Extended Abstract Track

Generalizable Representation Geometry for Grating Stimuli in Primary Visual Cortex and Artificial Neural Networks

Editors: List of editors' names

Abstract

Humans and other animals display a remarkable ability to generalize learned knowledge to novel domains (out-of-distribution, OOD). This capability is thought to depend on the format of neural population representations, but which geometrical properties support OOD generalization—and which learning objectives give rise to them—remain unclear. We analyze mouse V1 population responses to static grating orientations and show that a decoder trained within a restricted orientation domain can generalize to held-out domains. The quality of generalization correlates with both the dimensionality and curvature of the underlying representation manifold. Notably, similar OOD-generalizable geometry emerges in a deep neural network (PredNet) trained for next-frame prediction on natural videos. These results identify possible geometric properties underpinning OOD generalization and suggest predictive learning as a plausible route to acquire generalizable representational geometry.

Keywords: Out-of-distribution generalization, Predictive learning, mouse V1

1. Extended Abstract

Humans and other animals routinely generalize across domains—for example, adapting driving skills learned at low speed to high-speed highway driving. Such OOD generalization is believed to rely on specific formats of neural representations (Li et al., 2024). While prior work has advanced our understanding for discrete variables (e.g., object categories (Sorscher et al., 2022)), how neural representations support OOD generalization over continuous variables (e.g., orientation or speed) is less understood.

A prevailing hypothesis posits that low-dimensional, low-curvature neural manifolds facilitate OOD generalization (Chung and Abbott, 2021). Empirical hints of low-dimensional and low-curvature structure have appeared in studies of natural-video perception (Hénaff et al., 2019), V1 (Hénaff et al., 2021), and EEG (Sheahan et al., 2021); however, a direct link to OOD generalization has been lacking.

To address this, we studied mouse V1 responses (Stringer et al. (2021), Figure 1A) to static gratings spanning orientations. We partitioned trials into non-overlapping orientation ranges for training and testing, trained a circular ridge decoder on one range, and evaluated on held-out ranges. Decoders generalized robustly, with errors increasing smoothly as the held-out span grew, while remaining below chance across tested ranges (Figure 1B). To probe representational correlates of this OOD generalization, we projected population responses onto the first 200 principal components and fit a smooth orientation manifold via Gaussian process regression (GPR). We quantified dimensionality as the number of PCs required to reach a variance-explained threshold and curvature as the mean angle between adjacent tangents along a 300-point mesh on the fitted manifold. We found that both manifold dimensionality and curvature were correlated with cross-validated OOD error (three 60°

folds; Figure 1D), whereas in-distribution (random-split) errors showed no statistically significant correlation (Figure 1E), suggesting that low-dimensional, low-curvature geometry is a preferable biological neural representation format for OOD generalization.

