A Picture of the Space of Typical Learnable Tasks

Rahul Ramesh [1] **Jialin Mao** [1] **Itay Griniasty** [2] **Rubing Yang** [1] **Han Kheng Teoh** [2]
Mark K. Transtrum [3] **James P. Sethna** [2] **Pratik Chaudhari** [1]

Abstract

We develop information geometric techniques to understand the representations learned by deep networks when they are trained on different tasks using supervised, meta-, semi-supervised and contrastive learning. We shed light on the following phenomena that relate to the structure of the space of tasks: (1) the manifold of probabilistic models trained on different tasks using different representation learning methods is effectively low-dimensional; (2) supervised learning on one task results in a surprising amount of progress even on seemingly dissimilar tasks; progress on other tasks is larger if the training task has diverse classes; (3) the structure of the space of tasks indicated by our analysis is consistent with parts of the Wordnet phylogenetic tree; (4) episodic meta-learning algorithms and supervised learning traverse different trajectories during training but they fit similar models eventually; (5) contrastive and semi-supervised learning methods traverse trajectories similar to those of supervised learning. We use classification tasks constructed from the CIFAR-10 and Imagenet datasets to study these phenomena. Code is available at https://github.com/grasp-lyrl/picture_of_space_of_tasks.

1. Introduction

Exploiting data from related tasks to reduce the sample complexity of learning a desired task, is an idea that lies at the heart of burgeoning fields like transfer, multi-task, meta, few-shot, semi- and self-supervised learning. These algorithms, aided by the ability of deep networks to learn flexible representations, have shown an impressive ability to predict well on new tasks. These algorithms are very different from each other but it stands to reason they must be exploiting some structure in the space of tasks. We do not yet know what the structure in the space of tasks is precisely (see §4 for a discussion of related work). The goal of this paper is to characterize this structure and shed light on *why* these existing algorithms are successful.

We develop information-geometric techniques to analyze representations learned by different algorithms on different tasks. The key idea of this paper is to think of a deep network with weights w trained on a task as a probabilistic model

$$P_w(\vec{y}) = \prod_{n=1}^{N} p_w^n(y_n)$$

where $\vec{y} = (y_1, \ldots, y_N)$ denotes any sequence of outputs (each $y_n \in \{1, \ldots, C\}$ classes) on N independent and identically distributed samples and $p_w^n(y_n)$ denotes the probability that sample x_n belongs to class y_n as predicted by a deep network with weights w. We instantiate the technical machinery of information geometry using this NC-dimensional object to study different probabilistic models fitted to the task irrespective of which representation learning algorithm, e.g., supervised learning, meta-learning, etc., or what neural architecture was used to fit the probabilistic model. This construction circumvents the enormous diversity of algorithms, architectures with different feature spaces and training methods across these different sub-fields and provides us with a single space to study these models in — the prediction space of the model.

1.1. Contributions

We develop new theoretical and computational tools to study such probabilistic models in §2. This involves techniques to visualize these very high-dimensional objects, to compute geodesics on such manifolds, to interpolate checkpoints along training trajectories into continuous curves, and to map models trained on different tasks into a unique prediction space. We point these technical tools to understanding the structure of the space of learnable tasks and study different representation algorithms such as supervised, transfer, meta, semi- and self-supervised learning. We report the following findings in §3; a result pertaining to contrastive learning its provided in Appendix B.

(1) The manifold of probabilistic models trained on different tasks using different representation learning methods is effectively low-dimensional. This dimensional-

[1]University of Pennsylvania [2]Cornell University [3]Brigham Young University. Correspondence to: Rahul Ramesh <rahulram@seas.upenn.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

ity is very small: for Imagenet where our probabilistic models are in 10^7 dimensions, the top 3 dimensions preserve 80.02% of the pairwise distances between 2430 models trained on different sub-tasks of Imagenet.

(2) Supervised learning on one task results in a surprising amount of progress on seemingly dissimilar tasks (informally, "progress" means that the representation learned on one can be used to make accurate predictions on other tasks; this is defined in (4)); progress on other tasks is larger if the training task has diverse classes.

(3) Structure of the space of tasks indicated by our analysis is consistent with parts of the Wordnet phylogenetic tree.

(4) Episodic meta-learning algorithms and supervised learning traverse different trajectories in the space of probabilistic models during training but learn similar models eventually; the trajectory of episodic meta-learning for a small "way" is about $40\times$ longer in terms of its Riemann length than that of supervised learning.

(5) Contrastive and semi-supervised learning methods traverse similar trajectories to that of supervised learning in the space of probabilistic models.

(6) Fine-tuning a model upon a sub-task does not change the representation much if the model was trained for a large number of epochs.

We present evidence and analysis of these findings using multiple neural architectures and a large number of different image-classification tasks created from the CIFAR-10 and Imagenet datasets.

2. Methods

Modeling the task We define a task P as a joint distribution on inputs $x \in \mathbb{R}^d$ and outputs $y \in \{1, \ldots, C\}$ corresponding to C classes. Suppose we have N independent and identically distributed samples $\{(x_n, y_n^*)\}_{n=1}^N$ from P. Let $\vec{y} = (y_1, \ldots, y_N)$ denote any sequence of outputs on these N samples and $\vec{y^*}$ denote the sequence of ground-truth labels. We can model the task as

$$P_w(\vec{y}) = \prod_{n=1}^{N} p_w^n(y_n) \tag{1}$$

where w are the parameters of the model and we have used the shorthand $p_w^n(y_n) \equiv p_w(y_n \mid x_n)$. Let "truth" or $P_* \equiv P(\vec{y^*})$ denote the true probability distribution which corresponds to the ground-truth labels. Let "ignorance" or P_0 denote the probability distribution that corresponds to $p^n(y) = 1/C$ for all n and all $y \in \{1, \ldots, C\}$.

Bhattacharyya distance Given two models P_u and P_v parameterized by weights u and v respectively, the Bhattacharyya distance (Bhattacharyya, 1946) between them averaged over samples can be written as (see Appendix C)

$$\begin{aligned} \mathrm{d_B}(P_u, P_v) &:= -N^{-1} \log \sum_{\vec{y}} \prod_n \sqrt{p_u(y_n)\, p_v(y_n)} \\ &= -N^{-1} \sum_n \log \sum_c \sqrt{p_u^n(c)\, p_v^n(c)}. \end{aligned} \tag{2}$$

Our model (1) involves a product over the probabilities of N samples. Many distances, e.g., the Hellinger distance $2\left(1 - \prod_n \sum_c \sqrt{p_u^n(c)\, p_v^n(c)}\right)$, saturate for large N, this is because random high-dimensional vectors are nearly orthogonal. This makes it difficult to use such distances to understand high-dimensional probabilistic models. Bhattacharyya distance is well-behaved for large N due to the logarithm (Quinn et al., 2019; Teoh et al., 2020), and that is why it is well suited to our problem.

