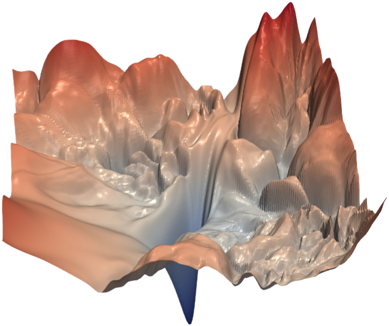
Picture of the space of typical learnable tasks

Rahul Ramesh[†], Jialin Mao[†], Itay Griniasty[§],
Han Kheng Teoh[§], Rubing Yang[†],
Mark K. Transtrum[¶], James P. Sethna[§], Pratik Chaudhari[†]

[†]University of Pennsylvania, [§]Cornell University, [¶]Brigham Young University

Motivation

Motivation



Why are neural networks able to find representations that capture the shared structure in data?

Prediction Space

We analyze the training trajectories of neural networks in prediction space.

Prediction Space

We analyze the training trajectories of neural networks in prediction space.

Consider a neural network with weights w and inputs $\{x_i\}_{i=1}^N$. The predictions

$$P_w = \begin{pmatrix} p_w(y=1 \mid x_1) & p_w(y=2 \mid x_1) & \cdots & p_w(y=C \mid x_1) \\ p_w(y=1 \mid x_2) & p_w(y=2 \mid x_2) & \cdots & p_w(y=C \mid x_2) \\ \vdots & \vdots & \vdots & \vdots \\ p_w(y=1 \mid x_N) & p_w(y=2 \mid x_N) & \cdots & p_w(y=C \mid x_N) \end{pmatrix}$$

is an $N \times C$ dimensional object.

Trajectories in Prediction Space

We convert training trajectories in weight space

$$(w_1, w_2, \cdots, w_T)$$

into trajectories in prediction space

$$(P_{w_1}, P_{w_2}, \cdots, P_{w_T}).$$

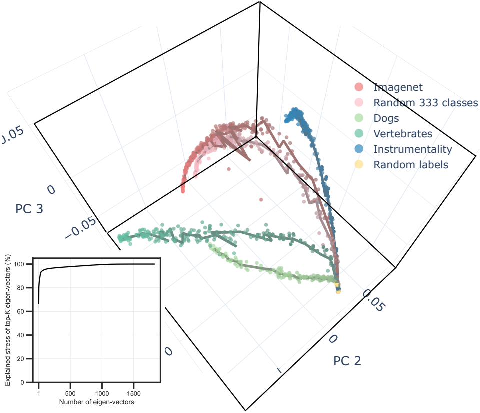
Trajectories in Prediction Space

We convert training trajectories in weight space

$$(w_1, w_2, \cdots, w_T)$$

into trajectories in prediction space

$$(P_{w_1}, P_{w_2}, \cdots, P_{w_T}).$$



InPCA reveals that the training trajectories are effectively low-dimensional in prediction space.

Information Geometry

We use tools from information geometry to study the prediction space.

Probabilistic model P_w is a point on a statistical manifold with the metric being the Fisher information metric.

Information Geometry

We use tools from information geometry to study the prediction space.

Probabilistic model P_w is a point on a statistical manifold with the metric being the Fisher information metric.

We consider

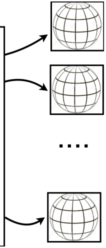
$$\sqrt{P_w} = \begin{pmatrix} \sqrt{p_w(y=1 \mid x_1)} & \sqrt{p_w(y=2 \mid x_1)} & \cdots & \sqrt{p_w(y=C \mid x_1)} \\ \sqrt{p_w(y=1 \mid x_2)} & \sqrt{p_w(y=2 \mid x_2)} & \cdots & \sqrt{p_w(y=C \mid x_2)} \\ \vdots & \vdots & \vdots & \vdots \\ \sqrt{p_w(y=1 \mid x_i)} & \sqrt{p_w(y=2 \mid x_i)} & \cdots & \sqrt{p_w(y=C \mid x_i)} \\ \vdots & \vdots & \vdots & \vdots \\ \sqrt{p_w(y=1 \mid x_N)} & \sqrt{p_w(y=2 \mid x_N)} & \cdots & \sqrt{p_w(y=C \mid x_N)} \end{pmatrix},$$

and note that L2 norm of each row 1.

Information Geometry

We think of $\sqrt{P_w}$ as a point on the product of N spheres.

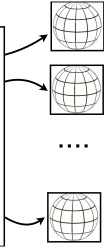
$$\begin{pmatrix} \sqrt{p_w(y=1 \mid x_1)} & \sqrt{p_w(y=2 \mid x_1)} & \cdots & \sqrt{p_w(y=C \mid x_1)} \\ \sqrt{p_w(y=1 \mid x_2)} & \sqrt{p_w(y=2 \mid x_2)} & \cdots & \sqrt{p_w(y=C \mid x_2)} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \sqrt{p_w(y=1 \mid x_N)} & \sqrt{p_w(y=2 \mid x_N)} & \cdots & \sqrt{p_w(y=C \mid x_N)} \end{pmatrix}$$



Information Geometry

We think of $\sqrt{P_w}$ as a point on the product of N spheres.

$$\begin{pmatrix} \sqrt{p_w(y=1 \mid x_1)} & \sqrt{p_w(y=2 \mid x_1)} & \cdots & \sqrt{p_w(y=C \mid x_1)} \\ \sqrt{p_w(y=1 \mid x_2)} & \sqrt{p_w(y=2 \mid x_2)} & \cdots & \sqrt{p_w(y=C \mid x_2)} \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots \\ \sqrt{p_w(y=1 \mid x_N)} & \sqrt{p_w(y=2 \mid x_N)} & \cdots & \sqrt{p_w(y=C \mid x_N)} \end{pmatrix}$$