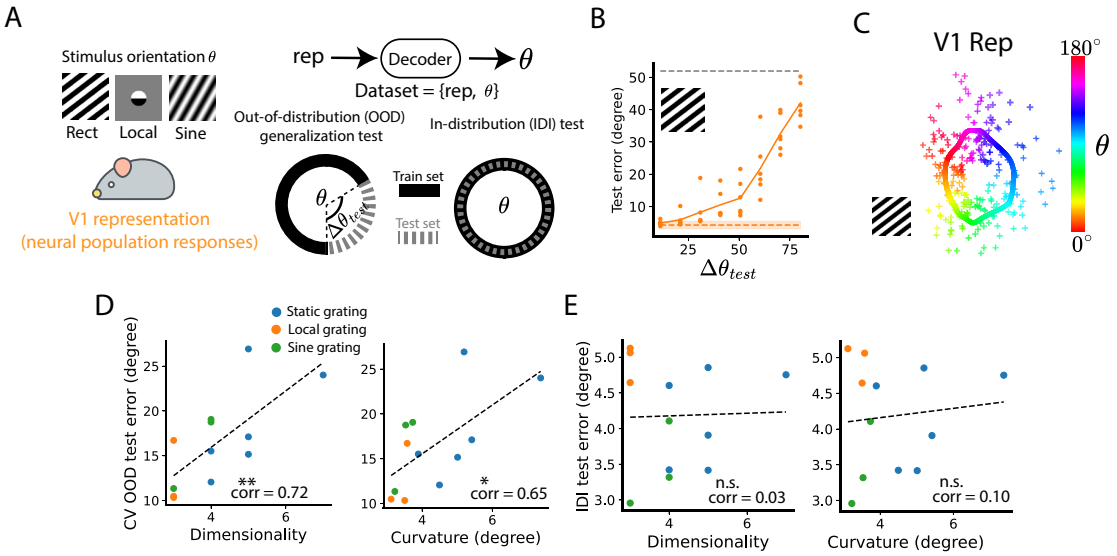

Figure 1: OOD generalization error correlates with the dimensionality and curvature of the representation manifold. (A) Neural responses from mouse V1 to three grating types across orientations were analyzed. Train/test schemes: OOD (non-overlapping orientation ranges) and IDI (random split). A decoder was trained to predict orientation from neural responses. (B) Each dot is one recording session at a given $\Delta\theta_{test}$; lines connect session averages. Orange dashed line: mean IDI error; orange shaded band: 95% confidence interval of the IDI mean. Gray dashed line: chance-level decoder. (C) One example fitted neural manifold (solid) in PCA space; crosses are test data. (D, E) Each dot is one recording session. Corr: Pearson correlation coefficient. $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $n.s.$ not significant; two-sided Pearson correlation test.

What learning objectives engender such geometry? Predictive learning has been shown to yield multiple biologically plausible features (Lotter et al., 2020) and is a powerful pretraining objective. Here we hypothesize that predictive learning promotes OOD generalization.

To test this hypothesis, we trained PredNet, a hierarchical predictive coding–inspired network, on next-frame prediction of natural driving videos from KITTI (Geiger et al. (2013), Figure 2A). After training, we presented sine grating stimuli of varying orientations and collected responses from each layer. Each layer contains on the order of 50,000 units; to make geometry analyses tractable, we projected responses into 50 principal components, which explained nearly all representational variance. Mirroring the V1 analysis, we split the data into three non-overlapping orientation ranges—two folds for training and the left-

out range for testing—and measured decoding error. Cross-validated OOD error decreased with depth in trained PredNet, but not in the untrained model. In contrast, in-distribution error did not account for the OOD trend, suggesting an intrinsic improvement in OOD generalization with depth rather than a trivial consequence of in-distribution performance (Figure 2B). Finally, manifold geometry mirrored this pattern: deeper trained layers exhibited reduced dimensionality and curvature, and these metrics correlated significantly with OOD performance (Figure 2C and D).

