, \boldsymbol{Y}) = \min_{\boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}} \theta(\boldsymbol{C}\boldsymbol{X}, \boldsymbol{C}\boldsymbol{Y}\boldsymbol{Q}), \tag{3}$$

where $\boldsymbol{C} = \boldsymbol{I} - \frac{1}{M}\mathbf{1}\mathbf{1}^\top$ is the *centering matrix* ($\mathbf{1}\mathbf{1}^\top$ is an $M \times M$ matrix of all ones). One may check that the columns of $\boldsymbol{CX}$ and $\boldsymbol{CY}$ are mean-centered, and that $\boldsymbol{C}$ is a symmetric, idempotent matrix with $\boldsymbol{C}^\top\boldsymbol{C} = \boldsymbol{C}^2 = \boldsymbol{C}$.

Closely related to the Riemannian shape distance is a quantity called the **Procrustes size-and-shape distance** [20] (which, for brevity, we will simply call the Procrustes distance):

$$\mathcal{P}(\boldsymbol{X}, \boldsymbol{Y}) = \min_{\boldsymbol{Q}^\top\boldsymbol{Q}=\boldsymbol{I}} \|\boldsymbol{CX} - \boldsymbol{CYQ}\|_F \tag{4}$$

After expanding eq. (4) and rearranging, we can see that the cosine of the Riemannian shape distance is related to the Procrustes distance between two matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ normalized by their centered Frobenius norms:

$$\cos\theta\ (\boldsymbol{X}, \boldsymbol{Y}) = 1 - \frac{1}{2}\mathcal{P}^2\left(\frac{\boldsymbol{X}}{\|\boldsymbol{CX}\|_F}, \frac{\boldsymbol{Y}}{\|\boldsymbol{CY}\|_F}\right). \tag{5}$$

It is straightforward to check that both Riemannian shape distance and Procrustes distance are invariant to translations and orthogonal transformations. The Riemannian shape distance is additionally invariant to isotropic scalings. Again, both of these distances are special cases of eq. (1).

We note that both shape distances are symmetric and satisfy the triangle inequality. That is, for any triplet of configuration matrices $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{M} \in \mathsf{R}^{M\ \ N}$ we have $\theta\ (\boldsymbol{X}, \boldsymbol{Y}) = \theta\ (\boldsymbol{Y}, \boldsymbol{X})$ and $\theta\ (\boldsymbol{X}, \boldsymbol{Y}) \leq \theta\ (\boldsymbol{X}, \boldsymbol{M}) + \theta\ (\boldsymbol{M}, \boldsymbol{Y})$, and likewise for $\mathcal{P}$. Formally, this means that θ and $\mathcal{P}$ define metric spaces over the equivalence classes defined by their nuisance transformations. Williams et al. [58] argued that these properties were advantageous for downstream analyses such as nearest-neighbor regression and clustering methods that leverage metric space structure.

Comparing shapes of unequal dimension. Importantly, the definitions of shape distance in eqs. (3) and (4) assume that we are comparing networks of equal size—i.e., that $N_x = N_y = N$. This corresponds to the setting of traditional shape theory, but is not an ideal assumption for our application since we often desire to compare representations across networks with different sizes or different numbers of experimentally measured neurons. Williams et al. [58] proposed procedures to either use PCA or zero padding to embed all networks into a common dimension. We will show these procedures are unnecessary since we can reformulate the shape distances in terms of:

$$\boldsymbol{\Sigma}_X = \boldsymbol{X}^\top\boldsymbol{CX} \quad , \quad \boldsymbol{\Sigma}_Y = \boldsymbol{Y}^\top\boldsymbol{CY} \quad \text{and} \quad \boldsymbol{\Sigma}_{XY} = \boldsymbol{X}^\top\boldsymbol{CY} \tag{6}$$

which are the covariance and cross-covariance matrices describing similarity between pairs of neural tuning functions within and across networks. Specifically, let us define:

$$\theta\ (\boldsymbol{X}, \boldsymbol{Y}) = \cos^{\ 1}\left[\frac{\|\boldsymbol{\Sigma}_{XY}\|}{\sqrt{\mathrm{Tr}[\boldsymbol{\Sigma}_X]\,\mathrm{Tr}[\boldsymbol{\Sigma}_Y]}}\right] \tag{7}$$

$$\mathcal{P}(\boldsymbol{X}, \boldsymbol{Y}) = \sqrt{\mathrm{Tr}[\boldsymbol{\Sigma}_X] + \mathrm{Tr}[\boldsymbol{\Sigma}_Y] - 2\|\boldsymbol{\Sigma}_{XY}\|} \tag{8}$$

where $\|\cdot\|$ denotes the nuclear matrix norm (also called the Schatten 1-norm), which is given by the sum of a matrix's singular values. Our main claim is the following:

Lemma 1. *When $N_x = N_y = N$, the definitions of θ and $\mathcal{P}$ given in eqs.* (3) *and* (4) *are equivalent to the definitions of θ and $\mathcal{P}$ given in eqs.* (7) *and* (8)*.*

which follows immediately from a well-known result of Schönemann [47] (Appendix 7.1).

Lemma 1 essentially shows that eqs. (7) and (8) are reasonable generalizations of shape distance that are well-defined when $N_x \neq N_y$. Although it may not be immediately obvious, we will see (due to theorem 1 in section 3) that the new definitions of shape distance in eqs. (7) and (8) continue to satisfy the triangle inequality, even when comparing networks with different sizes. The geometric intuition underlying shape distances—that of fitting a rotational alignment between two neural activation spaces—also carries over. Specifically, let us assume $N_x \leq N_y$ (without loss of generality). Then, we can isometrically embed the lower-dimensional representations into R^{N_y} by, for example, appending $N_y - N_x$ columns of zeros to $\boldsymbol{X}$. Then, we compute the shape distance as before, which involves finding the optimal orthogonal transformation in N_y dimensions to match the landmark points.

2.2 (Dis)similarity measures that quantify stimulus-by-stimulus relationships

Recall that the (dis)similarity measures summarized above relied on learning explicit alignment transformations on the neural activation space, R^N for a population of N neurons. We now turn to review the second, alternative category of measures, which avoid the need to fit any alignment. Instead, these methods quantify representational similarity by comparing summary statistics that are already invariant to nuisance transformations.

For instance, given N-dimensional network responses to M sampled inputs, we can compute a **representational dissimilarity matrix (RDM)** [25]: an $M \times M$ matrix of Euclidean distances between all pairs of evoked responses within the N-dimensional response space. Intuitively, RDMs encode a rich geometric summary of the network's representation that is invariant to rotations and translations of the neural activation space. In fact, since the Procrustes distance is invariant to translations, it is easy to show that $\boldsymbol{X}$ and $\boldsymbol{Y}$ have the same RDM if and only if the size-and-shape Procrustes distance between $\boldsymbol{X}$ and $\boldsymbol{Y}$ is zero. This already hints at deeper connections, which we reveal in section 3.

Quantifying similarity of RDMs across networks is common in cognitive/systems neuroscience, where the approach is broadly referred to as **representational similarity analysis (RSA)** [25]. Many variants of RSA use different approaches to compute the within-network RDMs (e.g. Mahalanobis vs. Euclidean distance) or different measures to compare two RDMs (e.g. Pearson or Spearman correlation scores) [41, 56, 49]. But all of these variants conceptually share the same core approach.

