Spectral analysis of representational similarity with limited neurons

Understanding how neural representations align between biological and artificial systems has emerged as a central challenge in computational neuroscience. While deep neural networks now reliably predict neural responses across multiple brain areas, their utility for understanding biological computation remains limited by our ability to accurately measure representational similarities. This limitation becomes particularly acute when working with sparse neural recordings, where traditional similarity metrics may fail to capture true representational relationships. Recent spectral analyses of similarity measures provided careful decomposition of neural similarities in the regime of small sample sizes. Here, we consider Centered Kernel Alignment (CKA) as a similarity measure and, using techniques from random matrix theory, identify what spectral aspects affect the representational similarity in the limited neuron regime. We show that the true CKA is underestimated when a small population of neurons is randomly sampled and compared with deterministic neural network representations. We find that increasing sample size may cause underestimating the true CKA. When the number of neurons is small, we demonstrate that only information up to a certain eigenvector threshold can be resolved. We develop a systematic method to denoise the CKA and demonstrate a similarity measure that is robust against changes in population size.

While deep neural networks achieve better neural predictivity [1], recent studies have shown that predictivity measures may suffer from finite size effects and hence require deeper analysis [2, 3]. Here, we analyze Centered Kernel Alignment (CKA) [4] in the context of comparing neural data with limited neuron recordings against deep network representations that are treated as deterministic models. CKA is also a special case of another commonly used method in neuroscience, Representational Similarity Analysis (RSA) [5], when the neural activations are constrained to unit norm. We denote the neural recording matrix as $\tilde{\mathbf{X}} \in \mathbb{R}^{P\times \tilde{N}}$, where P is the number of stimuli and $\tilde{N}$ is the size of the population from which only a subset of $N \ll \tilde{N}$ neurons are assumed to be observed. The model, on the other hand, is denoted by $\tilde{\mathbf{Y}} \in \mathbb{R}^{P\times M}$ activation matrix of M neurons to the same set of stimuli and assumed to be observed in its entirety. The true CKA between the entire neural population and model activations is given by:

$$\mathrm{CKA}(\tilde{\mathbf{\Sigma}}_X, \tilde{\mathbf{\Sigma}}_Y) = \frac{\mathrm{Tr}\left[\tilde{\mathbf{\Sigma}}_X \tilde{\mathbf{\Sigma}}_Y\right]}{\sqrt{\mathrm{Tr}\,\tilde{\mathbf{\Sigma}}_X^2\ \mathrm{Tr}\,\tilde{\mathbf{\Sigma}}_Y^2}}, \quad \tilde{\mathbf{\Sigma}}_X = \tilde{\mathbf{\Phi}}_X \tilde{\mathbf{\Phi}}_X^\top, \quad \tilde{\mathbf{\Sigma}}_Y = \tilde{\mathbf{\Phi}}_Y \tilde{\mathbf{\Phi}}_Y^\top, \tag{1}$$

where $\tilde{\mathbf{\Sigma}}$ denotes the population Gram matrix. In real-life neural recordings, we only have access to a subset of N neurons from this population and hence, we hope to approximate the true CKA from the empirical one defined as $\mathrm{CKA}(\mathbf{\Sigma}_X(N), \tilde{\mathbf{\Sigma}}_Y)$ where $\mathbf{\Sigma}_X(N)$ is the Gram matrix constructed out of $N \ll \tilde{N}$ neurons. As shown in Fig. 1, decreasing the number of neurons N causes the observed CKA to fall below its true value.