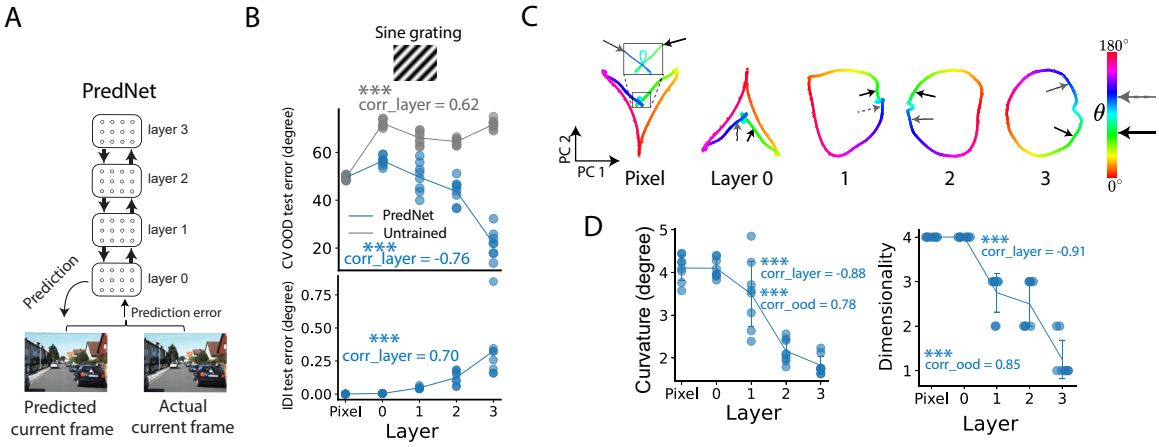

Figure 2: A neural network trained to perform next-frame prediction forms OOD-generalizable representation manifolds. (A) PredNet architecture. (B) Sine grating stimuli with varying orientations were presented to trained and untrained PredNets; unit responses were collected across layers. Blue/gray dots represent individual trained/untrained models. "Corr_layer" denotes the Pearson correlation between layer index and test error. (C) Fitted manifolds for each PredNet layer (projected into PCA space). (D) Dimensionality and curvature across layers. "Corr_ood" denotes the Pearson correlation between layer index and OOD performance.

Together, these results show that low-dimensional, low-curvature geometry in biological/artificial neural representations better support OOD generalization, and suggest predictive learning as a plausible route to acquire generalizable representational geometry.

References

SueYeon Chung and L. F. Abbott. Neural population geometry: An approach for understanding biological and artificial neural networks. *Current Opinion in Neurobiology*, 70:137–144, October 2021. ISSN 0959-4388. doi: 10.1016/j.conb.2021.10.010. URL https://www.sciencedirect.com/science/article/pii/S0959438821001227.

A Geiger, P Lenz, C Stiller, and R Urtasun. Vision meets robotics: The KITTI dataset. *Int. J. Rob. Res.*, 32(11):1231–1237, September 2013. ISSN 0278-3649. doi: 10.1177/0278364913491297. URL https://doi.org/10.1177/0278364913491297.

Olivier J. Hénaff, Robbe L.T. Goris, and Eero P. Simoncelli. Perceptual straightening of natural videos. *Nature Neuroscience*, 22(6):984–991, 2019. ISSN 15461726. doi: 10.1038/s41593-019-0377-4. URL http://dx.doi.org/10.1038/s41593-019-0377-4. Publisher: Springer US.

Olivier J. Hénaff, Yoon Bai, Julie A. Charlton, Ian Nauhaus, Eero P. Simoncelli, and Robbe L.T. Goris. Primary visual cortex straightens natural video trajectories. *Nature Communications*, 12(1), 2021. ISSN 20411723. doi: 10.1038/s41467-021-25939-z. URL http://dx.doi.org/10.1038/s41467-021-25939-z. Publisher: Springer US.

Qianyi Li, Ben Sorscher, and Haim Sompolinsky. Representations and generalization in artificial and brain neural networks. *Proceedings of the National Academy of Sciences*, 121 (27):e2311805121, July 2024. doi: 10.1073/pnas.2311805121. URL https://www.pnas.org/doi/10.1073/pnas.2311805121. Publisher: Proceedings of the National Academy of Sciences.

William Lotter, Gabriel Kreiman, and David Cox. A neural network trained for prediction mimics diverse features of biological neurons and perception. *Nature Machine Intelligence*, 2(4):210–219, April 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-0170-9. URL https://www.nature.com/articles/s42256-020-0170-9. Publisher: Nature Publishing Group.

Hannah Sheahan, Fabrice Luyckx, Stephanie Nelli, Clemens Teupe, and Christopher Summerfield. Neural state space alignment for magnitude generalization in humans and recurrent networks. *Neuron*, 109(7):1214–1226.e8, 2021. ISSN 10974199. doi: 10.1016/j.neuron.2021.02.004. URL https://doi.org/10.1016/j.neuron.2021.02.004. Publisher: Elsevier Inc.

Ben Sorscher, Surya Ganguli, and Haim Sompolinsky. Neural representational geometry underlies few-shot concept learning. *Proceedings of the National Academy of Sciences of the United States of America*, 119(43):1–12, 2022. ISSN 10916490. doi: 10.1073/pnas.2200800119.

Carsen Stringer, Michalis Michaelos, Dmitri Tsyboulski, Sarah E. Lindo, and Marius Pachitariu. High-precision coding in visual cortex. *Cell*, 184(10):2767–2778.e15, May 2021. ISSN 10974172. doi: 10.1016/j.cell.2021.03.042. URL https://doi.org/10.1016/j.cell.2021.03.042. Publisher: Elsevier Inc.