An alternative to computing RDMs is to compute $M \times M$ *kernel matrices*, which use a positive definite kernel function [46] to compute a similarity score between all pairs of evoked network responses. Specifically, we will focus on *centered linear kernel matrices* (with centering matrix $\boldsymbol{C}$ defined as in the previous section), as they are the most popular in practice:

$$\boldsymbol{K}_X = \boldsymbol{C}\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{C} \quad \text{and} \quad \boldsymbol{K}_Y = \boldsymbol{C}\boldsymbol{Y}\boldsymbol{Y}^T\boldsymbol{C}\,. \tag{9}$$

Note that $\boldsymbol{K}_X$ and $\boldsymbol{K}_Y$ are simply covariance matrices across stimuli; they are natural counterparts to the covariance matrices across neurons, $\boldsymbol{\Sigma}_X$ and $\boldsymbol{\Sigma}_Y$ defined in eq. (6). Intuitively, $\boldsymbol{K}_X$ and $\boldsymbol{K}_Y$ hold all the information necessary to compute an RDM since the squared Euclidean distance between stimulus i and j is given by: $[\boldsymbol{K}_X]_{ii} + [\boldsymbol{K}_X]_{jj} - [\boldsymbol{K}_X]^2_{ij}$. Like RDMs, the kernel matrices are invariant to rotations, reflections, and translations. An influential paper by Kornblith et al. [23] proposed **centered kernel alignment (CKA)** [7, 6] as a measure of similarity between kernel matrices:

$$\mathrm{CKA}(\boldsymbol{K}_X, \boldsymbol{K}_Y) = \frac{\mathrm{Tr}[\boldsymbol{K}_X\boldsymbol{K}_Y]}{\sqrt{\mathrm{Tr}[\boldsymbol{K}_X^2]\,\mathrm{Tr}[\boldsymbol{K}_Y^2]}} \tag{10}$$

which is the cosine of the angle between $\boldsymbol{K}_X$ and $\boldsymbol{K}_Y$ (see eq. 2). Shahbazi et al. [50] pointed out that CKA does not exploit the fact that $\boldsymbol{K}_X$ and $\boldsymbol{K}_Y$ are positive semidefinite (PSD) matrices, and they propose an alternative metric based on the **Riemannian distance on PSD matrices**. Yet another measure on the linear kernel matrices is the **normalized Bures similarity (NBS)**, defined as [37, 54]:

$$\mathrm{NBS}(\boldsymbol{K}_X, \boldsymbol{K}_Y) = \frac{\mathcal{F}(\boldsymbol{K}_X, \boldsymbol{K}_Y)}{\sqrt{\mathrm{Tr}[\boldsymbol{K}_X]\,\mathrm{Tr}[\boldsymbol{K}_Y]}} \tag{11}$$

with

$$\mathcal{F}(\boldsymbol{K}_X, \boldsymbol{K}_Y) = \mathrm{Tr}[(\boldsymbol{K}_X^{1/2}\boldsymbol{K}_Y\boldsymbol{K}_X^{1/2})^{1/2}]. \tag{12}$$

The quantity $\mathcal{F}(\boldsymbol{K}_X, \boldsymbol{K}_Y)$ is known as the *fidelity* and is used extensively in quantum information theory as a measure of the distinguishability of quantum states [35, 57]. We will also make use of a related quantity known as the **Bures distance** on PSD matrices [4]:

$$\mathcal{B}(\boldsymbol{K}_X, \boldsymbol{K}_Y) = \sqrt{\mathrm{Tr}[\boldsymbol{K}_X] + \mathrm{Tr}[\boldsymbol{K}_Y] - 2\,\mathrm{Tr}\left[\left(\boldsymbol{K}_X^{1/2}\boldsymbol{K}_Y\boldsymbol{K}_X^{1/2}\right)^{1/2}\right]}. \tag{13}$$

It is well-known that the Bures distance is equal to the 2-Wasserstein distance between two mean-centered normal distributions with covariances given by $\boldsymbol{K}_X$ and $\boldsymbol{K}_Y$ ([42], Remark 2.30). Thus, one can interpret $\mathcal{B}(\boldsymbol{K}_X, \boldsymbol{K}_Y)$ as the cost of optimally transporting mass between two normal densities. This connection could allow one to exploit the large collection of existing knowledge of optimal transportation, as in [34], although this is beyond the scope of the present work.

3 Duality of Shape and Bures Distances

In section 2, we saw that many measures of representational (dis)similarity either identify an explicit mapping between neural dimensions or directly compare stimulus-by-stimulus (dis)similarities. Each perspective has its own conceptual appeal. The former encourages us to reason about geometric features in the space of neural activations, such as curvature or tangling of manifold structure which feature prominently in theories of neural computation [45, 16, 15]. The latter avoids the need to align neural axes, and connects to a rich literature in psychology that leverages pairwise similarity judgements to interrogate the structure of cognition [3, 12]. Our sense is that many researchers at the intersection of neuroscience, cognitive science, and interpretable deep learning tend to develop a personal preference for one perspective over the other. But are these approaches really so distinct?

We now turn to one of our main results, which highlights a specific case where these two perspectives produce the same quantitative result—an example of a *duality* [2]. Specifically, the theorem below states that the Procrustes distance ($\mathcal{P}$, eqs. 4 and 8) is equal to the Bures distance between linear kernel matrices ($\mathcal{B}$, eq. 13). Furthermore, the normalized Bures similarity (NBS, eq. 11) is equal to the cosine of the Riemannian shape distance (θ , eqs. 3 and 7).

Theorem 1. *Let $\boldsymbol{K}_X$ and $\boldsymbol{K}_Y$ be centered linear kernel matrices as in eq.* (9)*. Then,*

$$\mathcal{B}(\boldsymbol{K}_X, \boldsymbol{K}_Y) = \mathcal{P}(\boldsymbol{X}, \boldsymbol{Y}) \tag{14}$$

and furthermore,

$$\textit{NBS}(\boldsymbol{K}_X, \boldsymbol{K}_Y) = \cos\theta \ \ (\boldsymbol{X}, \boldsymbol{Y}) \ . \tag{15}$$

The proof follows from the fact that for centered linear kernel matrices, $\mathrm{Tr}[\boldsymbol{K}_X] = \mathrm{Tr}[\boldsymbol{\Sigma}_X]$ and $\mathrm{Tr}[\boldsymbol{K}_Y] = \mathrm{Tr}[\boldsymbol{\Sigma}_Y]$, and the following lemma:

Lemma 2. *For centered linear kernel matrices, $\mathcal{F}(\boldsymbol{K}_X, \boldsymbol{K}_Y) = \|\boldsymbol{\Sigma}_{XY}\|$.*

It is easy to see that the lemma implies the theorem. E.g., from the definitions in eqs. (8) and (13):

$$\mathcal{B}^2(\boldsymbol{K}_X, \boldsymbol{K}_Y) = \mathrm{Tr}[\boldsymbol{K}_X] + \mathrm{Tr}[\boldsymbol{K}_Y] - 2\mathcal{F}(\boldsymbol{K}_X, \boldsymbol{K}_Y)$$
$$\mathcal{P}^2(\boldsymbol{X}, \boldsymbol{Y}) = \mathrm{Tr}[\boldsymbol{\Sigma}_X] + \mathrm{Tr}[\boldsymbol{\Sigma}_Y] - 2\|\boldsymbol{\Sigma}_{XY}\|$$

So eq. (14) follows from observing that the three terms on the right hand sides above are equal due to lemma 2. A similar argument shows that eq. (15) follows from lemma 2 as well. Thus, all that remains is to prove lemma 2, which we do in Appendix 7.2.

The proof of theorem 1 is straightforward, and somewhat similar results have already appeared in mathematical literature [34]. However, this result may have gone unnoticed by researchers at the intersection of machine learning and neuroscience because prior similar statements have appeared in a technical literature focused on distinct problems.

Theorem 1 enables us to draw upon an extensive literature to theoretically characterize shape/Bures distances. For example, it is well known that $\mathcal{B}$ and $\cos^{-1}$ NBS both satisfy the criteria of a metric space, including the triangle inequality [40]. Thus, we can immediately conclude that the generalized definitions of Procrustes and Riemannian shape distance in eqs. (7) and (8) are also metrics, even though most classical work on shape theory does not typically consider datasets with unequal dimensions ($N_x \neq N_y$).

4 Asymptotic Analysis of Shape/Bures Distances

In the previous section, we provide a concrete link between two previously disconnected perspectives of representaitonal (dis)similarity. Beyond conceptual appeal, does this advance unlock any practical benefits for future research? To demonstrate the utility of our result, we investigate how shape/Bures distance changes as more inputs are sampled ($M \to \infty$) and as the size of the neural population increases ($N \to \infty$). As shown below, both of these regimes are of interest to researchers in neuroscience and deep learning, and the duality established in theorem 1 enables immediate insights.

