A Appendix

Accompanies Harvey et. al. (2024) "What Representational Similarity Measures Imply about Decodable Information."

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 $\mathcal{P}(\tilde{\boldsymbol{X}}, \tilde{\boldsymbol{Y}})$. Tildes are omitted in this section for clarity, but we note that this result is applied in the main text to the normalized representations $\tilde{\boldsymbol{X}}$ and $\tilde{\boldsymbol{Y}}$.

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

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

where $\operatorname{vec}(\boldsymbol{X})$ 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 [34], 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 [16] 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 [34], 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 $\text{Tr}\,\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^\top = \text{Tr}\,\tilde{\boldsymbol{Y}}\tilde{\boldsymbol{Y}}^\top = 1$, 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, [16] 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 $\frac{1}{\text{Tr}\,\boldsymbol{A}\boldsymbol{A}^\top}\boldsymbol{A}\boldsymbol{A}^\top$ and $\frac{1}{\text{Tr}\,\boldsymbol{B}\boldsymbol{B}^\top}\boldsymbol{B}\boldsymbol{B}^\top$ 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.

A.3 Generalizations and Interpretations as $M \to \infty$

In the main text, we considered quantifying representational similarity across a finite set of M stimulus conditions. Further, when considering the average distance in decoding performance, we have assumed that $\mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top] = \boldsymbol{I}$. Here, we briefly discuss how to relax both constraints, leaving a full investigation to future work. Readers will need familiarity with the basic principles of kernel methods and Gaussian processes in machine learning (see e.g. [31, 36]) to follow along with certain results in this section.

To begin, we must revisit and refine our theoretical framework developed in section 2. Under proposition 2, we treated the neural response matrices, $\boldsymbol{X}$ and $\boldsymbol{Y}$, as constants while we treated the decoder targets, $\boldsymbol{z}$, as a random variable. In this section we will treat $\boldsymbol{X}$, $\boldsymbol{Y}$ and $\boldsymbol{z}$ as joint random variables, as has been done in prior work [27, 3]. Specifically, let $\mathcal{U}$ denote the space

of possible stimulus inputs (e.g. the space of natural images or grammatically correct English sentences). We are interested in quantifying similarity between two neural networks, $f : \mathcal{U} \mapsto \mathbb{R}^{N_X}$ and $g : \mathcal{U} \mapsto \mathbb{R}^{N_Y}$, across this stimulus space. Let $P_{\boldsymbol{u}}$ denote a distribution with support on $\mathcal{U}$, and let $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M$ denote independent samples from $P_{\boldsymbol{u}}$. As before, we collect the network responses into matrices $\boldsymbol{X} \in \mathbb{R}^{M \times N_X}$ and $\boldsymbol{Y} \in \mathbb{R}^{M \times N_Y}$, which we assume to be mean centered, $\mathbb{E}[f(\boldsymbol{u})] = \mathbb{E}[g(\boldsymbol{u})] = \boldsymbol{0}$. The rows of these matrices are given by $\boldsymbol{x}_i = f(\boldsymbol{u}_i)$ and $\boldsymbol{y}_i = g(\boldsymbol{u}_i)$, and we assume that $\|f(\boldsymbol{u}_i)\|_2 \leq \infty$ for all inputs indexed by $i \in \{1, ..., M\}$. This is a common assumption that can be interpreted in the neuroscience sense as assuming neural firing rates are bounded from above (they cannot become arbitrarily large).

Finally, we define a decoding task as a function $\eta : \mathcal{U} \mapsto \mathbb{R}$. For M finite samples (as considered in the main text), the decoder target is the vector with elements consisting of η evaluated at each sampled input: $\boldsymbol{z} = [\eta(\boldsymbol{u}_1) \quad \ldots \quad \eta(\boldsymbol{u}_M)]^\top$. As in sections 3 and 4, we will be interested in characterizing the *average* decoder alignment over multiple tasks. In the main text we considered $\mathbb{E}_{\boldsymbol{z}} \langle \boldsymbol{X}\boldsymbol{w}^*, \boldsymbol{Y}\boldsymbol{v}^* \rangle$, but we would like to now generalize this to an expectation of an inner product between two *functions* over the input space. We consider η to be drawn from a Gaussian process, $\eta \sim \mathcal{GP}(q)$, with a covariance operator $q : \mathcal{U} \times \mathcal{U} \mapsto \mathbb{R}$. Intuitively, q defines the difficulty of the distribution over regression tasks by specifying smoothness with respect to the input space $\mathcal{U}$. For example, the popular squared exponential or radial basis function (RBF) kernel:

$$q(\boldsymbol{u}, \boldsymbol{u}') = \exp\left(-\frac{\|\boldsymbol{u} - \boldsymbol{u}'\|_2^2}{2\gamma}\right) \tag{54}$$

comes equipped with a length scale parameter $\gamma > 0$ that determines the smoothness of functions sampled $\eta \sim \mathcal{GP}(q)$. In the limit that $\gamma \to 0$ we get the Kronecker delta function:

$$\delta(\boldsymbol{u}, \boldsymbol{u}') = \begin{cases} 1 & \boldsymbol{u} = \boldsymbol{u}' \\ 0 & \boldsymbol{u} \neq \boldsymbol{u}' \end{cases} \tag{55}$$

In the remainder of this section, we show that the average decoding similarity converges to an expected inner product between kernels defined by the two neural systems. Specifically, let us define $k : \mathcal{U} \times \mathcal{U} \mapsto \mathbb{R}$ and $h : \mathcal{U} \times \mathcal{U} \mapsto \mathbb{R}$ as follows:

$$k(\boldsymbol{u}, \boldsymbol{u}') = f(\boldsymbol{u})^\top \boldsymbol{G}(f)^{-1} f(\boldsymbol{u}') \qquad \text{and} \qquad h(\boldsymbol{u}, \boldsymbol{u}') = g(\boldsymbol{u})^\top \boldsymbol{G}(g)^{-1} g(\boldsymbol{u}') \tag{56}$$

where we have generalized eq. (4) as:

$$\boldsymbol{G}(f) = a\mathbb{E}[f(\boldsymbol{u}) f(\boldsymbol{u})^\top] + b\boldsymbol{I} \tag{57}$$

We denote by $L^2(\mathcal{U})$ the space of L^2-integrable functions from the set $\mathcal{U}$ to $\mathbb{R}$. Consider the integral operators $\mathcal{K} : L^2(\mathcal{U}) \to L^2(\mathcal{U})$ and $\mathcal{H} : L^2(\mathcal{U}) \to L^2(\mathcal{U})$ associated with the kernels in eq. (56). We define the functions $\eta_{\mathcal{K}}(\boldsymbol{u})$ and $\eta_{\mathcal{H}}(\boldsymbol{u})$ as these operators acting on the decoding task function $\eta(\boldsymbol{u})$:

$$[\mathcal{K}\eta](\boldsymbol{u}) = \int k(\boldsymbol{u}, \boldsymbol{u}') \eta(\boldsymbol{u}') p(\boldsymbol{u}') d\boldsymbol{u}' \equiv \eta_{\mathcal{K}}(\boldsymbol{u}) \tag{58}$$

$$[\mathcal{H}\eta](\boldsymbol{u}) = \int h(\boldsymbol{u}, \boldsymbol{u}') \eta(\boldsymbol{u}') p(\boldsymbol{u}') d\boldsymbol{u}' \equiv \eta_{\mathcal{H}}(\boldsymbol{u}) \tag{59}$$

The functions $\eta_{\mathcal{K}}(\boldsymbol{u})$ and $\eta_{\mathcal{H}}(\boldsymbol{u})$ are analogous to the vectors $\boldsymbol{X}\boldsymbol{w}^* = \boldsymbol{K}_X \in \mathbb{R}^M$ and $\boldsymbol{Y}\boldsymbol{v}^* = \boldsymbol{K}_Y \in \mathbb{R}^M$, and can be thought of as the 'decoded' functions. We will define the infinite-dimensional decoding similarity as the inner product:

$$S = \langle \eta_{\mathcal{K}}, \eta_{\mathcal{H}} \rangle_{L^2} = \int_{\mathcal{U}} \eta_{\mathcal{K}}(\boldsymbol{u}) \eta_{\mathcal{H}}(\boldsymbol{u}) p(\boldsymbol{u}) d\boldsymbol{u} \tag{60}$$