Remark 1 (Models with different intermediate representations can have zero Bhattacharyya distance). Two models can have different internal representations and yet define identical probabilistic models. For example, a representation and a rotated version of the same representation can define identical probabilistic models if this rotation is undone before the output. The Bhattacharyya distance (2) only depends on the output probabilities and would be zero if the probabilistic models are identical. Focusing the theory on the probabilistic model that makes the predictions as opposed to the feature space therefore allows us to capture many symmetries in the prediction space.

Distances between trajectories of probabilistic models Consider a trajectory $(w(k))_{k=0,\ldots,T}$ that records the weights after T updates of the optimization algorithm, e.g., stochastic gradient descent. This trajectory corresponds to a trajectory of probabilistic models $\tilde{\tau}_w = (P_{w(k)})_{k=0,\ldots,T}$. We are interested in calculating distances between such training trajectories. First, consider $\tilde{\tau}_u = (u(0), u(1), u(2), \ldots, u(T))$ and another trajectory $\tilde{\tau}_v \equiv (u(0), u(2), u(4), \ldots, u(T), u(T), \ldots, u(T))$ which trains twice as fast but to the same end point. If we define the distance between these trajectories as, say, $\sum_k \mathrm{d_B}(P_{u(k)}, P_{v(k)})$, then the distance between $\tilde{\tau}_u$ and $\tilde{\tau}_v$ will be non-zero—even if they are fundamentally the same. This issue is more pronounced when we calculate distances between training trajectories of different tasks. It arises because we are recording each trajectory using a different time coordinate, namely its own training progress.

To compare two trajectories correctly, we need a notion of time that can allow us to uniquely index any trajectory. The geodesic between the start point P_0 and the true distribution P_* is a natural candidate for this purpose since it is unique. Geodesics are locally length-minimizing curves in a metric space. For the product manifold in (1), we can obtain a closed-form formula for the geodesic by noticing that for each sample, the vector $\left(\sqrt{p_u^n(c)}\right)_{c=1,\ldots,C}$ lies

on a C-dimensional unit sphere. The geodesic connecting two models P_u and P_v under the Fisher information metric which is induced by the Bhattacharyya distance is just the great circle on the sphere (Ito & Dechant, 2020, Eq. 47):

$$\sqrt{P^{\lambda}_{u,v}} = \prod_{n=1}^{N} \left(\frac{\sin\left((1-\lambda)\mathrm{d}^n_{\mathrm{G}}\right)}{\sin\left(\mathrm{d}^n_{\mathrm{G}}\right)} \sqrt{p^n_u} + \frac{\sin\left(\lambda \mathrm{d}^n_{\mathrm{G}}\right)}{\sin\left(\mathrm{d}^n_{\mathrm{G}}\right)} \sqrt{p^n_v} \right), \quad (3)$$

where $\lambda \in [0,1]$ and $\mathrm{d}^n_{\mathrm{G}} = \cos^{-1}\left(\sum_c \sqrt{p^n_u(c)}\sqrt{p^n_v(c)}\right)$ is one half of the great-circle distance between $p^n_u(\cdot)$ and $p^n_v(\cdot)$. Any probabilistic model P_w on a trajectory $\tilde{\tau}_w$ can now be **re-indexed by a new "time" that we call "progress"**:

$$[0,1] \ni t_w = \arg \inf_{\lambda \in [0,1]} \mathrm{d}_{\mathrm{G}}(P_w, P^{\lambda}_{0,*}). \quad (4)$$

It indicates the distance of P_w to the truth P_* measured in terms of the closest point $P^{t_w}_{0,*}$ on the geodesic to P_w. We solve (4) using bisection search (Brent, 1971). Observe that using the same expression as (3), we can also interpolate between two successive recorded points $P_{w(k)}$ and $P_{w(k+1)}$ of a trajectory by calculating $P^{\lambda}_{w(k),w(k+1)}$ for different values of $\lambda \in [0,1]$. This is useful because different networks train with very different speeds on different tasks, especially in early stages of training. This allows us to effectively convert a sequence of models $\tilde{\tau}_w = (P_{w(k)})_{k=0,\ldots,T}$ into a continuous curve $\tau_w = (P_{w(t)})_{t\in[0,1]}$. We calculate the distance between continuous curves τ_u, τ_v as

$$\mathrm{d}_{\mathrm{traj}}(\tau_u, \tau_v) = \int_0^1 \mathrm{d}_{\mathrm{B}}(P_{u(t)}, P_{v(t)})\, \mathrm{d}t\,; \quad (5)$$

which is approximated using a uniform grid on $[0,1]$.

Riemann length of a trajectory Divergences like the Bhattacharyya distance or the Kullback-Leibler (KL) divergence (which is the cross-entropy loss up to a constant) can be used to define a Riemannian structure in the space of probabilistic models (Amari, 2016). The distance between two infinitesimally different models P_w and $P_{w+\mathrm{d}w}$ is

$$\mathrm{d}s^2 = 4\mathrm{d}_{\mathrm{B}}(P_w, P_{w+\mathrm{d}w}) = \langle \mathrm{d}w\,, g(w)\,\mathrm{d}w\rangle + \mathcal{O}(\|\mathrm{d}w\|^2),$$

where $g(w) = N^{-1}\sum_{\vec{y}}(P_w)^{-1}\partial^2 P_w$ is the Fisher Information Matrix (Quinn, 2019, Section A.3). This Fisher Information Matrix (FIM) is therefore the metric of the space of the probability distributions and weights w play the role of the coordinates in this space. Up to a scalar factor, the Bhattacharyya distance and the KL-divergence induce the same FIM. The Riemann length of a trajectory τ_w is the integral of these infinitesimal lengths:

$$\mathrm{Length}(\tau_w) = 2\int_0^1 \sqrt{\mathrm{d}_{\mathrm{B}}(P_{w(t)}, P_{w(t+\mathrm{d}t)})}; \quad (6)$$

it is equal to the integral of FIM-weighted incremental distance traveled in the weight space. Observe that we do not need the FIM to calculate the length. We can think of the length of a trajectory taken by a model to reach the solution P_* compared to the length of the geodesic as a measure of the inefficiency of the training procedure since the geodesic is the curve with the shortest length. This inefficiency can arise because: (a) not all probability distributions along the geodesic can be parametrized by our model class (approximation error), and (b) the training process may take steps that are misaligned with the geodesic (e.g., due to the loss function, mini-batch updates, supervised vs. some other form of representation learning, etc.).