Surprisingly, in the case of a power-law-like eigenspectrum—a characteristic consistently observed across many large-scale neural recordings [6]—it is the eigenvectors, rather than the eigenvalues, that drive this drop in CKA. Assuming that the spectrum of the true Gram matrices decay with power law, we find that the empirical CKA can be approximated by the following overlap matrices

$$Q_{ij} = \langle \mathbf{u}_X^i, \tilde{\mathbf{u}}_X^j \rangle^2, \quad M_{ia} = \langle \mathbf{u}_X^i, \tilde{\mathbf{u}}_Y^a \rangle^2, \quad \tilde{M}_{ia} = \langle \tilde{\mathbf{u}}_X^i, \tilde{\mathbf{u}}_Y^a \rangle^2, \tag{2}$$

where $\tilde{\mathbf{u}}_X^i$ and $\tilde{\mathbf{u}}_Y^a$ are the eigenvectors of the true Gram matrices of $\tilde{\mathbf{\Sigma}}_X$ and $\tilde{\mathbf{\Sigma}}_Y$, and $\mathbf{u}_X^i$ are the eigenvectors of the empirical Gram matrix $\mathbf{\Sigma}_X$. Applying a recent study on eigenvector overlap in high dimensions [7], we obtain a formula for the mean behavior of the empirical CKA. In Fig. 1, we compare our theoretical formula to the empirical CKA as a function of increasing N and find perfect agreement. The drop in empirical CKA is attributed to the eigenvectors of sample Gram matrix $\mathbf{\Sigma}_X$ getting more random and hence making two models less aligned. In fact, the alignment worsens even for fixed N when the number of samples increases. We demonstrate this in Fig. 1 where an increasing number of samples hurt alignment.

Next, we analyzed how many eigenvectors can be reliably estimated given the data limitations. In Fig. 2, we show the overlap of sample and true eigenvectors $Q_{ii} = \langle \mathbf{u}_X^i, \tilde{\mathbf{u}}_X^i \rangle^2$ for different N. We found

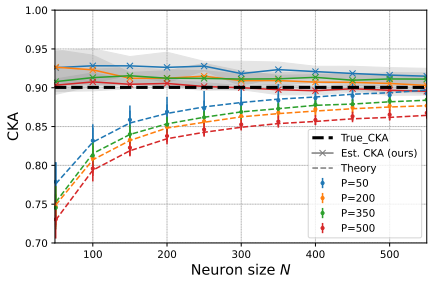

Figure 1: **CKA with finite neurons and samples.** As a test bed, we compare two identical DNN layers trained on CIFAR-10 with different initializations and treat one of them as neural recording. Dashed lines and dots with error bars, resp., indicate theoretical and empirical CKAs. Solid lines and the grey region indicate our estimate of true CKA and its standard deviation. The CKA deviates from its true value both for decreasing N and increasing P.

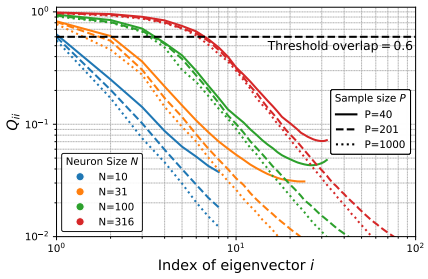

Figure 2: **Overlap of the i-th empirical and true eigenvector.** Theoretical curves for the overlap between empirical vs true eigenvector ($Q_{ii} = \langle \mathbf{u}_X^i, \tilde{\mathbf{u}}_X^i \rangle^2$) for a Gram matrix with power eigenvalues $\lambda_i = i^{-1}$. Our theory can generally inform practitioners how many eigenvectors are resolvable given N. Overlap is $O(1)$ up to some index, then fastly decays. Here the threshold is set to 0.6 as in black horizontal dotted line for comparison.

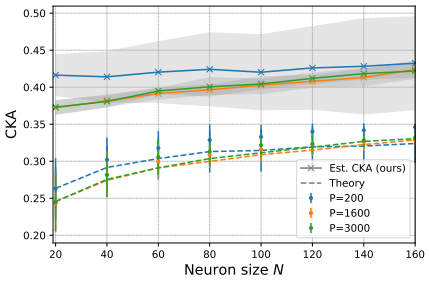

Figure 3: **CKA between macaque IT data and DNNs.** Same experiment as Fig. 1 but with real neural recordings [8] with 3200 samples and 168 neurons compared to the last convolutional layer of an ImageNet trained WideResNet50. The empirical CKA agrees well with the theory. Our true CKA estimate has a larger variance but gives relatively consistent scores across different P and N.

that the first few eigenvectors can be reliably estimated up to a threshold, beyond which the overlap suddenly decays. This informs practitioners about which eigenvectors are meaningfully contributing to the representation similarity with a limited number of neurons.