Before proceeding, we must introduce normalization factors into our definitions of Procrustes and Bures distances, since $\mathcal{P}(\boldsymbol{X}, \boldsymbol{Y})$ and $\mathcal{B}(\boldsymbol{K}_X, \boldsymbol{K}_Y)$ will diverge as $N, M \to \infty$. Thus, we define the normalized Procrustes distance, ρ, and normalized Bures distance, b, as:

$$\rho(\boldsymbol{X}, \boldsymbol{Y}) = \cancel{\rho}\frac{1}{\sqrt{NM}}\mathcal{P}(\boldsymbol{X}, \boldsymbol{Y}) \quad \text{and} \quad b(\boldsymbol{K}_X, \boldsymbol{K}_Y) = \cancel{\rho}\frac{1}{\sqrt{NM}}\mathcal{B}(\boldsymbol{K}_X, \boldsymbol{K}_Y) \tag{16}$$

Note that we have assumed that $N_x = N_y = N$ (i.e. the networks have the same number of neurons) since we are interested in taking $N \to \infty$ anyways. The motivation behind the normalizing factor of $1/\sqrt{MN}$ will be made clear by the discussion below.

Interpretation as $M \to \infty$. Thus far, we have considered network representations of M input stimuli. In most cases, these inputs are viewed as from a broader input distribution. For example, to compare visual representations, we input M random natural images into a pair of networks and measure their similarity. As $M \to \infty$ we would desire that $\rho(\boldsymbol{X}, \boldsymbol{Y})$ and $b(\boldsymbol{K}_X, \boldsymbol{K}_Y)$ converge to a constant value that reflects the underlying structure of the input distribution. But in this limit $\boldsymbol{K}_X$ and $\boldsymbol{K}_Y$ become, loosely speaking, infinite-dimensional matrices. Thus it may not be immediately obvious how to interpret $b(\boldsymbol{K}_X, \boldsymbol{K}_Y)$ without tools from functional analysis.

On the other hand, due to the law of large numbers, we have in this limit that:

$$\tfrac{1}{M}\boldsymbol{\Sigma}_X \to \boldsymbol{\Sigma}_X \qquad \tfrac{1}{M}\boldsymbol{\Sigma}_Y \to \boldsymbol{\Sigma}_Y \qquad \tfrac{1}{M}\boldsymbol{\Sigma}_{XY} \to \boldsymbol{\Sigma}_{XY} \tag{17}$$

where $\boldsymbol{\Sigma}_X$, $\boldsymbol{\Sigma}_Y$, and $\boldsymbol{\Sigma}_{XY}$ are the true covariances and cross-covariances across neurons. Thus,

$$\lim_{M \to \infty} \rho(\boldsymbol{X}, \boldsymbol{Y}) = \sqrt{\tfrac{1}{N}} \cdot \sqrt{\mathrm{Tr}[\boldsymbol{\Sigma}_X] + \mathrm{Tr}[\boldsymbol{\Sigma}_Y] - 2\|\boldsymbol{\Sigma}_{XY}\|} \tag{18}$$

which is in some sense the "true" Procrustes distance we are aiming to approximate, pre-multiplied by a factor of $1/\sqrt{N}$. Theorem 1 allows us to immediately conclude that $b(\boldsymbol{K}_X, \boldsymbol{K}_Y)$ converges to the same value as $\rho(\boldsymbol{X}, \boldsymbol{Y})$ in this limit.

Interpretation as $N \to \infty$. Experimental neuroscientists often record a small random sample of N neurons (e.g. 100-1000 cells) from brain regions that are much larger, often by several orders of magnitude. How large do we need N to be in order to achieve a good estimate of the "true" representational (dis)similarity [24]? In analogy to our logic above, the "true" (dis)similarity is given by considering the limit that $N \to \infty$. This limiting regime is also of interest to the theory of deep learning, in which one often deals with networks with "infinitely wide" layers [13, 17, 28].

In the limit that $N \to \infty$, shape distances become hard to conceptualize without leveraging advanced mathematics. Conceptually, one can imagine fitting an infinite-dimensional rotation matrix that aligns the neural axes, or else calculating eq. (8) on infinite-dimensional covariances. On the other hand, in analogy to eq. (17), we have:

$$\tfrac{1}{N}\boldsymbol{K}_X \to \boldsymbol{K}_X \qquad \tfrac{1}{N}\boldsymbol{K}_Y \to \boldsymbol{K}_Y \tag{19}$$

in the limit that $N \to \infty$ due to the law of large numbers. Intuitively, $\boldsymbol{K}_X$ and $\boldsymbol{K}_Y$ are the "true" covariances describing the expected similarity between pairs of stimuli across the fully observed neural population. Further, it is easy to verify that the normalized Bures distance converges to the appropriate value, up to a factor of $1/\sqrt{M}$. That is,

$$\lim_{N \to \infty} b(\boldsymbol{X}, \boldsymbol{Y}) = \sqrt{\tfrac{1}{M}} \cdot \sqrt{\mathrm{Tr}[\boldsymbol{K}_X] + \mathrm{Tr}[\boldsymbol{K}_Y] - 2\mathcal{F}(\boldsymbol{K}_X, \boldsymbol{K}_Y)} \tag{20}$$

Theorem 1 allows us to immediately conclude that $\rho(\boldsymbol{X}, \boldsymbol{Y})$ also converges to this value in this limit.

Interpretation as $M \to \infty$ and $N \to \infty$. In summary, theorem 1 makes it easy to see that normalized shape and Bures distances converge to reasonable values when either $M \to \infty$ or $N \to \infty$. Deriving this insight only required us to leverage a basic law of large numbers—namely, that empirical covariance matrices converge to the true covariance in the limit of infinite samples.

An obvious question is whether the shape/Bures distances converge to reasonable quantities when *both* M and N are taken to infinity. A rigorous analysis of this scenario requires a more delicate approach that leverages concepts from functional analysis. Nonetheless, it can be shown that as $M, N \to \infty$, the Bures distance between covariance matrices converges to the Bures distance between an associated positive semidefinite covariance operator [43, 33].

5 Theoretical and Numerical Comparisons with CKA

We have seen that shape and Bures distances enjoy a special duality. Namely, they can be interpreted as the distance found after optimally aligning the neural activation spaces (see eqs. 3 and 4) or a direct distance between stimulus-by-stimulus kernel matrices. This special property does not appear to be

shared by other representational (dis)similarity measures, which are typically compatible with only one of these perspectives. However, the shape and Bures distances only represent a very small fraction of a much larger landscape of (dis)similarity measures, which we surveyed briefly in section 2. Do the shape and Bures distances meaningfully differ from these alternative approaches?

In this section, we investigate the relationship between NBS and CKA [6, 7, 23], a particularly popular approach in the deep learning literature. By comparing their definitions in eqs. (10) and (11), one may guess that CKA is closely related to NBS (and therefore also to Riemannian shape distance by theorem 1). However, we will show that CKA scores between networks can differ substantially (e.g. two- to three-fold) from NBS scores. We also derive upper and lower bounds that relate CKA and NBS; an exercise which confirms their rather loose relationship. Overall, we conclude that one should not expect CKA and NBS to behave similarly in practical scenarios.