Mapping a model trained on one task to another task using "imprinting" In this paper, we will consider different tasks $\{P^k\}_{k=1,\ldots,}$ with the same input domain but possibly different number of classes C^k. Given a model P^1_w parametrized by weights w for task P^1, we are interested in evaluating its learned representation on another task, say, P^2. Let $w = (w_1, w_2)$ be the weights for the backbone and the classifier respectively. The logits are $\mathbb{R}^{C^1} \ni {w_2}^{\top}\varphi(x; w_1)$ corresponding to an input x and features of the penultimate layer $\varphi(x; w_1)$. The network's output $p_w(c \mid x_n)$ for $c = 1, \ldots, C^1$ is computed using a softmax applied to the logits. If we have learned w from one task P^1, then we can re-initialize each row of the classifier weights $(w_2)'_c$ for $c = 1, \ldots, C^2$ to maximize the cosine similarity with the average feature of samples from task P^2 with ground-truth class c:

$$(w_2)'_c = h/\|h\|_2 \quad \text{where } h = \textstyle\sum_{\{x: y^*_x = c\}} \varphi(x; w_1). \quad (7)$$

The new network $w = (w_1, {w_2}')$ can be used to make predictions on P^2. Using imprinting, we can map a trajectory τ^1_w of a network being trained on P^1 to another task P^2 by mapping each point along the trajectory; let us denote this mapped trajectory by $\tau^{1\to 2}_w$.

Remark 2 (Imprinting versus training the final layer or probing). There are many ways of performing such a mapping, e.g., one could fine-tune the weights using data from P^2, linear probing (Shi et al., 2016), etc. The technique described above is known as "imprinting" (Hu et al., 2015; Qi et al., 2018; Dhillon et al., 2020). In this paper, we will be mapping thousands of models across different trajectories to other tasks. Training the final layer, or a new classifier, for all these models is cumbersome and imprinting provides a simple way around this issue. Note that imprinting is not equivalent to training the classifier w_2 (with backbone w_1 fixed) using samples from the other task but we found that imprinted weights work well in practice (see Appendix D).

How to choose an appropriate task to map different models to? Consider the training trajectory τ^1_u of a model being trained on P^1 and another trajectory τ^2_v of a model being trained on P^2. Using (7), we can map these trajectories to the other task to get $\tau^{1\to 2}_u$ and $\tau^{2\to 1}_v$. This allows us

to calculate, for instance, $\mathrm{d}_{\text{traj}}(\tau_u^{1\to 2}, \tau_v^2)$ using (5) which is the distance of the trajectory of the model trained on P^1 and then mapped to P^2 with respect to the trajectory of a model trained on task P^2. If the two learning tasks P^1 and P^2 are very different, (e.g., Animals in CIFAR-10 and Vehicles in CIFAR-10), then this distance will be large.

Quantities like $\mathrm{d}_{\text{traj}}(\tau_u^{1\to 2}, \tau_v^2)$ or $\mathrm{d}_{\text{traj}}(\tau_v^{2\to 1}, \tau_u^1)$ are reasonable candidates to study similarities between tasks P^1 and P^2, but they are not equal to one another. We are also interested in doing such calculations with models trained on many different tasks, and mapping them to each other will lead to an explosion of quantities. To circumvent this, we map to a unique task whose output space is the union of the output spaces of the individual tasks, e.g., to study P^1 (Animals) and P^2 (Vehicles), we will map both trajectories to P^{U} which is all of CIFAR-10. We will use

$$\mathrm{d}_{\text{traj}}(\tau_u^{1\to \mathrm{U}}, \tau_v^{2\to \mathrm{U}}) \tag{8}$$

as the distance between trajectories trained on P^1 and P^2.

Visualizing a high-dimensional probabilistic model in lower-dimensions We use a visualization technique called intensive principal component analysis (InPCA) (Quinn et al., 2019) that embeds a probabilistic model into a lower-dimensional space. For m probability distributions, consider a matrix $D \in \mathbb{R}^{m\times m}$ with entries $D_{uv} = \mathrm{d}_{\mathrm{B}}(P_u, P_v)$ and

$$W = -LDL/2 \tag{9}$$

where $L_{ij} = \delta_{ij} - 1/m$ is the centering matrix. An eigen-decomposition of $W = U\Sigma U^\top$ where the eigenvalues are sorted in descending order of their magnitudes $|\Sigma_{00}| \geq |\Sigma_{11}| \geq \ldots$ allows us to compute the embedding of these m probability distributions into an m-dimensional space as $\mathbb{R}^{m\times m} \ni X = U\sqrt{\Sigma}$. Unlike standard PCA where eigenvalues are non-negative, eigenvalues of InPCA can be both positive and negative, i.e., the lower-dimensional space is a Minkowski space (Quinn et al., 2019). This allows the InPCA embedding to be an isometry, i.e., pairwise distances are preserved:

$$\textstyle\sum_{i=1}^m (X_u^i - X_v^i)^2 = \mathrm{d}_{\mathrm{B}}(P_u, P_v) \geq 0 \tag{10}$$

for embeddings X_u, X_v of two distributions P_u, P_v. We can measure how well pairwise distances are preserved by a k-dimensional sub-space using the "explained stress" χ_k (Cox & Cox, 2008):

$$\chi_k = 1 - \frac{\|W - \sum_{i=1}^k \Sigma_{ii}\, U_i U_i^\top\|_{\mathrm{F}}}{\|W\|_{\mathrm{F}}} = 1 - \sqrt{\frac{\sum_{i=k+1}^m \Sigma_{ii}^2}{\sum_{i=1}^m \Sigma_{ii}^2}}. \tag{11}$$

Just like standard PCA, if we preserve all the eigenvectors (i.e., $k = m$), then (10) holds exactly and $\chi_k = 1$. But if we use fewer eigenvectors then pairwise distances can be distorted. See Appendix F for more details of the explained stress. Appendix A.2 describes how we implement InPCA for high-dimensional probabilistic models.

3. Results

We next describe our findings using the theoretical ideas developed in the previous section. We present a broad range of evidence and analysis using a large number of representation learning techniques, multiple neural architectures and a large number of different image-classification tasks created from the CIFAR-10 and ImageNet datasets. Experiments in this paper required about 30,000 GPU-hours. Appendix A describes the setup for these experiments in detail. See Appendix H for a discussion of some frequently asked questions. One more result, Result 7: Contrastive learning methods trained on different datasets learn similar representations, is presented in Appendix B.

Remark 3. All the analysis in this paper (except Figures 4 and 5) was conducted using the test data. All models were trained using the training data, but all mapped models, distances between trajectories, quantitative evaluation of progress and InPCA embeddings were computed using the test dataset. The reason for this is that we would like to study the geometry of tasks as evidenced by samples that were not a part of training. To emphasize, we do not develop any new algorithms for learning in this paper. Therefore using the test data to quantify relationships between tasks is reasonable; see similar motivations in Kaplun et al. (2022) or Ilyas et al. (2022) among others. Our findings remain valid when training data is used for analysis; this is because in most of our experiments, a representation is trained on one task but makes predictions on a completely new task after mapping.