Then we have, using the definitions in eq. (58) and eq. (59),

$$S = \int_{\mathcal{U}} \int_{\mathcal{U}} \int_{\mathcal{U}} k(\boldsymbol{u}, \boldsymbol{u}')\eta(\boldsymbol{u}')\eta(\boldsymbol{u}'')h(\boldsymbol{u}, \boldsymbol{u}'')\, p(\boldsymbol{u})d\boldsymbol{u}\, p(\boldsymbol{u}')d\boldsymbol{u}'\, p(\boldsymbol{u}'')d\boldsymbol{u}''. \tag{61}$$

This integral measures the similarity between two neural systems f and g for a particular choice of decoding task function η. However, following the main text, we would like to take the expectation over some ensemble of decoding task functions η.

$$\mathbb{E}_\eta[S] = \mathbb{E}_\eta \langle \eta_{\mathcal{K}}, \eta_{\mathcal{H}} \rangle_{L^2} \tag{62}$$

Assuming the conditions to invoke Fubini's theorem regarding changing the ordering of integration are met, we have

$$\mathbb{E}_\eta[S] = \int_{\mathcal{U}} \int_{\mathcal{U}} \int_{\mathcal{U}} k(\boldsymbol{u}, \boldsymbol{u}')q(\boldsymbol{u}', \boldsymbol{u}'')h(\boldsymbol{u}, \boldsymbol{u}'')\, p(\boldsymbol{u})d\boldsymbol{u}\, p(\boldsymbol{u}')d\boldsymbol{u}'\, p(\boldsymbol{u}'')d\boldsymbol{u}''. \tag{63}$$

We now recognize this integral as this as the trace of a composition of integral operators,

$$\mathbb{E}_\eta[S] = \mathrm{Tr}[\mathcal{K}\mathcal{Q}\mathcal{H}] \tag{64}$$

where $\mathcal{Q} : L^2(\mathcal{U}) \to L^2(\mathcal{U})$ is the covariance operator associated with the Gaussian process covariance function q. This quantity is finite, since all of the operators involved are Hilbert-Schmidt and trace-class.

We now would like to show that the finite-dimensional notion of "decoding similarity" introduced in the main text (appropriately normalized) can be thought of as a plug-in estimator of the quantity eq. (64), and that this estimator converges to $\mathbb{E}_\eta[S]$ as $M \to \infty$.

A 'plug-in' estimator for this could be to use the Gram matrices for each kernel and matrix multiplication to approximate the integrals. First, we sample M examples from the input distribution $\boldsymbol{u} \sim P_{\boldsymbol{u}}$ and construct the Gram matrices $\boldsymbol{Q}_{ij} = q(\boldsymbol{u}_i, \boldsymbol{u}_j)$ and

$$\boldsymbol{K}_{ij} = k(\boldsymbol{u}_i, \boldsymbol{u}_j) = f(\boldsymbol{u}_i)^\top \boldsymbol{G}(f)^{-1} f(\boldsymbol{u}_j) \tag{65}$$

$$\boldsymbol{H}_{ij} = h(\boldsymbol{u}_i, \boldsymbol{u}_j) = g(\boldsymbol{u}_i)^\top \boldsymbol{G}(g)^{-1} g(\boldsymbol{u}_j). \tag{66}$$

Then we have, by the strong law of large numbers,

$$\int_{\mathcal{U}} \int_{\mathcal{U}} \int_{\mathcal{U}} k(\boldsymbol{u}, \boldsymbol{u}')q(\boldsymbol{u}', \boldsymbol{u}'')h(\boldsymbol{u}, \boldsymbol{u}'')\, p(\boldsymbol{u})d\boldsymbol{u}\, p(\boldsymbol{u}')d\boldsymbol{u}'\, p(\boldsymbol{u}'')d\boldsymbol{u}'' = \lim_{M\to\infty} \frac{1}{M^3} \sum_{ijk} K_{ij}Q_{jk}H_{ki}$$

or

$$\mathrm{Tr}[\mathcal{K}\mathcal{Q}\mathcal{H}] = \lim_{M\to\infty} \frac{1}{M^3} \mathrm{Tr}\, \boldsymbol{K}\boldsymbol{Q}\boldsymbol{H}. \tag{67}$$