5.1 Relationship between CKA and Euclidean geometry

We begin our comparison by noting a relationship between CKA and Euclidean distance:

$$\mathrm{CKA}(\boldsymbol{K}_X, \boldsymbol{K}_Y) = 1 - \frac{1}{2}\left\| \frac{\boldsymbol{K}_X}{\|\boldsymbol{K}_X\|_F} - \frac{\boldsymbol{K}_Y}{\|\boldsymbol{K}_Y\|_F} \right\|_F^2. \tag{21}$$

Likewise, by noting that $\mathrm{NBS}(\boldsymbol{K}_X, \boldsymbol{K}_Y) = \mathcal{F}(\boldsymbol{K}_X / \operatorname{Tr} \boldsymbol{K}_X, \boldsymbol{K}_Y / \operatorname{Tr} \boldsymbol{K}_Y)$ and rearranging the definition of Bures distance, we observe that NBS has an analogous relationship to Bures distance:

$$\mathrm{NBS}(\boldsymbol{K}_X, \boldsymbol{K}_Y) = 1 - \frac{1}{2}\mathcal{B}^2\Big(\frac{\boldsymbol{K}_X}{\operatorname{Tr} \boldsymbol{K}_X}, \frac{\boldsymbol{K}_Y}{\operatorname{Tr} \boldsymbol{K}_Y}\Big) \tag{22}$$

which can be compared with eq. (5). Thus, the central conceptual difference between CKA and NBS is the choice of metric on the space of PSD matrices. Intuitively, measuring similarity with CKA (instead of NBS) is akin to measuring distance using a Euclidean (instead of Bures) geometry on PSD matrices.

Many previous works have argued that using a Euclidean geometry to compare or estimate PSD matrices is suboptimal for certain analyses [1, 10], including recent work by Shahbazi et al. [50] in the context of comparing neural representations. To gain some intuition, consider the problem of interpolating between two kernel matrices. In a Euclidean geometry, one obtains $\alpha \boldsymbol{K}_X + (1-\alpha)\boldsymbol{K}_Y$ which is PSD if $0 \leq \alpha \leq 1$. However, if one extrapolates beyond these bounds (e.g. by choosing $\alpha > 1$), the resulting matrix may contain negative eigenvalues. Such extrapolations are used when CKA is optimized by gradient descent, as done in several prior works [9, 8].

The problems mentioned above do not arise if one extrapolates along geodesics defined by the Bures distance. Indeed, the Bures distance between a PSD matrix and a matrix with negative eigenvalues is not well-defined. However, it is not the main point of this paper to contend that Bures geometry is inherently superior. One can show that the topologies induced by the Euclidean and Bures distances coincide (see Lemma 3.2 in [55]). Determining whether the Euclidean geometry is "good enough" for a particular application should be considered carefully on a case-by-case basis.

5.2 Upper and lower bounds on NBS in terms of CKA

Having discussed a central conceptual difference between NBS and CKA, we turn our attention to a more practical question: How big are the potential discrepancies between CKA and NBS? Figure 2 shows NBS plotted against CKA for randomly sampled pairs of PSD matrices. This figure suggests that, while there is not a one-to-one relationship between the two quantities, the two similarity measures constrain each other to an allowed envelope. We can derive bounds on this envelope by expressing squared NBS and CKA in terms of matrix norms:

$$\mathrm{NBS}(\boldsymbol{K}_X, \boldsymbol{K}_Y)^2 = \frac{\|\boldsymbol{\Sigma}_{XY}\|^2}{\|\boldsymbol{K}_X\|\ \|\boldsymbol{K}_Y\|}, \quad \mathrm{CKA}(\boldsymbol{K}_X, \boldsymbol{K}_Y) = \frac{\|\boldsymbol{\Sigma}_{XY}\|_F^2}{\|\boldsymbol{K}_X\|_F \|\boldsymbol{K}_Y\|_F} \tag{23}$$

(Note that we have exploited theorem 1 to reformulate the numerator of NBS.) The elementary matrix norm inequalities $\|\boldsymbol{A}\|_F \leq \|\boldsymbol{A}\| \leq \sqrt{\mathrm{rank}(\boldsymbol{A})}\|\boldsymbol{A}\|_F$ and the subsequent observation $\operatorname{Tr}[\boldsymbol{K}_X \boldsymbol{K}_Y] \leq \mathcal{F}(\boldsymbol{K}_X, \boldsymbol{K}_Y)^2$ lead us to the following envelope (setting $r(\boldsymbol{A}) = \mathrm{rank}(\boldsymbol{A})$):

$$\frac{\mathrm{CKA}(\boldsymbol{K}_X, \boldsymbol{K}_Y)}{\sqrt{r(\boldsymbol{K}_X) r(\boldsymbol{K}_Y)}} \leq \mathrm{NBS}(\boldsymbol{K}_X, \boldsymbol{K}_Y)^2 \leq \min[r(\boldsymbol{K}_X), r(\boldsymbol{K}_Y)]\, \mathrm{CKA}(\boldsymbol{K}_X, \boldsymbol{K}_Y) \tag{24}$$

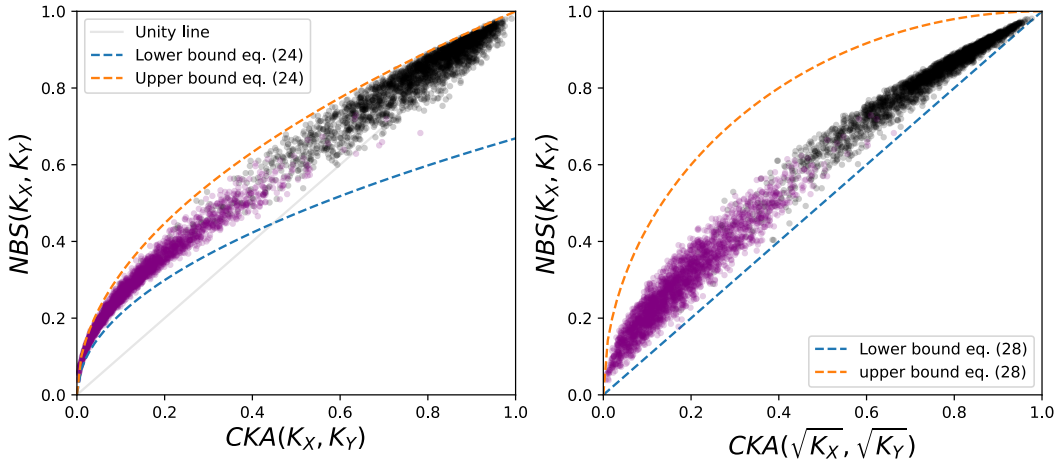

Figure 2: Comparing CKA and NBS. Purple points represent the similarity between pairs of matrices generated by sampling two Wishart distributions, $\sqrt{\boldsymbol{K}_X} \sim W_{10}(\boldsymbol{I}, 1)$ and $\sqrt{\boldsymbol{K}_Y} \sim W_{10}(\boldsymbol{I}, 5)$. Black points are generated by sampling $\sqrt{\boldsymbol{K}_X} \sim W_{10}(\boldsymbol{I}, 1)$ and setting $\sqrt{\boldsymbol{K}_Y} = \sqrt{\boldsymbol{K}_X} + \epsilon$, where $\epsilon \sim W_{10}(\boldsymbol{I}, 4)$. (a) CKA bounds NBS within an envelope determined by the matrix ranks; see eq. (24). (b) CKA($\sqrt{\boldsymbol{K}_X}, \sqrt{\boldsymbol{K}_Y}$) bounds NBS($\boldsymbol{K}_X, \boldsymbol{K}_Y$) with inequalities given by eq. (28).

Both of these bounds are saturated when $\text{rank}(\boldsymbol{K}_X) = \text{rank}(\boldsymbol{K}_Y) = 1$. Figure 2 (a) demonstrates that while NBS is bound to an envelope by CKA set by the matrix ranks, the allowed discrepency between these two can still be large compared with the total range of $[0, 1]$.

5.3 Connecting NBS and CKA through Uhlmann's theorem

The equality of the Procrustes and Bures distances can be further understood by noticing that for any particular fixed PSD $\boldsymbol{K}_X$, there are infinitely many matrices $\boldsymbol{X}$ for which $\boldsymbol{K}_X = \boldsymbol{X}\boldsymbol{X}^\top$. This set of matrices are related by orthogonal transformations– for two $M \times N_x$ matrices $\boldsymbol{X}$ and $\boldsymbol{X}^0$ that satisfy $\boldsymbol{X}\boldsymbol{X}^\top = \boldsymbol{K}_X = \boldsymbol{X}^0\boldsymbol{X}^{0\top}$, we have $\boldsymbol{X}\boldsymbol{U} = \boldsymbol{X}^0$ with $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{I}$. This set includes rectangular matrices of dimension $M \times N_x$, where $N_x \geq \text{rank}(\boldsymbol{K}_x)$. The unique PSD square root $\sqrt{\boldsymbol{K}_X}$ represents a particular square member of this set.