Result 1: The manifold of models trained on different tasks, and using different representation learning methods, is effectively low-dimensional We trained multiple models on 6 different sub-tasks of ImageNet (from 5 random initializations each) to study the dimensionality of the manifold of probabilistic models along the training trajectories (100 points equidistant in progress (4)) after mapping all models to all ImageNet classes ($\sim 10^8$ dimensions). We use the explained stress (defined in Appendix F), to measure if the distances are preserved by the first k dimensions of the embedding of the models. The first 3 dimensions of InPCA (Figure 1a) preserve 80.02% of the explained stress (Figure 1b shows more dimensions). This is therefore a remarkably low-dimensional manifold. It is not exactly low-dimensional because the explained stress is not 100%, but it is an effectively low-dimensional manifold. This also indicates that the individual manifolds of models trained on one task are low-dimensional, even if they start from different random initializations in the weight space. Such low-dimensional manifolds are seen in *all* our experiments, irrespective of the specific method used for representation

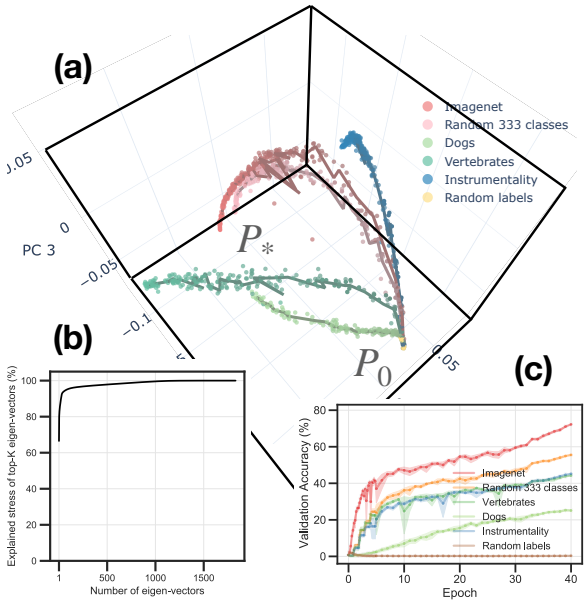


Figure 1. **(a)** Visualization of training trajectories of models trained on 6 tasks from ImageNet. Each point is one network, bold lines connect points along the average trajectory of each task (across 5 random weight initializations). Trajectories move towards the truth P_*, which corresponds to the ground-truth labels. Training on one task makes a remarkable amount of progress on unseen, seemingly dissimilar, classes. Trajectories of models trained on a random set of 333 classes are similar to those of the entire ImageNet. Some classes (Instrumentality) are closer to this trajectory while others such as Vertebrates and Dogs are farther away. Dogs is a semantic subset of Vertebrates; it splits at the beginning but seems to eventually reach a similar representation as one of the intermediate points of Vertebrates.

(b) Percentage explained stress (11) captured by subspace spanned by the top k InPCA eigenvectors.

(c) Validation accuracy on different tasks vs. epochs.

learning, namely, supervised, transfer (fine-tuning), meta, semi-supervised and contrastive learning.

Remark 4 (A detailed description of how we plot trajectories of representations). We provide a non-mathematical description of how the theory in §2 was used to draw Figure 1a below. We train 5 different networks (random seeds for initialization) for each of the 6 tasks, and record 61 model checkpoints during training; this gives 1830 checkpoints for this experiment. We re-index all checkpoints to calculate their progress using Equations (3) and (4). We then interpolate between each consecutive pair of the 61 checkpoints along each trajectory using (3). The training trajectory can now be sampled at any progress $t_w \in [0, 1]$. We next calculate the "average trajectory" of the 5 networks (random seeds) of each task by averaging the output probabilities in (1) at a fixed value of t_w; 100 different values of t_w spread uniformly between $[0, 1]$ are chosen. These 100

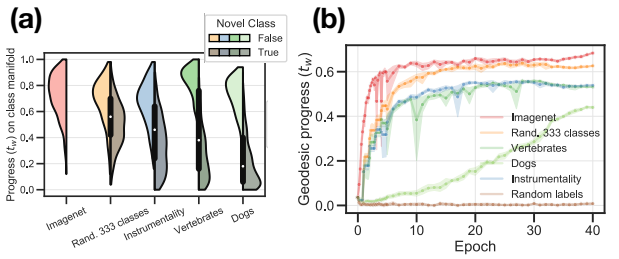


Figure 2. **(a)** Progress made by each model on classes seen during training (left half, lighter shade) and on novel classes (right half, darker shade). We compute t_w^c which is the progress t_w of images restricted to a single class c. This quantity t_w^c measures the quality of the representation for class c. Violin plots denote the distribution of t_w^c indicate that we make more progress on classes seen during training. If the model sees a larger diversity of classes (like with random 333 classes), more progress is made on the novel classes. Surprisingly, even if we train on just the "Dogs", we make some progress on novel classes.

(b) Progress t_w (4) on the Y-axis against the number of epochs of training on the X-axis. The progress t_w increases with more epochs of training—all models make non-trivial progress towards the truth P_* ($t_w = 1$). Even if we train on only Dogs (118 classes) we make progress on the entire ImageNet.

points along the average trajectory of each of the 6 tasks are also embedded together with the 1830 checkpoints (i.e., $m = 2430$ in (9)). Figure 1a plots the top three dimensions obtained from InPCA. To clarify, the explained stress of the top 2430 dimensions would be exactly 100%.

Result 2: Supervised learning on one task results in a surprising amount of progress on seemingly dissimilar tasks. Progress on other tasks is larger if the training task has diverse classes. We studied the progress t_w (4) made by models (Figure 2b) trained on tasks from Result 1. Training on the task "Dogs" makes non-trivial progress on other tasks, even seemingly dissimilar ones like "Instruments" which contains vehicles, devices and clothing. In fact, it makes a remarkable amount of progress on the *entire* ImageNet, about 63.38% of the progress of a model trained directly on ImageNet. Progress is larger for larger phyla of ImageNet (Vertebrates and Instruments). But what is surprising is that if we train on a random subset of 333 classes (a third of ImageNet), then the progress on the entire ImageNet is very large (92%). This points to a strong shared structure among classes even for large datasets such as ImageNet. Note that this *does not* mean that tasks such as Vertebrates and Instruments are similar to each other. Even if training trajectories are similar for a while, they do bifurcate eventually and the final models are indeed different (see Figure 3b and Remark 5 on how to interpret it).

In Figure 2a, we studied the projections of models trained on one task onto the geodesics of unseen classes calculated using (3) evaluated at the progress t_w (4)). We find

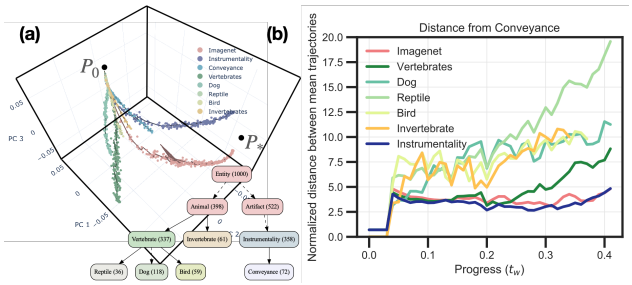


Figure 3. **(a)** Trajectories of models trained on different phyla of Wordnet (inset). The model manifold is again effectively low-dimensional (78.72% explained stress in 3 dimensions).