However, the 'true' Gram matrices generated from the kernels in eq. (65) and eq. (66) in practice are also estimated, as the $N \times N$ regularization matrices $\boldsymbol{G}(f)$ and $\boldsymbol{G}(g)$ are potentially estimated from the same M samples from $P_{\boldsymbol{u}}$. We estimate $\boldsymbol{G}(f)$ and $\boldsymbol{G}(g)$ using the plug-in estimates of the $N \times N$ covariance matrices $\mathbb{E}[f(\boldsymbol{u})f(\boldsymbol{u})^\top]$ and $\mathbb{E}[g(\boldsymbol{u})g(\boldsymbol{u})^\top]$.

$$\boldsymbol{G}(\boldsymbol{X}) = \frac{\alpha}{M}\boldsymbol{X}^\top\boldsymbol{X} + \beta\boldsymbol{I} \xrightarrow[M\to\infty]{} \alpha\mathbb{E}[f(\boldsymbol{u})f(\boldsymbol{u})^\top] + \beta\boldsymbol{I} = \boldsymbol{G}(f) \tag{68}$$

$$\boldsymbol{G}(\boldsymbol{Y}) = \frac{\alpha}{M}\boldsymbol{Y}^\top\boldsymbol{Y} + \beta\boldsymbol{I} \xrightarrow[M\to\infty]{} \alpha\mathbb{E}[g(\boldsymbol{u})g(\boldsymbol{u})^\top] + \beta\boldsymbol{I} = \boldsymbol{G}(g) \tag{69}$$

More precisely, it can be shown that the empirical regularization matrices $\boldsymbol{G}(\boldsymbol{X})$ and $\boldsymbol{G}(\boldsymbol{Y})$ concentrates around $\boldsymbol{G}(f)$ and $\boldsymbol{G}(g)$ in operator norm:

$$\|\boldsymbol{G}(\boldsymbol{X}) - \boldsymbol{G}(f)\|_{op} \xrightarrow[M\to\infty]{} 0 \tag{70}$$

$$\|\boldsymbol{G}(\boldsymbol{Y}) - \boldsymbol{G}(g)\|_{op} \xrightarrow[M\to\infty]{} 0 \tag{71}$$

Bounds on the rate of this convergence in operator norm can be obtained using elementary matrix concentration inequalities. We now must argue that $\boldsymbol{G}(\boldsymbol{X})^{-1}$ concentrates to $\boldsymbol{G}(f)^{-1}$ and similarly $\boldsymbol{G}(\boldsymbol{Y})^{-1}$ concentrates to $\boldsymbol{G}(g)^{-1}$. Since the matrices $\boldsymbol{G}(\boldsymbol{X})$ and $\boldsymbol{G}(f)$ are symmetric and positive definite, we have

$$\|\boldsymbol{G}(\boldsymbol{X})^{-1}\|_{op} = \max_{v:\|v\|\leq 1} \|(\frac{\alpha}{M}\boldsymbol{X}^\top\boldsymbol{X} + \beta\boldsymbol{I})^{-1}v\|_2 = \frac{1}{\lambda_{min}(\boldsymbol{G}(\boldsymbol{X}))} \leq \frac{1}{\beta} \tag{72}$$

$$\|\boldsymbol{G}(f)^{-1}\|_{op} = \max_{v:\|v\|\leq 1} \|(\alpha\mathbb{E}[f(\boldsymbol{u})f(\boldsymbol{u})^\top] + \beta\boldsymbol{I})^{-1}v\|_2 = \frac{1}{\lambda_{min}(\boldsymbol{G}(f))} \leq \frac{1}{\beta} \tag{73}$$