We will now assume $N_x = N_y = N$, so that we can compute the Hilbert-Schmidt inner product between $\boldsymbol{X}$ and $\boldsymbol{Y}$ to measure their overlap. Intuitively, a meaningful measure of similarity between $\boldsymbol{K}_X$ and $\boldsymbol{K}_Y$ could be the maximum inner product over all $\boldsymbol{X}$ and $\boldsymbol{Y}$ that are consistent with $\boldsymbol{K}_X$ and $\boldsymbol{K}_Y$. Since these possible neural representations are all related by a orthogonal transformation, we can fix $\boldsymbol{X}$ and $\boldsymbol{Y}$ arbitrarily and optimize their overlap over the set of orthogonal transformations $\mathcal{U}(N)$:

$$\begin{aligned} &\max_{\boldsymbol{X},\boldsymbol{Y}} \{|\operatorname{Tr}[\boldsymbol{X}^\top\boldsymbol{Y}]| : \boldsymbol{X}\boldsymbol{X}^\top = \boldsymbol{K}_X, \boldsymbol{Y}\boldsymbol{Y}^\top = \boldsymbol{K}_Y\} \\ &= \max_{U} \{|\operatorname{Tr}[\boldsymbol{X}^\top\boldsymbol{Y}\boldsymbol{U}]| : \boldsymbol{U} \in \mathcal{U}(N)\} \\ &= \|\boldsymbol{X}^\top\boldsymbol{Y}\| \quad = \mathcal{F}(\boldsymbol{K}_X, \boldsymbol{K}_Y). \end{aligned} \tag{25}$$

In the last equality we have used the result of lemma 2. This result, known in the quantum information community as Uhlmann's theorem [57], shows that the fidelity between covariance matrices can be understood as the solution to maximizing the overlap between neural representation matrices that are consistent with those covariance matrices. This result does not depend on the dimension N, provided it is at least the maximum rank of $\boldsymbol{K}_X$ and $\boldsymbol{K}_Y$ (adding extra dimensions beyond $\max(\text{rank}(\boldsymbol{K}_X), \text{rank}(\boldsymbol{K}_Y))$ does not affect the solution to this problem). Equation (25) implies that NBS is simply the same maximum overlap between 'consistent' neural representation matrices, normalized to lie in the interval $[0, 1]$. Without loss of generality, we can write the maximization problem in eq. (25) in terms of the unique $M \times M$ PSD square roots of the covariance matrices:

$$\mathcal{F}(\boldsymbol{K}_X, \boldsymbol{K}_Y) = \max_U \{|\operatorname{Tr}[(\boldsymbol{X}\boldsymbol{X}^\top)^{\frac{1}{2}}(\boldsymbol{Y}\boldsymbol{Y}^\top)^{\frac{1}{2}}\boldsymbol{U}]| : \boldsymbol{U} \in \mathcal{U}(M)\} \tag{26}$$

Choosing $\boldsymbol{U} = \boldsymbol{I}$ leads to the inequality:

$$NBS(\boldsymbol{K}_X, \boldsymbol{K}_Y) \geq \frac{\operatorname{Tr}[\boldsymbol{K}_X^{1/2}\boldsymbol{K}_Y^{1/2}]}{\sqrt{\operatorname{Tr}\boldsymbol{K}_X \operatorname{Tr}\boldsymbol{K}_Y}} = CKA(\boldsymbol{K}_X^{\frac{1}{2}}, \boldsymbol{K}_Y^{\frac{1}{2}}). \tag{27}$$

where we have dropped the absolute value signs because the Hilbert-Schmidt inner product between two PSD matrices is non-negative. The inequality is saturated when $\sqrt{\boldsymbol{K}_X}$ commutes with $\sqrt{\boldsymbol{K}_Y}$. This inequality shows that the CKA between two particular neural representation matrices, namely the PSD square roots $\sqrt{\boldsymbol{K}_X}$ and $\sqrt{\boldsymbol{K}_Y}$, appears as a suboptimal solution to the maximization in eq. (26). NBS($\boldsymbol{K}_X, \boldsymbol{K}_Y$), on the other hand, represents this same Hilber-Schmidt inner product maximized over all neural representation matrices consistent with $\boldsymbol{K}_X$ and $\boldsymbol{K}_Y$.

Lastly, we can derive an upper bound on NBS in terms of $CKA(\sqrt{\boldsymbol{K}_X}, \sqrt{\boldsymbol{K}_X})$ using the Fuchs-van de Graaf inequalities from quantum information theory (see section 7.3). We are lead to another set of inequalities which bound the deviation of $CKA(\sqrt{\boldsymbol{K}_X}, \sqrt{\boldsymbol{K}_X})$ from $NBS(\boldsymbol{K}_X, \boldsymbol{K}_Y)$:

$$1 - NBS(\boldsymbol{K}_X, \boldsymbol{K}_X) \leq 1 - CKA(\boldsymbol{K}_X^{1/2}, \boldsymbol{K}_Y^{1/2}) \leq \sqrt{1 - NBS(\boldsymbol{K}_X, \boldsymbol{K}_X)^2} \tag{28}$$

These upper and lower bounds are represented in fig. 2 (b) as the orange and blue dashed lines.

6 Discussion

Differences in neural representations are quantified by a wide variety of methods in the current literature [22]. In many cases, the relationships between these various quantities is unclear both conceptually and quantitatively. While prior works have made attempts to mathematically relate various (dis)similarity measures, our work greatly expands the scope of these comparisons. In particular, we show that two independently proposed approaches—shape distances [20, 58] and the normalized Bures similarity (NBS; [37, 54])—are essentially identical (see theorem 1). A notable feature of this equivalence is that the two methods are motivated from very different perspectives. The Procrustes and Riemannian shape distances can be viewed as the residual distance that is left after neural dimensions are aligned by an optimal rotation. In contrast, NBS directly compares the structure of two kernel matrices, $\boldsymbol{K}_X$ and $\boldsymbol{K}_Y$, without any alignment. Superficially, NBS looks similar to CKA (see eq. 23), but we have seen that these similarity scores utilize fundamentally different geometries (eqs. 21 and 22) and can produce discrepant quantitative outcomes as we demonstrated both analytically and numerically (see fig. 2).

NBS and Bures distance are rooted in a rich literature in quantum information theory [39, 35, 57]. Indeed, the bound on CKA derived in section 5.3 follows a classic result in this area known as Uhlmann's theorem. Similarly, the Bures geometry on PSD manifolds has been extensively studied in the context of optimal transport, yielding both theoretical insights and practical algorithms [5, 27]. Our work only scratches the surface of these connections, and future studies should seek to import additional findings from these well-developed nearby fields.

Our main result shows a duality between shape and Bures distances, and an important open question is whether similar dualities can be found for other (dis)similarity measures. If shape and Bures distances represent a truly unique link between the two major perspectives on the problem (summarized in sections 2.1 and 2.2), this provides a concrete motivation for their adoption. In short, these methods can enjoy the conceptual and practical advantages of each perspective, depending on the circumstance.

References

[1] Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. "Geometric Means in a Novel Vector Space Structure on Symmetric Positive-Definite Matrices". *SIAM Journal on Matrix Analysis and Applications* 29.1 (2007), pp. 328–347.

[2] Michael Atiyah. *Duality in Mathematics and Physics*. Lecture notes from the Institut de Matematica de la Universitat de Barcelona (IMUB). 2007.

[3] Richard Beals, David H Krantz, and Amos Tversky. "Foundations of multidimensional scaling." *Psychological review* 75.2 (1968), p. 127.

[4] Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. "On the Bures–Wasserstein distance between positive definite matrices". *Expositiones Mathematicae* 37.2 (2019), pp. 165–191.