(b) We analyze the trajectories in Figure 3(a) and obtain a quantitative description of how trajectories of different tasks diverge from each other during training; the procedure is explained in in Remark 5. The plot depicts the Bhattacharyya distance between the mean trajectories (over random initializations) on different tasks, and the mean trajectory of Conveyance. This distance is normalized by the average of the tube radii (maximum distance of one of the 5 trajectories from the mean, computed at each progress) of the two trajectories. Such quantities allow us to make precise statements about the differences between representations and show some very surprising conclusions. Trajectories of tasks that are nearby in Wordnet are also nearby in terms of their learned representations. Further, trajectories of ImageNet (pink) are closer to Conveyance (as expected), but those of Vertebrates (red) are equally far away for more than 60% ($t_w \approx 0.25$) of the progress. In other words, training on Vertebrates (reptiles, dog, bird) makes a remarkable progress on Conveyance (cars, planes).

that a model trained on the entire ImageNet makes uneven progress on the various classes (but about 80% progress across them, progress is highly correlated with test error of different classes). Models trained on the 6 individual tasks also make progress on other unseen classes. As before, training on Instruments, Vertebrates, Dogs makes smaller progress on unseen classes compared to training on a random subset of 333 classes. This is geometric evidence that the more diverse the training dataset, the better the generalization to *unseen* classes/tasks; this phenomenon has been widely noticed and utilized to train models on multiple tasks, as we discuss further in Result 4.

Result 3: The structure of the space of tasks indicated by our visualization technique is consistent with parts of the Wordnet phylogenetic tree. To obtain a more fine-grained characterization of how the geometry in the space of learnable tasks reflects the semantics of these tasks, we selected two particular phyla of ImageNet (Animals, Artifacts) and created sub-tasks using classes that belong to these phyla (Figure 3a). Trajectories of models trained on Instruments and Conveyance are closer together than those of Animals. Within the Animals phylum, trajectories of Vertebrates (Dog, Reptile, Bird) are closer together than those of Invertebrates (Figure 3b for quantitative metrics). Effectively, we can recover a part of the phylogenetic tree of Wordnet using our training trajectories. We speculate that this may point to some shared structure between visual features of images and natural language-based semantics of the corresponding categories which was used to create Wordnet (Miller, 1998) of the corresponding categories. Such alignment with a natural notion of relatedness also demonstrates the soundness and effectiveness of our technical machinery.

Remark 5 (Building a precise and quantitative characterization of trajectories of representations). The precise way to understand statements like those in Result 3 is using the quantitative analysis reported in Figure 3b and Figure A3. To expand upon the caption, the X-axis of the plot is progress. For multiple models (5 random seeds) trained on two tasks (say Conveyance and Dogs), we have calculated the mean (across random seeds) of the interpolated trajectories at different progress. At each specific progress, we have plotted the distance between the mean model trained on Conveyance (say task 1) and Dogs (say task 2) divided by the average tube radii (which is the maximum of the distance of the model corresponding to one seed from the mean):

$$2\mathrm{d_B}(\tau_{\text{mean}}^{1\to U}, \tau_{\text{mean}}^{2\to U}) / \textstyle\sum_{k=1,2} \max_a [\mathrm{d_B}(\tau_a^{k\to U}, \tau_{\text{mean}}^{k\to U})].$$

The is a measure of how far away the trajectories of these two models are. If it is less than 1, then the "tubes" corresponding to models trained on tasks 1 and task 2 intersect.

Let us emphasize that we have performed such analyses for all experiments in this paper (see Figure A3); while the InPCA embedding gives an easy-to-understand visual description of these results for high-dimensional probabilistic models, the information geometric techniques developed in this paper enable us to make these descriptions precise and quantitative. We also include a similar step-by-step guide on how to interpret Figure A1b in Appendix B.

Result 4: Episodic meta-learning algorithms traverse very different trajectories during training but they fit a similar model eventually. Meta-learning methods build a representation which can be adapted to a new task (Thrun & Pratt, 2012). We studied a common variant, the so-called episodic training methods (Bengio et al., 1992), in the context of few-shot learning methods (Vinyals et al., 2016). In these methods, each mini-batch consists of samples from C^w out of C classes (called "way") split into two parts: a "support set" D_s of s samples/class (called "shot"), and a "query set" D_q of q samples/class. Typical methods, say prototypical networks of Snell et al. (2017b), implement a clustering loss on features of the query samples using averaged features of the support samples $\varphi_c = s^{-1} \sum_{\{x \in D_s, y^*(x)=c\}} \varphi(x; w_1)$ for all $c = 1, \ldots, C^w$ as the cluster centroids. If features φ lie on an ℓ_2 ball of

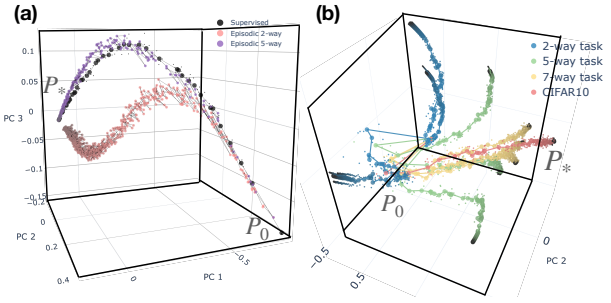


Figure 4. **(a)** Training trajectories for supervised learning (black), 2-way (pink) and 5-way episodic meta-learning (purple). Trajectories of 5-way meta-learning are very similar to those of supervised learning and eventually reach very similar models and high test accuracy. In contrast, 2-way meta-learning has a much longer trajectory (about 40× longer in Riemann length than black) and does not reach a good test accuracy (on all 10 CIFAR-10 classes). Representations are similar during early parts of training even if these are quite different learning mechanisms. **(b)** Trajectories of 2-way (blue), 5-way (green), 7-way (yellow) tasks trained using cross-entropy loss compared to supervised learning (red). For large "way", trajectories are similar to supervised learning but they quickly deviate from the red trajectories for small ways.

radius 1, then doing so is akin to maximizing the cosine similarity between cluster centroids and features of query samples. The same clustering loss with the learned backbone w_1 is used to predict on unseen classes (using "few" support samples to compute centroids) at test time.

To understand the representations learned by episodic meta-learning methods, we compared trajectories of episodic meta-learning to the trajectory taken by supervised learning in Figure 4. Supervised learning uses the cross-entropy loss over all the C classes while episodic meta-learning optimizes a loss that considers all k-way classification tasks (where k is typically smaller than C), its objective differs from that used for supervised learning. Since the two objectives are different, it comes as a surprise that both arrive at the same solution; see Figure 4a,b and Figure A2 for distances between trajectories. But the Riemann trajectory length of episodic training is about 40× longer than that of supervised learning. It is worth noting that the explained stress is only 40.96% in Figure 4a because of larger fluctuations for episodic learning in other directions. Therefore, episodic meta-learning has a qualitatively different training trajectory in the prediction space than supervised learning. The implications of this are consistent with recent literature which has noticed that the performance of few-shot learning methods using supervised learning (followed by fine-tuning) is comparable to, or better than, episodic meta-learning (Dhillon et al., 2020; Kolesnikov et al., 2020; Fakoor et al., 2020). Indeed, a supervised learned representation also minimizes the clustering loss.