where the last inequality on the right hand side is a consequence of the Courant-Fischer theorem and the positive-semidefiniteness of the covariance matrix. Similar relations hold for $\boldsymbol{G}(\boldsymbol{Y})$ and $\boldsymbol{G}(g)$. Now we would like to study the quantity

$$\|(\boldsymbol{G}(\boldsymbol{X})^{-1} - \boldsymbol{G}(f)^{-1})\|_{op} = \|[(\frac{\alpha}{M}\boldsymbol{X}^\top\boldsymbol{X} + \beta\boldsymbol{I})^{-1} - (\alpha\mathbb{E}[f(\boldsymbol{u})f(\boldsymbol{u})^\top] + \beta\boldsymbol{I})^{-1}]\|_{op} \tag{74}$$

Multiplying the quantity inside the norm by identity and invoking the bound in eq. (72), we have

$$\|(\boldsymbol{G}(\boldsymbol{X})^{-1} - \boldsymbol{G}(f)^{-1})\|_{op} \leq \frac{1}{\beta}\|\boldsymbol{I} - (\frac{\alpha}{M}\boldsymbol{X}^\top\boldsymbol{X} + \beta\boldsymbol{I})(\alpha\mathbb{E}[f(\boldsymbol{u})f(\boldsymbol{u})^\top] + \beta\boldsymbol{I})^{-1}\|_{op} \tag{75}$$

Pulling a factor of $(\alpha\mathbb{E}[f(\boldsymbol{u})f(\boldsymbol{u})^\top] + \beta\boldsymbol{I})^{-1}$ out of everything in the norm on the right hand side:

$$\begin{aligned}\|(\boldsymbol{G}(\boldsymbol{X})^{-1} - \boldsymbol{G}(f)^{-1})\|_{op} \leq \frac{1}{\beta}\|\big[&(\alpha\mathbb{E}[f(\boldsymbol{u})f(\boldsymbol{u})^\top] + \beta\boldsymbol{I}) \\ &- (\frac{\alpha}{M}\boldsymbol{X}^\top\boldsymbol{X} + \beta\boldsymbol{I})\big](\alpha\mathbb{E}[f(\boldsymbol{u})f(\boldsymbol{u})^\top] + \beta\boldsymbol{I})^{-1}\|_{op}\end{aligned}$$

and using the submultiplicative property of the operator norm,

$$\|(\boldsymbol{G}(\boldsymbol{X})^{-1} - \boldsymbol{G}(f)^{-1})\|_{op} \leq \frac{|\alpha|}{\beta}\|\mathbb{E}[f(\boldsymbol{u})f(\boldsymbol{u})^\top] - \frac{1}{M}\boldsymbol{X}^\top\boldsymbol{X}\|_{op}\|(\alpha\mathbb{E}[f(\boldsymbol{u})f(\boldsymbol{u})^\top] + \beta\boldsymbol{I})^{-1}\|_{op}.$$

Using eq. (72) again, we have

$$\|(\boldsymbol{G}(\boldsymbol{X})^{-1} - \boldsymbol{G}(f)^{-1})\|_{op} \leq \frac{|\alpha|}{\beta^2}\|\mathbb{E}[f(\boldsymbol{u})f(\boldsymbol{u})^\top] - \frac{1}{M}\boldsymbol{X}^\top\boldsymbol{X}\|_{op} \tag{76}$$

or

$$\|(\boldsymbol{G}(\boldsymbol{X})^{-1} - \boldsymbol{G}(f)^{-1})\|_{op} \leq \frac{|\alpha|}{\beta^2}\|\boldsymbol{G}(\boldsymbol{X}) - \boldsymbol{G}(f)\|_{op} \tag{77}$$

with an analogous bound holding for $\boldsymbol{G}(\boldsymbol{Y})$ and $\boldsymbol{G}(g)$.