[5] Sinho Chewi, Tyler Maunu, Philippe Rigollet, and Austin J Stromme. "Gradient descent algorithms for Bures-Wasserstein barycenters". *Conference on Learning Theory*. PMLR. 2020, pp. 1276–1304.

[6] Corinna Cortes, Mehryar Mohri, and Afshin Rostamizadeh. "Algorithms for learning kernels based on centered alignment". *J. Mach. Learn. Res.* 13.1 (2012), pp. 795–828.

[7] Nello Cristianini, John Shawe-Taylor, André Elisseeff, and Jaz Kandola. "On Kernel-Target Alignment". *Advances in Neural Information Processing Systems*. Ed. by T. Dietterich, S. Becker, and Z. Ghahramani. Vol. 14. MIT Press, 2001.

[8] Joel Dapello, Kohitij Kar, Martin Schrimpf, Robert Baldwin Geary, Michael Ferguson, David Daniel Cox, and James J. DiCarlo. "Aligning Model and Macaque Inferior Temporal Cortex Representations Improves Model-to-Human Behavioral Alignment and Adversarial Robustness". *The Eleventh International Conference on Learning Representations*. 2023.

[9] MohammadReza Davari, Stefan Horoi, Amine Natik, Guillaume Lajoie, Guy Wolf, and Eugene Belilovsky. "Reliability of CKA as a Similarity Measure in Deep Learning". *The Eleventh International Conference on Learning Representations*. 2023.

[10] Ian L. Dryden, Alexey Koloydenko, and Diwei Zhou. "Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging". *The Annals of Applied Statistics* 3.3 (2009), pp. 1102–1123.

[11] Lyndon R. Duong, Jingyang Zhou, Josue Nassar, Jules Berman, Jeroen Olieslagers, and Alex H. Williams. "Representational dissimilarity metric spaces for stochastic neural networks". *International Conference on Learning Representations*. 2023.

[12] Shimon Edelman. "Representation is representation of similarities". *Behavioral and brain sciences* 21.4 (1998), pp. 449–467.

[13] Alexander G. de G. Matthews, Mark Rowland, Jiri Hron, Richard E. Turner, and Zoubin Ghahramani. *Gaussian Process Behaviour in Wide Deep Neural Networks*. 2018.

[14] Juan A Gallego, Matthew G Perich, Raeed H Chowdhury, Sara A Solla, and Lee E Miller. "Long-term stability of cortical population dynamics underlying consistent behavior". *Nature neuroscience* 23.2 (2020), pp. 260–270.

[15] Anne Harrington, Vasha DuTell, Ayush Tewari, Mark Hamilton, Simon Stent, Ruth Rosenholtz, and William T. Freeman. "Exploring perceptual straightness in learned visual representations". *The Eleventh International Conference on Learning Representations*. 2023.

[16] Olivier J Hénaff, Yoon Bai, Julie A Charlton, Ian Nauhaus, Eero P Simoncelli, and Robbe LT Goris. "Primary visual cortex straightens natural video trajectories". *Nature communications* 12.1 (2021), p. 5982.

[17] Arthur Jacot, Franck Gabriel, and Clément Hongler. "Neural tangent kernel: Convergence and generalization in neural networks". *Advances in neural information processing systems* 31 (2018).

[18] Haydn T. Jones, Jacob M. Springer, Garrett T. Kenyon, and Juston S. Moore. "If you've trained one you've trained them all: inter-architecture similarity increases with robustness". *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*. Ed. by James Cussens and Kun Zhang. Vol. 180. Proceedings of Machine Learning Research. PMLR, 2022, pp. 928–937.

[19] D G Kendall. "The Diffusion of Shape". *Adv. Appl. Probab.* 9.3 (1977), pp. 428–430.

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

[21] Tim C. Kietzmann, Courtney J. Spoerer, Lynn K. A. Sörensen, Radoslaw M. Cichy, Olaf Hauk, and Nikolaus Kriegeskorte. "Recurrence is required to capture the representational dynamics of the human visual system". *Proceedings of the National Academy of Sciences* 116.43 (2019), pp. 21854–21863.

[22] Max Klabunde, Tobias Schumacher, Markus Strohmaier, and Florian Lemmerich. *Similarity of Neural Network Models: A Survey of Functional and Representational Measures*. 2023.

[23] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. "Similarity of Neural Network Representations Revisited". *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 3519–3529.

[24] Nikolaus Kriegeskorte and Jörn Diedrichsen. "Inferring brain-computational mechanisms with models of activity measurements". *Philosophical Transactions of the Royal Society B: Biological Sciences* 371.1705 (2016), p. 20160278.

[25] Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. "Representational similarity analysis - connecting the branches of systems neuroscience". *Front. Syst. Neurosci.* 2 (2008), p. 4.

[26] Nikolaus Kriegeskorte, Marieke Mur, Douglas A Ruff, Roozbeh Kiani, Jerzy Bodurka, Hossein Esteky, Keiji Tanaka, and Peter A Bandettini. "Matching categorical object representations in inferior temporal cortex of man and monkey". *Neuron* 60.6 (2008), pp. 1126–1141.

[27] Alexey Kroshnin, Vladimir Spokoiny, and Alexandra Suvorikova. "Statistical inference for Bures–Wasserstein barycenters". *The Annals of Applied Probability* 31.3 (2021), pp. 1264–1298.

[28] Jaehoon Lee, Jascha Sohl-dickstein, Jeffrey Pennington, Roman Novak, Sam Schoenholz, and Yasaman Bahri. "Deep Neural Networks as Gaussian Processes". *International Conference on Learning Representations*. 2018.

[29] Yixuan Li, Jason Yosinski, Jeff Clune, Hod Lipson, and John Hopcroft. "Convergent Learning: Do different neural networks learn the same representations?" *Proceedings of the 1st International Workshop on Feature Extraction: Modern Questions and Challenges at NIPS 2015*. Ed. by Dmitry Storcheus, Afshin Rostamizadeh, and Sanjiv Kumar. Vol. 44. Proceedings of Machine Learning Research. Montreal, Canada: PMLR, 2015, pp. 196–212.

[30] Yeong-Cherng Liang, Yu-Hao Yeh, Paulo E M F Mendonça, Run Yan Teh, Margaret D Reid, and Peter D Drummond. "Quantum fidelity measures for mixed states". *Rep. Prog. Phys.* 82.7 (2019), p. 076001.

[31] 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.

[32] Luigi Malagò, Luigi Montrucchio, and Giovanni Pistone. "Wasserstein Riemannian geometry of Gaussian densities". *Information Geometry* 1.2 (2018), pp. 137–179.

[33] Anton Mallasto and Aasa Feragen. "Learning from uncertain curves: The 2-Wasserstein metric for Gaussian processes". *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc., 2017.

[34] Valentina Masarotto, Victor M. Panaretos, and Yoav Zemel. "Procrustes Metrics on Covariance Operators and Optimal Transportation of Gaussian Processes". *Sankhya A* 81.1 (2019), pp. 172–213.

[35] Paulo E. M. F. Mendon ça, Reginaldo d. J. Napolitano, Marcelo A. Marchiolli, Christopher J. Foster, and Yeong-Cherng Liang. "Alternative fidelity measure between quantum states". *Phys. Rev. A* 78 (5 2008), p. 052330.

[36] Ari Morcos, Maithra Raghu, and Samy Bengio. "Insights on representational similarity in neural networks with canonical correlation". *Advances in neural information processing systems* 31 (2018).

[37] Boris Muzellec and Marco Cuturi. "Generalizing Point Embeddings using the Wasserstein Space of Elliptical Distributions". *Advances in Neural Information Processing Systems 2018*. 2018.

[38] Thao Nguyen, Maithra Raghu, and Simon Kornblith. "Do Wide and Deep Networks Learn the Same Things? Uncovering How Neural Network Representations Vary with Width and Depth". *International Conference on Learning Representations*. 2021.

[39] Michael A Nielsen and Isaac L Chuang. "Quantum computation and quantum information". *Phys. Today* 54.2 (2001), p. 60.