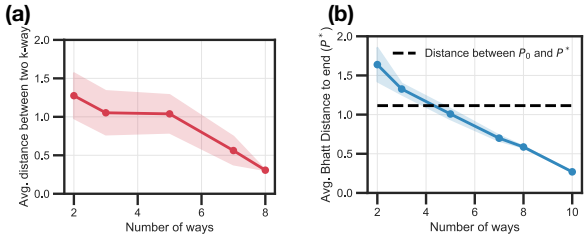


Figure 5. **(a)** Average distance between two k-way meta-learning trajectories decreases with k, this is a geometric evidence of the variance of predictions of learned representations. **(b)** Training with a small way leads to models that predict poorly on test data (large distances from truth). These embeddings were calculated using the training dataset. The rationale being that we wanted to show how different meta-learning and supervised learning are during training.

In order to understand why few-shot accuracy of episodic training is better with a large way (Gidaris & Komodakis, 2018), we trained models on different 2-way 5-way and 7-way tasks using the cross-entropy loss (Figure 4b). We find that the radius of the tube that encapsulates the models of 2-way tasks around their mean trajectory is very large, almost as large as the total length of the trajectory, i.e., different models trained with a small way tasks traverse very different trajectories. Tube radius decreases as the way increases (Figure 5a). Further, the distance of models from the truth P_* (which is close to the end point of the supervised learning model) is higher for a small way (Figure 5b). This is geometric evidence of the widely used empirical practice of using a large way in episodic meta-learning. Observe in Figure 5b that as the way increases, the trajectory becomes more and more similar to that of supervised learning. See Figure A3 for a quantitative analysis of these trajectories.

Result 5: Contrastive and semi-supervised learning methods traverse trajectories similar to those of supervised learning. Contrastive learning (Becker & Hinton, 1992) learns representations without using ground-truth labels (Gutmann & Hyvärinen, 2010; Chen et al., 2020a). It has been extremely effective for self-supervised learning (Doersch & Zisserman, 2017; Kolesnikov et al., 2019), e.g., prediction accuracy with 1–10% labeled data is close to that of supervised learning using all data (Chen et al., 2020b). Semi-supervised methods (Berthelot et al., 2019; Sohn et al., 2020) learn representations when ground-truth labels are available for only a small fraction of the data (0.1–1%). These methods achieve a prediction accuracy within 5% of the accuracy achieved through supervised learning. We compared representations learned using contrastive and semi-supervised learning with those from supervised learning to understand why these methods are so effective.

Consider a task P and a set of augmentations G (e.g., cropping, resizing, blurring, color/contrast/brightness dis-

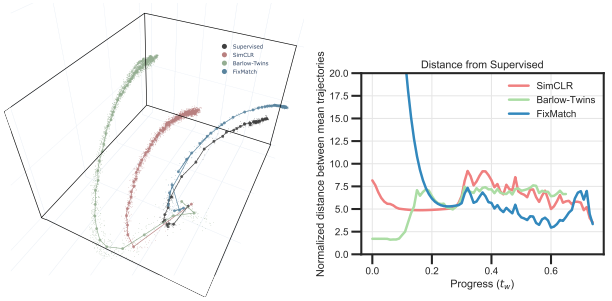


Figure 6. We consider 4 methods for training on CIFAR10: supervised learning, SimCLR (Chen et al., 2020a), Barlow-twins (Zbontar et al., 2021) and Fixmatch (Sohn et al., 2020). Fixmatch has access to 2500 labeled samples and 47500 unlabeled samples. SimCLR and Barlow-twins use 50,000 unlabeled samples for training.
(a) We plot the trajectories for supervised, semi-supervised and contrastive learning. The trajectory of semi-supervised learning (Fixmatch) eventually resembles supervised learning in comparison to contrastive learning methods. All methods result in remarkably similar trajectories although some of these methods are trained using only unlabeled data.
(b) Normalized distance of trajectories corresponding to contrastive and semi-supervised learning to the trajectory of supervised learning. Semi-supervised learning (Fixmatch) deviates considerably from the other methods at the beginning. We speculate that this is because the trajectory of Fixmatch is influenced by the 2500 labeled samples. As, training progresses, Fixmatch becomes increasingly similar to supervised learning as evidenced by the dip in the blue line for larger values of progress (t_w).

tortion etc.). Given inputs (say images) x from P, contrastive learning forces the representation $\varphi(g(x); w_1)$ and $\varphi(g'(x); w_1)$ (shortened to $\varphi(g(x))$ below) of the same input for two different augmentations g, g' to be similar. And forces it to be different from representations of other augmented inputs x' (Zbontar et al., 2021; Bachman et al., 2019; Dosovitskiy et al., 2014). Semi-supervised learning methods have access to both labeled inputs x_l and unlabeled inputs x_u. More recent methods are usually trained to fit the labeled inputs using the cross-entropy loss while enforcing consistent predictions across all augmentations (Tarvainen & Valpola, 2017; Berthelot et al., 2019) for any unlabeled input.

We compare the representations of semi-supervised (Fixmatch (Sohn et al., 2020)), contrastive (SimCLR (Chen et al., 2020a), Barlow-twins (Zbontar et al., 2021)) and supervised learning in Figure 6. All three trajectories are similar to the trajectory of supervised learning. We find that the trajectory of semi-supervised learning deviates from the supervised learning trajectory initially, but the two are very similar for larger values of progress (t_w). This points to a remarkable ability of semi and self-supervised learning methods to learn representations that are similar to those of supervised learning; it is not just that the accuracy of these methods is similar, they also learn similar probabilistic models.

Result 6: Fine-tuning a pre-trained model on a sub-task does not change the representation much. To understand how models train on multiple tasks, we selected two binary classification sub-tasks of CIFAR-10 (Airplane vs. Automobile, and Bird vs. Cat).

We selected models at different stages of standard supervised learning on CIFAR-10 (i.e., using 10-way output and softmax cross-entropy loss) and fine-tuned each of these models on two sub-tasks (the entire network is fine-tuned without freezing the backbone). As Figure 7 shows, models that were fine-tuned from earlier parts of the trajectory travel a large distance and move away from trajectories of the supervised learned CIFAR-10 models. As we fine-tune later and later models, the distance traveled away from the trajectory is smaller and smaller, i.e., changes in the representation are smaller. For a fully-trained CIFAR-10 model which interpolates the training data, the distance traveled by fine-tuning is very small (the points are almost indistinguishable in the picture); this is because both P^1 and P^2 are subsets of CIFAR-10.