With this knowledge, we define $\hat{\boldsymbol{K}}$ and $\hat{\boldsymbol{H}}$ as the empirical Gram matrices constructed using the estimates $\boldsymbol{G}(\boldsymbol{X})$ and $\boldsymbol{G}(\boldsymbol{Y})$, and conclude that, for every $\{i, j\}$,

$$\hat{\boldsymbol{K}}_{ij} = f(\boldsymbol{u}_i)^\top \boldsymbol{G}(\boldsymbol{X})^{-1} f(\boldsymbol{u}_j) \xrightarrow[M\to\infty]{} f(\boldsymbol{u}_i)^\top \boldsymbol{G}(f)^{-1} f(\boldsymbol{u}_j) = \boldsymbol{K}_{ij} \tag{78}$$

$$\hat{\boldsymbol{H}}_{ij} = g(\boldsymbol{u}_i)^\top \boldsymbol{G}(\boldsymbol{Y})^{-1} g(\boldsymbol{u}_j)^\top \xrightarrow[M\to\infty]{} g(\boldsymbol{u}_i)^\top \boldsymbol{G}(g)^{-1} g(\boldsymbol{u}_j) = \boldsymbol{H}_{ij}. \tag{79}$$

where we recognize the matrices $\hat{\boldsymbol{K}}$ and $\hat{\boldsymbol{H}}$ as $M\boldsymbol{K}_X$ and $M\boldsymbol{K}_Y$ in the main text.

Lastly, we would now like to study

$$\left| \frac{1}{M^3} \operatorname{Tr} \hat{\boldsymbol{K}}\boldsymbol{Q}\hat{\boldsymbol{H}} - \operatorname{Tr} \mathcal{K}\mathcal{Q}\mathcal{H} \right| \tag{80}$$

taking note that the quantity $\frac{1}{M^3} \operatorname{Tr} \hat{\boldsymbol{K}}\boldsymbol{Q}\hat{\boldsymbol{H}}$ is a scaled version of eq. (13) in the main text. The triangle inequality implies a relationship with the trace of the true Gram matrices:

$$\left| \frac{1}{M^3} \operatorname{Tr} \hat{\boldsymbol{K}}\boldsymbol{Q}\hat{\boldsymbol{H}} - \operatorname{Tr} \mathcal{K}\mathcal{Q}\mathcal{H} \right| \leq \left| \frac{1}{M^3} \operatorname{Tr} \hat{\boldsymbol{K}}\boldsymbol{Q}\hat{\boldsymbol{H}} - \frac{1}{M^3} \operatorname{Tr} \boldsymbol{K}\boldsymbol{Q}\boldsymbol{H} \right| + \left| \frac{1}{M^3} \operatorname{Tr} \boldsymbol{K}\boldsymbol{Q}\boldsymbol{H} - \operatorname{Tr} \mathcal{K}\mathcal{Q}\mathcal{H} \right|. \tag{81}$$

The first term on the right hand side can be bounded:

$$\begin{aligned}
\left| \frac{1}{M^3} \operatorname{Tr} \hat{\boldsymbol{K}}\boldsymbol{Q}\hat{\boldsymbol{H}} - \frac{1}{M^3} \operatorname{Tr} \boldsymbol{K}\boldsymbol{Q}\boldsymbol{H} \right| &\leq \frac{1}{M^3} \sum_{ijk} \left| \hat{\boldsymbol{K}}_{ij}\boldsymbol{Q}_{jk}\hat{\boldsymbol{H}}_{ki} - \boldsymbol{K}_{ij}\boldsymbol{Q}_{jk}\boldsymbol{H}_{ki} \right| \\
&= \frac{1}{M^3} \sum_{ijk} \left| \boldsymbol{Q}_{jk} \right| \left| \hat{\boldsymbol{K}}_{ij}\hat{\boldsymbol{H}}_{ki} - \boldsymbol{K}_{ij}\boldsymbol{H}_{ki} \right| \\
&= \frac{1}{M^3} \sum_{ijk} \left| \boldsymbol{Q}_{jk} \right| \left| (\hat{\boldsymbol{K}}_{ij} - \boldsymbol{K}_{ij})\hat{\boldsymbol{H}}_{ki} + \boldsymbol{K}_{ij}(\hat{\boldsymbol{H}}_{ki} - \boldsymbol{H}_{ki}) \right| \\
&= \frac{1}{M^3} \sum_{ijk} \left| \boldsymbol{Q}_{jk} \right| \Big(\left| \hat{\boldsymbol{K}}_{ij} - \boldsymbol{K}_{ij} \right| \left| \hat{\boldsymbol{H}}_{ki} \right| + \left| \boldsymbol{K}_{ij} \right| \left| \hat{\boldsymbol{H}}_{ki} - \boldsymbol{H}_{ki} \right| \Big).
\end{aligned}$$