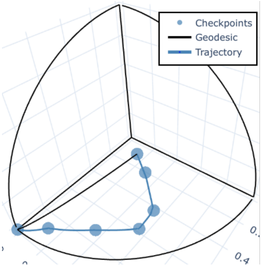


The geodesic under the FIM is exactly the great circle distance, i.e.,

$$\sqrt{P_{u,v}^{\lambda}} = \frac{\sin((1-\lambda)d_G)}{\sin(d_G)}\sqrt{P_u} + \frac{\sin(\lambda d_G)}{\sin(d_G)}\sqrt{P_v}, \qquad \lambda \in [0,1].$$

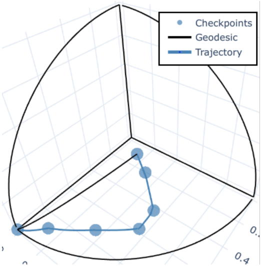
Computational Info. Geometry

Equipped with these ideas, we study the trajectories with the following tools:



Computational Info. Geometry

Equipped with these ideas, we study the trajectories with the following tools:



Geometric progress

$$t_w = \inf_{\lambda \in [0,1]} d_G(P_w, P_{0,*}^{\lambda})$$

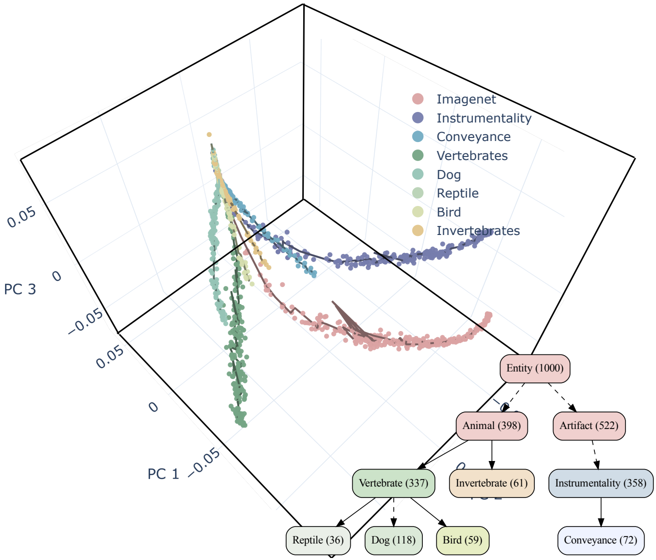
Riemann Length

$$L = 2\int_0^1 \sqrt{d_B(P_{w(t)}, P_{w(t+\mathrm{d}t)})}$$

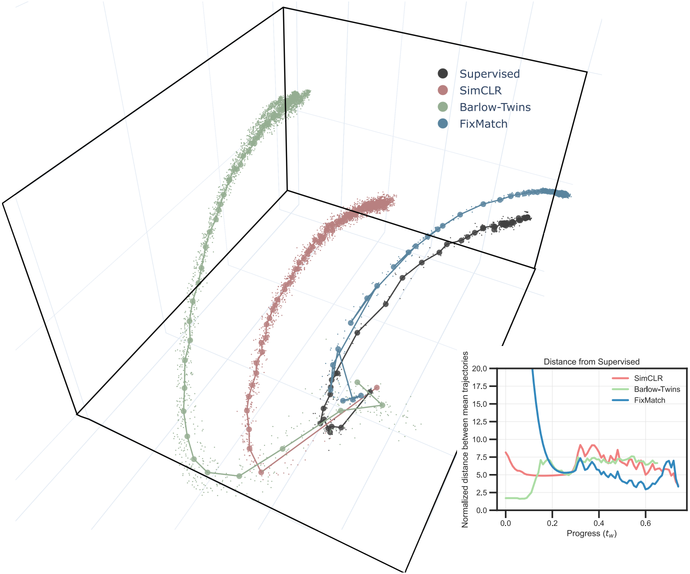
Comparing curves

$$d_{\text{traj}}(\tau_u, \tau_v) = \int_0^1 d_B(P_{u(t)}, P_{v(t)})\mathrm{d}t$$

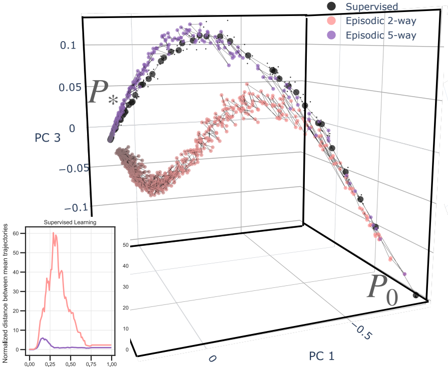
Results - Training on different tasks

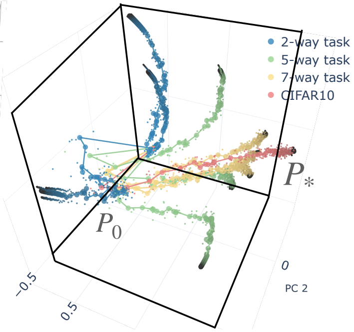


Results - Self-supervised learning



Results - Episodic meta-learning





Conclusion

We develop tools to study the training trajectories of representations in prediction space.

Conclusion

We develop tools to study the training trajectories of representations in prediction space.

These trajectories are effectively low-dimensional in prediction space, but we don’t fully understand why.

Conclusion

We develop tools to study the training trajectories of representations in prediction space.

These trajectories are effectively low-dimensional in prediction space, but we don't fully understand why.

ArXiv:2210.17011

github.com/grasp-lyrl/picture_of_space_of_tasks