Algorithms for transfer learning train on a source task before fine-tuning the model on the target task. If two tasks share a large part of their training trajectory, then we may start the fine-tuning from many shared intermediate points—there are many such points. If the chosen point is farther along in terms of progress then the efficiency resulting from using the source task is higher because the trajectory required to fit the target task is shorter; such trajectories were used in (Gao & Chaudhari, 2021) to define a distance between tasks. As we saw in Result 2, trajectories of different tasks bifurcate after a shared part. The resultant deviation less for related tasks and more for dissimilar tasks (Figure 7a, Figure 1a,c). Therefore it is difficult to know *a priori* from which point one should start the fine-tuning from without knowing the manifold of the target task. In particular, our geometric picture indicates that fine-tuning from a fully-trained model can be detrimental to the accuracy on the target task. This has been noticed in a number of places in the transfer learning literature, e.g., Li et al. (2020), and has also been studied theoretically (Gao & Chaudhari, 2020).

4. Related Work and Discussion

Understanding the space of learnable tasks A large body of work has sought to characterize relationships between tasks, e.g., domain specific methods (Zamir et al., 2018; Cui et al., 2018; Pennington et al., 2014), learning theoretic work (Baxter, 2000; Maurer, 2006; Ben-David et al., 2010; Ramesh & Chaudhari, 2022; Tripuraneni et al., 2020; Hanneke & Kpotufe, 2020; Caruana, 1997), random matrix models (Wei et al., 2022), neural tangent kernel models (Malladi et al., 2022) and information-theoretic analy-

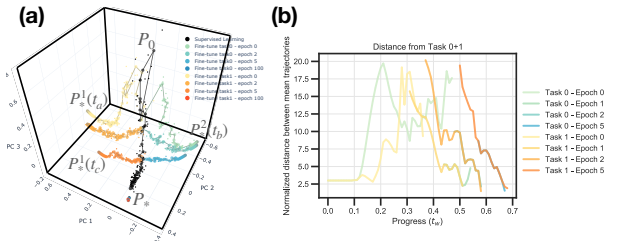


Figure 7. **(a)** Fine-tuning trajectories on Airplane vs. Automobile, and Bird vs. Cat sub-tasks of CIFAR-10 (warm and cold hues) pre-trained from different points along the trajectory of supervised learning. If the pretrained model has progressed further towards the truth P_*, then fine-tuning it on a sub-task does not change the representation much. The final trajectory (fine-tuning from epoch 100) is indistinguishable from P_*. **(b)** Bhattacharyya distance between the mean trajectories normalized by the average of the tube radii (like Figure 3b). models (say, fine-tuned after epoch 5 on task 1) go *backwards* in terms of progress, i.e., they unlearn the pre-trained representation in order to fit the new task. This occurs as early as epoch 1 here. It suggests that learning occurs extremely rapidly at the beginning and determines the efficiency of fine-tuning. Some curves here are not visible because they are overlapping heavily.

ses (Jaakkola & Haussler, 1999; Achille et al., 2019a;b). Broadly speaking, this work has focused on understanding the accuracy of a model on a new task when it is trained upon a related task, e.g., relationships between tasks are characterized using the excess risk of a hypothesis. Our methods also allow us to say things like "task P^1 is far from P^2 as compared to P^3". But they can go further. We can glean a global picture of the geometric structure in the space of tasks and quantify statements such as "the divergence between P^1 and P^2 eventually is more than that of P^1 and P^3, but representations learned on these tasks are similar for 30% of the way".

There is strong structure in typical inputs, e.g., recent work on understanding generalization (Yang et al., 2022; Bartlett et al., 2020) as well as older work such as Simoncelli & Olshausen (2001); Field (1994); Marr (2010) has argued that visual data is effectively low-dimensional. Our works suggests that tasks also share a low-dimensional structure. Just like the effective low-dimensionality of inputs enables generalization on one task, effective low-dimensionality of the manifold of models trained on different tasks could perhaps explain generalization to new tasks.

Relationships between tasks in neuroscience Our results are conceptually close to those on organization and representation of semantic knowledge (Mandler & McDonough, 1993). Such work has primarily used simple theoretical models, e.g., linear dynamics of Saxe et al. (2019) (who also use MDS). Our tools are very general and paint a similar picture of ontologies of complex tasks. Concept formalization and specialization over age (Vosniadou & Brewer, 1992) also resembles our experiment in how fine-tuning models trained for longer periods changes the representation marginally. Our broad goals are similar to those of Sorscher et al. (2021) but our techniques are very different.

Information Geometry has a rich body of sophisticated ideas (Amari, 2016), but it has been difficult to wield it computationally, especially for high-dimensional models like deep networks. Our model in (1) is a finite-dimensional probability distribution, in contrast to the standard object in information geometry which is an infinite-dimensional probability distribution defined over the entire domain. This enables us to compute embeddings of manifolds, geodesics, projections etc. We are not aware of similar constructions in the literature.

Visualizing training trajectories of deep networks InPCA is a variant of multi-dimensional scaling (MDS, see Cox & Cox (2008)), with the difference being that InPCA retains the negative eigenvalues which preserves pairwise distances (Quinn et al., 2019). A large number of works have investigated trajectories of deep networks and the energy landscape during or after training using dimensionality reduction techniques (Horoi et al., 2021; Li et al., 2018; Huang et al., 2020). Gur-Ari et al. (2018); Antognini & Sohl-Dickstein (2018) studied the dimensionality of training trajectories. The key distinction here with respect to this body of work is that we study the prediction space, not the weight space. While the weight space has symmetries (Freeman & Bruna, 2017; Garipov et al., 2018) and nontrivial dynamics (Tanaka & Kunin, 2021; Chaudhari & Soatto, 2018), the prediction space, i.e,. $[0,1]^{N\times C} \ni \{p_w(c \mid x_i)\}$, completely characterizes the output of a probabilistic model. In comparison, the loss or the error which are typically used to reason about relationships between tasks, are coarse summaries of the predictions. Any two models, irrespective of their architecture, training methodology, or even the task that they were trained on, can be studied rigorously using our techniques.

Acknowledgments

RR, JM, RY and PC were supported by grants from the National Science Foundation (IIS-2145164, CCF-2212519), the Office of Naval Research (N00014-22-1-2255), and cloud computing credits from Amazon Web Services. IG was supported by the National Science Foundation (DMREF-89228, EFRI-1935252) and Eric and Wendy Schmidt AI in Science Postdoctoral Fellowship. HKT was supported by the National Institutes of Health (1R01NS116595-01). JPS was supported by the National Science Foundation (DMR-1719490), MKT was supported by the National Science Foundation (DMR-1753357). The authors would like to acknowledge Jay Spendlove for helpful comments on this material.

References

Achille, A., Lam, M., Tewari, R., Ravichandran, A., Maji, S., Fowlkes, C. C., Soatto, S., and Perona, P. Task2vec: Task embedding for meta-learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 6430–6439, 2019a.