Since the empirical Gram matrices contain unresolvable (poorly estimated) eigenvectors, we devised a systematic way to denoise the information to estimate the true CKA. We do so by estimating the true overlap $\tilde{\mathbf{M}}$ by regressing from $\mathbf{Q}$ to empirical overlap $\mathbf{M}$. We show that $\mathbf{Q}$ can be calculated from theory [7] and $\tilde{\mathbf{M}}$ can be estimated by minimizing $\|\mathbf{M} - \mathbf{Q}\tilde{\mathbf{M}}\|$. We test our method on ANN activations in Fig.1 and find that the estimated CKA is pretty close to the true CKA even for small numbers of neurons and samples. Finally, we apply our theory to comparing macaque IT data [8] against the last convolutional layer of an ImageNet trained WideResNet50 in Fig. 3. We find that our theory correctly captures the behavior of the empirical CKA with limited neurons. Interestingly, we observe that the mean of our CKA estimate increases with neuron size in the large sample (P) regime, but remains consistent for smaller sample sizes.

In this work, we analyzed a common neural similarity metric, the CKA, when neural data is compared to large artificial models in the neuron-limited regime. Assuming a power-law spectrum, we found that the CKA is dominated by the overlaps between empirical and true eigenvectors, and developed a theory for the empirical CKA based on the spectral properties of the data. We showed that the CKA may produce inaccurate estimates due to overlaps with random eigenvectors, and devised a method to correct for this effect. This work demonstrates that the naive similarity measures should be carefully analyzed in the context of limited neuron and sample sizes, and corrected accordingly to ensure their reliability.

[1] Schrimpf*, Kubilius* et al. 2018. BioRxiv. Brain-score: Which artificial neural network for object recognition is most brain-like?
[2] Canatar*, Feather* et al. 2024. NeurIPS. A spectral theory of neural prediction and alignment.
[3] Pospisil et al. 2024. ICLR. Estimating shape distances on neural representations with limited samples.
[4] Kornblith et al. 2019. ICML. Representational similarity analysis-connecting the branches of systems neuroscience.
[5] Kriegeskorte et al. 2008. Front. in Sys. Neuro. Representational similarity analysis - connecting the branches of systems neuroscience.
[6] Stringer et al. 2019. Nature. High-dimensional geometry of population responses in visual cortex.
[7] Bun et al. 2019. Phys. Rev. E. Overlaps between eigenvectors of correlated random matrices.
[8] Majaj et al. 2015. J. Neurosci. Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance.

[1] 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*, page 407007, 2018.

[2] Abdulkadir Canatar, Jenelle Feather, Albert Wakhloo, and SueYeon Chung. A spectral theory of neural prediction and alignment. *Advances in Neural Information Processing Systems*, 36, 2024.

[3] Dean A Pospisil, Brett W Larsen, Sarah E Harvey, and Alex H Williams. Estimating shape distances on neural representations with limited samples. In *The Twelfth International Conference on Learning Representations*.

[4] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pages 3519–3529. PMLR, 2019.

[5] Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, page 4, 2008.

[6] Carsen Stringer, Marius Pachitariu, Nicholas Steinmetz, Matteo Carandini, and Kenneth D Harris. High-dimensional geometry of population responses in visual cortex. *Nature*, 571(7765):361–365, 2019.

[7] Joël Bun, Jean-Philippe Bouchaud, and Marc Potters. Overlaps between eigenvectors of correlated random matrices. *Physical Review E*, 98(5):052145, 2018.

[8] Najib J Majaj, Ha Hong, Ethan A Solomon, and James J DiCarlo. Simple learned weighted sums of inferior temporal neuronal firing rates accurately predict human core object recognition performance. *Journal of Neuroscience*, 35(39):13402–13418, 2015.