[40] Michael A. Nielsen and Isaac L. Chuang. *Quantum Computation and Quantum Information*. Cambridge University Press, 2000.

[41] Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. "A toolbox for representational similarity analysis". *PLoS computational biology* 10.4 (2014), e1003553.

[42] Gabriel Peyré and Marco Cuturi. "Computational Optimal Transport: With Applications to Data Science". *Foundations and Trends® in Machine Learning* 11.5-6 (2019), pp. 355–607.

[43] Davide Pigoli, John A. D. Aston, Ian L. Dryden, and Piercesare Secchi. "Distances and inference for covariance operators". *Biometrika* 101.2 (2014), pp. 409–422.

[44] 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).

[45] Abigail A Russo, Sean R Bittner, Sean M Perkins, Jeffrey S Seely, Brian M London, Antonio H Lara, Andrew Miri, Najja J Marshall, Adam Kohn, Thomas M Jessell, et al. "Motor cortex embeds muscle-like commands in an untangled population response". *Neuron* 97.4 (2018), pp. 953–966.

[46] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[47] Peter H Schönemann. "A generalized solution of the orthogonal procrustes problem". *Psychometrika* 31.1 (1966), pp. 1–10.

[48] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. "Brain-score: Which artificial neural network for object recognition is most brain-like?" *BioRxiv* (2018), p. 407007.

[49] Heiko H Schütt, Alexander D Kipnis, Jörn Diedrichsen, and Nikolaus Kriegeskorte. "Statistical inference on representational geometries". *eLife* 12 (2023). Ed. by John T Serences and Timothy E Behrens, e82566.

[50] Mahdiyar Shahbazi, Ali Shirali, Hamid Aghajan, and Hamed Nili. "Using distance on the Riemannian manifold to compare representations in brain and in models". *NeuroImage* 239 (2021), p. 118271.

[51] Katherine R. Storrs, Tim C. Kietzmann, Alexander Walther, Johannes Mehrer, and Nikolaus Kriegeskorte. "Diverse Deep Neural Networks All Predict Human Inferior Temporal Cortex Well, After Training and Fitting". *Journal of Cognitive Neuroscience* 33.10 (2021), pp. 2044–2064.

[52] 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 arXiv:2310.13018* (2023).

[53] David Sussillo, Mark M Churchland, Matthew T Kaufman, and Krishna V Shenoy. "A neural network that finds a naturalistic solution for the production of muscle activity". *Nature neuroscience* 18.7 (2015), pp. 1025–1033.

[54] Shuai Tang, Wesley J. Maddox, Charlie Dickens, Tom Diethe, and Andreas Damianou. *Similarity of Neural Networks with Gradients*. 2020.

[55] Yann Thanwerdas and Xavier Pennec. "Bures–Wasserstein Minimizing Geodesics between Covariance Matrices of Different Ranks". *SIAM Journal on Matrix Analysis and Applications* 44.3 (2023), pp. 1447–1476.

[56] Alexander Walther, Hamed Nili, Naveed Ejaz, Arjen Alink, Nikolaus Kriegeskorte, and Jörn Diedrichsen. "Reliability of dissimilarity measures for multi-voxel pattern analysis". *Neuroimage* 137 (2016), pp. 188–200.

[57] John Watrous. *The theory of quantum information*. Cambridge university press, 2018.

[58] Alex H. Williams, Erin Kunz, Simon Kornblith, and Scott W. Linderman. "Generalized Shape Metrics on Neural Representations". *Advances in Neural Information Processing Systems*. Vol. 34. 2021.

7 Proofs

We use the following notation: if $\boldsymbol{A}$ is a positive semidefinite matrix, then it admits a decomposition $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ and a unique positive semidefinite square root $\boldsymbol{A}^{1/2} = \boldsymbol{U}\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^\top$.

7.1 Proof of Lemma 1

Proof of lemma 1. We begin with the Procrustes distance:

$$\begin{aligned}
\Big[\mathcal{P}(\boldsymbol{X},\boldsymbol{Y})\Big]^2 &= \min_{\boldsymbol{Q}^\top\boldsymbol{Q}=\boldsymbol{I}} \|\boldsymbol{C}\boldsymbol{X} - \boldsymbol{C}\boldsymbol{Y}\boldsymbol{Q}\|_F^2 \\
&= \min_{\boldsymbol{Q}^\top\boldsymbol{Q}=\boldsymbol{I}} \Big(\operatorname{Tr}[\boldsymbol{X}^\top\boldsymbol{C}^\top\boldsymbol{C}\boldsymbol{X}] + \operatorname{Tr}[\boldsymbol{Y}^\top\boldsymbol{C}^\top\boldsymbol{C}\boldsymbol{Y}] - 2\operatorname{Tr}[\boldsymbol{X}^\top\boldsymbol{C}^\top\boldsymbol{C}\boldsymbol{Y}\boldsymbol{Q}]\Big) \\
&= \operatorname{Tr}[\boldsymbol{X}^\top\boldsymbol{C}\boldsymbol{X}] + \operatorname{Tr}[\boldsymbol{Y}^\top\boldsymbol{C}\boldsymbol{Y}] - 2\max_{\boldsymbol{Q}^\top\boldsymbol{Q}=\boldsymbol{I}} \operatorname{Tr}\Big[\boldsymbol{X}^\top\boldsymbol{C}\boldsymbol{Y}\boldsymbol{Q}\Big] \\
&= \operatorname{Tr}\left[\boldsymbol{\Sigma}_{XX}\right] + \operatorname{Tr}\left[\boldsymbol{\Sigma}_{YY}\right] - 2\max_{\boldsymbol{Q}^\top\boldsymbol{Q}=\boldsymbol{I}} \operatorname{Tr}\left[\boldsymbol{\Sigma}_{XY}\boldsymbol{Q}\right] \\
&= \operatorname{Tr}\left[\boldsymbol{\Sigma}_{XX}\right] + \operatorname{Tr}\left[\boldsymbol{\Sigma}_{YY}\right] - 2\|\boldsymbol{\Sigma}_{XY}\|
\end{aligned}$$

where the step in the last line is the well-known result of Schönemann [47]. Similarly for the Riemannian shape distance:

$$\begin{aligned}
\theta\ (\boldsymbol{X},\boldsymbol{Y}) &= \min_{\boldsymbol{Q}^\top\boldsymbol{Q}=\boldsymbol{I}} \cos^{\ 1}\left(\frac{\operatorname{Tr}[\boldsymbol{X}^\top\boldsymbol{C}^\top\boldsymbol{C}\boldsymbol{Y}\boldsymbol{Q}]}{\sqrt{\operatorname{Tr}[\boldsymbol{X}^\top\boldsymbol{C}^\top\boldsymbol{C}\boldsymbol{X}]\operatorname{Tr}[\boldsymbol{Y}^\top\boldsymbol{C}^\top\boldsymbol{C}\boldsymbol{Y}]}}\right) \\
&= \min_{\boldsymbol{Q}^\top\boldsymbol{Q}=\boldsymbol{I}} \cos^{\ 1}\left(\frac{\operatorname{Tr}[\boldsymbol{X}^\top\boldsymbol{C}\boldsymbol{Y}\boldsymbol{Q}]}{\sqrt{\operatorname{Tr}[\boldsymbol{X}^\top\boldsymbol{C}\boldsymbol{X}]\operatorname{Tr}[\boldsymbol{Y}^\top\boldsymbol{C}\boldsymbol{Y}]}}\right) \\
&= \cos^{\ 1}\left(\frac{\max_{\boldsymbol{Q}^\top\boldsymbol{Q}=\boldsymbol{I}} \operatorname{Tr}[\boldsymbol{\Sigma}_{XY}\boldsymbol{Q}]}{\sqrt{\operatorname{Tr}[\boldsymbol{\Sigma}_{XX}]\operatorname{Tr}[\boldsymbol{\Sigma}_{YY}]}}\right) \\
&= \cos^{\ 1}\left(\frac{\|\boldsymbol{\Sigma}_{XY}\|}{\sqrt{\operatorname{Tr}[\boldsymbol{\Sigma}_{XX}]\operatorname{Tr}[\boldsymbol{\Sigma}_{YY}]}}\right)
\end{aligned}$$

as claimed by the lemma. □

7.2 Proof of Lemma 2

Proof. First, the nonzero singular values of $\boldsymbol{X}^\top\boldsymbol{Y}$ are equal to the square root of the nonzero eigenvalues of $\boldsymbol{A} = \boldsymbol{X}^\top\boldsymbol{Y}\boldsymbol{Y}^\top\boldsymbol{X}$. Thus,

$$\|\boldsymbol{X}^\top\boldsymbol{Y}\|_* = \operatorname{Tr}[(\boldsymbol{X}^\top\boldsymbol{Y}\boldsymbol{Y}^\top\boldsymbol{X})^{1/2}] = \operatorname{Tr}[\boldsymbol{A}^{1/2}] \tag{29}$$