Using the definitions of $\hat{\boldsymbol{K}}_{ij}$, $\boldsymbol{K}_{ij}$, $\boldsymbol{H}_{ij}$, and $\hat{\boldsymbol{H}}_{ij}$,

$$\begin{aligned}
\left| \frac{1}{M^3} \operatorname{Tr} \hat{\boldsymbol{K}}\boldsymbol{Q}\hat{\boldsymbol{H}} - \frac{1}{M^3} \operatorname{Tr} \boldsymbol{K}\boldsymbol{Q}\boldsymbol{H} \right| \leq\ & \frac{1}{M^3} \|(\boldsymbol{G}(\boldsymbol{X})^{-1} - \boldsymbol{G}(f)^{-1})\|_{op} \sum_{ijk} \left| \boldsymbol{Q}_{jk} \right| \left| \hat{\boldsymbol{H}}_{ki} \right| \|f(\boldsymbol{u}_i)\| \|f(\boldsymbol{u}_j)\| \\
& + \frac{1}{M^3} \|(\boldsymbol{G}(\boldsymbol{Y})^{-1} - \boldsymbol{G}(g)^{-1})\|_{op} \sum_{ijk} \left| \boldsymbol{Q}_{jk} \right| \left| \boldsymbol{K}_{ij} \right| \|g(\boldsymbol{u}_i)\| \|g(\boldsymbol{u}_j)\|.
\end{aligned}$$

Since $\left|\boldsymbol{Q}_{jk}\right|$, $\left|\hat{\boldsymbol{H}}_{ki}\right|$, and $\left|\boldsymbol{K}_{ij}\right|$ are assumed to be finite for all $i, j, k \in \{1, ..., M\}$, and the norms of the neural responses $\|f(\boldsymbol{u}_i)\|$ and $\|g(\boldsymbol{u}_i)\|$ can be assumed to be bounded[4] for all $i \in \{1, ..., M\}$, we can conclude by referencing eq. (77) that the right hand side approaches 0 as $M \to \infty$. This argument can be made precise by considering the rate of convergence of $\hat{\boldsymbol{K}}_{ij} \to \boldsymbol{K}_{ij}$ and $\hat{\boldsymbol{H}}_{ij} \to \boldsymbol{H}_{ij}$ that is inherited from the concentration of $\boldsymbol{G}(\boldsymbol{X}) \to \boldsymbol{G}(f)$ and $\boldsymbol{G}(\boldsymbol{Y}) \to \boldsymbol{G}(g)$.

The second term on the right hand side of eq. (81) is precisely the convergence of a sum over 'true' Gram matrices to the trace of a composition of integral operators described in eq. (67), so this term also approaches 0 as $M \to \infty$.

Putting it all together

[4] Similar to the treatment in [27], this could be interpreted as a resource constraint on the neural responses.

$$\frac{1}{M^3}\,\mathrm{Tr}\,\hat{\boldsymbol{K}}\boldsymbol{Q}\hat{\boldsymbol{H}} \xrightarrow[M\to\infty]{} \mathrm{Tr}[\mathcal{K}\mathcal{Q}\mathcal{H}] = \mathbb{E}_\eta[S]. \tag{82}$$

where we can recognize the plug-in estimate as

$$\frac{1}{M^3}\,\mathrm{Tr}\,\hat{\boldsymbol{K}}\boldsymbol{Q}\hat{\boldsymbol{H}} = \frac{1}{M}\,\mathrm{Tr}\,\boldsymbol{K}_X\boldsymbol{K}_z\boldsymbol{K}_Y = \frac{1}{M}\mathbb{E}\langle \boldsymbol{X}\boldsymbol{w}^*, \boldsymbol{Y}\boldsymbol{v}^*\rangle. \tag{83}$$

in the notation used in the main text (compare with eq. (13)).