Next, we argue that every nonzero eigenvalue of $\boldsymbol{A} \in \mathsf{R}^{N\ N}$ is also an eigenvalue of $\boldsymbol{B} = \boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{Y}\boldsymbol{Y}^\top \in \mathsf{R}^{M\ M}$. To see this, suppose $\lambda \neq 0$ is an eigenvalue of $\boldsymbol{A}$ with eigenvector $\boldsymbol{v} \in \mathsf{R}^N$. Then, $\boldsymbol{w} = \boldsymbol{X}\boldsymbol{v} \in \mathsf{R}^M$ is an eigenvector of $\boldsymbol{B}$ with the same eigenvalue, since:

$$\boldsymbol{X}^\top\boldsymbol{Y}\boldsymbol{Y}^\top\boldsymbol{X}\boldsymbol{v} = \lambda\boldsymbol{v} \tag{30}$$

$$\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{Y}\boldsymbol{Y}^\top\boldsymbol{X}\boldsymbol{v} = \lambda\boldsymbol{X}\boldsymbol{v} \qquad \text{(multiply both sides on the left by } \boldsymbol{X}\text{.)} \tag{31}$$

$$\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{Y}\boldsymbol{Y}^\top\boldsymbol{w} = \lambda\boldsymbol{w} \qquad \text{(let } \boldsymbol{w} = \boldsymbol{X}\boldsymbol{v}\text{.)} \tag{32}$$

Notice that eq. (30) together with $\lambda \neq 0$ implies that $\boldsymbol{w} = \boldsymbol{X}\boldsymbol{v} \neq \boldsymbol{0}$. Further, $\boldsymbol{B}$ does not contain any additional nonzero eigenvalues or additional repeated eigenvalues, since:

$$\operatorname{rank}(\boldsymbol{B}) = \operatorname{rank}(\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{Y}\boldsymbol{Y}^\top) \tag{33}$$

$$\leq \operatorname{rank}(\boldsymbol{X}^\top\boldsymbol{Y}) \qquad \text{(matrix product rank inequality)} \tag{34}$$

$$= \operatorname{rank}(\boldsymbol{X}^\top\boldsymbol{Y}\boldsymbol{Y}^\top\boldsymbol{X}) \qquad \text{(rank of a matrix and its Gram matrix are equal)} \tag{35}$$

$$= \operatorname{rank}(\boldsymbol{A}). \tag{36}$$

Thus, we've shown that the non-zero eigenvalues of $\boldsymbol{A}$ and $\boldsymbol{B}$ are equal.

Next, we define $\boldsymbol{C} = (\boldsymbol{X}\boldsymbol{X}^\top)^{1/2}\boldsymbol{Y}\boldsymbol{Y}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{1/2} \in \mathsf{R}^{M\ \ M}$, and argue that it has the same eigenvalue spectrum as $\boldsymbol{B}$. To see this, suppose that $\lambda \neq 0$ is an eigenvalue of $\boldsymbol{C}$ with eigenvector $\boldsymbol{z} \in \mathsf{R}^M$. Then, $\boldsymbol{w} = (\boldsymbol{X}\boldsymbol{X}^\top)^{1/2}\boldsymbol{z}$ is an eigenvector of $\boldsymbol{B}$ with the same eigenvalue:

$$(\boldsymbol{X}\boldsymbol{X}^\top)^{1/2}\boldsymbol{Y}\boldsymbol{Y}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{1/2}\boldsymbol{z} = \lambda\boldsymbol{z} \tag{37}$$

$$\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{Y}\boldsymbol{Y}^\top(\boldsymbol{X}\boldsymbol{X}^\top)^{1/2}\boldsymbol{z} = \lambda(\boldsymbol{X}\boldsymbol{X}^\top)^{1/2}\boldsymbol{z} \quad \text{(multiply on the left by } (\boldsymbol{X}\boldsymbol{X}^\top)^{1/2}\text{.)} \tag{38}$$

$$\boldsymbol{X}\boldsymbol{X}^\top\boldsymbol{Y}\boldsymbol{Y}^\top\boldsymbol{w} = \lambda\boldsymbol{w} \qquad \text{(let } \boldsymbol{w} = (\boldsymbol{X}\boldsymbol{X}^\top)^{1/2}\boldsymbol{z}\text{.)} \tag{39}$$

Combining this with our argument above, we conclude that the non-zero eigenvalues of $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ are equal. Furthermore, from the definition $\boldsymbol{C}$ and the definition of the fidelity in eq. (12), we have $\mathrm{Tr}[\boldsymbol{C}^{1/2}] = \mathcal{F}(\boldsymbol{X}\boldsymbol{X}^\top, \boldsymbol{Y}\boldsymbol{Y}^\top)$. We can therefore conclude the proof since, recalling eq. (29), we have:

$$\|\boldsymbol{X}^\top\boldsymbol{Y}\|_* = \mathrm{Tr}[\boldsymbol{A}^{1/2}] = \mathrm{Tr}[\boldsymbol{B}^{1/2}] = \mathrm{Tr}[\boldsymbol{C}^{1/2}] = \mathcal{F}(\boldsymbol{X}\boldsymbol{X}^\top, \boldsymbol{Y}\boldsymbol{Y}^\top) \tag{40}$$

as claimed by the lemma. □

7.3 Applying the Fuchs-van de Graaf inequalities to NBS and CKA

One of the Fuchs-van de Graaf inequalities tells us how the fidelity bounds the nuclear norm of the difference between positive semidefinite matrices ρ and σ with trace equal to 1 (known as the trace distance) [57]:

$$\|\rho - \sigma\|_* \leq 2\sqrt{1 - \mathcal{F}(\rho, \sigma)^2} \tag{41}$$

Rewriting $\rho = \boldsymbol{K}_X / \mathrm{Tr}[\boldsymbol{K}_X]$ and $\sigma = \boldsymbol{K}_Y / \mathrm{Tr}[\boldsymbol{K}_Y]$ allows us to recognize $\mathcal{F}(\rho, \sigma)$ as $NBS(\boldsymbol{K}_X, \boldsymbol{K}_Y)$. Using the norm inequality for positive semi definite operators $\boldsymbol{A}$ and $\boldsymbol{B}$, $\|\boldsymbol{A} - \boldsymbol{B}\|_* \geq \|\sqrt{\boldsymbol{A}} - \sqrt{\boldsymbol{B}}\|_F^2$ and expanding, we have:

$$1 - NBS(\boldsymbol{K}_X, \boldsymbol{K}_X) \leq 1 - CKA(\boldsymbol{K}_X^{1/2}, \boldsymbol{K}_Y^{1/2}) \leq \sqrt{1 - NBS(\boldsymbol{K}_X, \boldsymbol{K}_X)^2} \tag{42}$$

where the first inequality on the left hand side is from eq. (27). The right hand side inequality can be rewritten

$$NBS(\boldsymbol{K}_X, \boldsymbol{K}_X) \leq \sqrt{1 - (1 - CKA(\boldsymbol{K}_X^{1/2}, \boldsymbol{K}_Y^{1/2})^2}. \tag{43}$$

which defines the orange dashed curve in fig. 2 (b).