Achille, A., Mbeng, G., and Soatto, S. Dynamics and Reachability of Learning Tasks. *arXiv:1810.02440 [cs, stat]*, May 2019b.

Amari, S.-i. *Information Geometry and Its Applications*, volume 194 of *Applied Mathematical Sciences*. Tokyo, 2016.

Antognini, J. and Sohl-Dickstein, J. Pca of high dimensional random walks with comparison to neural network training. *Advances in Neural Information Processing Systems*, 31, 2018.

Bachman, P., Hjelm, R. D., and Buchwalter, W. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

Baxter, J. A model of inductive bias learning. *Journal of artificial intelligence research*, 12:149–198, 2000.

Becker, S. and Hinton, G. E. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356):161–163, 1992.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. A theory of learning from different domains. *Machine learning*, 79(1):151–175, 2010.

Bengio, S., Bengio, Y., Cloutier, J., and Gecsei, J. On the optimization of a synaptic learning rule. In *Preprints Conf. Optimality in Artificial and Biological Neural Networks*, pp. 6–8, 1992.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., and Raffel, C. A. MixMatch: A holistic approach to semi-supervised learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Bhattacharyya, A. On a measure of divergence between two multinomial populations. *Sankhyā: the indian journal of statistics*, pp. 401–406, 1946.

Bostock, M. Imagenet hierarchy. https://observablehq.com/@mbostock/imagenet-hierarchy, 2018.

Brent, R. P. An algorithm with guaranteed convergence for finding a zero of a function. *The computer journal*, 14(4): 422–425, 1971.

Caruana, R. Multitask learning. *Machine learning*, 28(1): 41–75, 1997.

Chaudhari, P. and Soatto, S. Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks. In *Proc. of International Conference of Learning and Representations (ICLR)*, 2018.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pp. 1597–1607, 2020a.

Chen, T., Kornblith, S., Swersky, K., Norouzi, M., and Hinton, G. E. Big self-supervised models are strong semi-supervised learners. *Advances in Neural Information Processing Systems*, 33:22243–22255, 2020b.

Cox, M. A. and Cox, T. F. Multidimensional scaling. In *Handbook of Data Visualization*, pp. 315–347. 2008.

Cui, Y., Song, Y., Sun, C., Howard, A., and Belongie, S. Large scale fine-grained categorization and domain-specific transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4109–4118, 2018.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Dhillon, G. S., Chaudhari, P., Ravichandran, A., and Soatto, S. A baseline for few-shot image classification. In *Proc. of International Conference of Learning and Representations (ICLR)*, 2020.

Doersch, C. and Zisserman, A. Multi-task self-supervised visual learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2051–2060, 2017.

Dosovitskiy, A., Springenberg, J. T., Riedmiller, M., and Brox, T. Discriminative unsupervised feature learning with convolutional neural networks. *Advances in neural information processing systems*, 27, 2014.

Fakoor, R., Chaudhari, P., Soatto, S., and Smola, A. J. Meta-Q-Learning. In *Proc. of International Conference of Learning and Representations (ICLR)*, 2020.

Field, D. J. What is the goal of sensory coding? *Neural computation*, 6(4):559–601, 1994.

Freeman, C. D. and Bruna, J. Topology and geometry of half-rectified network optimization. In *ICLR*, 2017.

Gao, Y. and Chaudhari, P. A free-energy principle for representation learning. In *Proc. of International Conference of Machine Learning (ICML)*, 2020.

Gao, Y. and Chaudhari, P. An information-geometric distance on the space of tasks. In *Proc. of International Conference of Machine Learning (ICML)*, 2021.

Garipov, T., Izmailov, P., Podoprikhin, D., Vetrov, D., and Wilson, A. G. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 8803–8812, 2018.

Gidaris, S. and Komodakis, N. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4367–4375, 2018.

Gur-Ari, G., Roberts, D. A., and Dyer, E. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018.

Gutmann, M. and Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 297–304. JMLR Workshop and Conference Proceedings, 2010.

Hanneke, S. and Kpotufe, S. A no-free-lunch theorem for multitask learning. *arXiv preprint arXiv:2006.15785*, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Horoi, S., Huang, J., Wolf, G., and Krishnaswamy, S. Visualizing high-dimensional trajectories on the loss-landscape of ANNs. *arXiv preprint arXiv:2102.00485*, 2021.

Hu, J., Lu, J., and Tan, Y.-P. Deep transfer metric learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 325–333, 2015.

Huang, W., Yi, M., and Zhao, X. Towards the generalization of contrastive self-supervised learning. *arXiv preprint arXiv:2111.00743*, 2021.

Huang, W. R., Emam, Z., Goldblum, M., Fowl, L., Terry, J. K., Huang, F., and Goldstein, T. Understanding generalization through visualizations. 2020.

Ilyas, A., Park, S. M., Engstrom, L., Leclerc, G., and Madry, A. Datamodels: Predicting predictions from training data. *arXiv preprint arXiv:2202.00622*, 2022.

Ito, S. and Dechant, A. Stochastic time evolution, information geometry, and the Cramér-Rao bound. *Physical Review X*, 10(2):021056, 2020.

Jaakkola, T. and Haussler, D. Exploiting generative models in discriminative classifiers. In *Advances in Neural Information Processing Systems*, pp. 487–493, 1999.

Kaplun, G., Ghosh, N., Garg, S., Barak, B., and Nakkiran, P. Deconstructing distributions: A pointwise framework of learning. *arXiv preprint arXiv:2202.09931*, 2022.

Kolesnikov, A., Zhai, X., and Beyer, L. Revisiting self-supervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1920–1929, 2019.

Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., and Houlsby, N. Big Transfer (BiT): General Visual Representation Learning. *arXiv:1912.11370 [cs]*, May 2020.

Krizhevsky, A. *Learning Multiple Layers of Features from Tiny Images*. PhD thesis, Computer Science, University of Toronto, 2009.

Leclerc, G., Ilyas, A., Engstrom, L., Park, S. M., Salman, H., and Madry, A. ffcv. `https://github.com/libffcv/ffcv/`, 2022. commit xxxxxxx.

Li, H., Xu, Z., Taylor, G., and Goldstein, T. Visualizing the loss landscape of neural nets. In *ICLR*, 2018.

Li, H., Chaudhari, P., Yang, H., Lam, M., Ravichandran, A., Bhotika, R., and Soatto, S. Rethinking the hyperparameters for fine-tuning. In *Proc. of International Conference of Learning and Representations (ICLR)*, 2020.

Malladi, S., Wettig, A., Yu, D., Chen, D., and Arora, S. A kernel-based view of language model fine-tuning. *arXiv preprint arXiv:2210.05643*, 2022.

Mandler, J. M. and McDonough, L. Concept formation in infancy. *Cognitive development*, 8(3):291–318, 1993.

Marr, D. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. 2010.

Maurer, A. Bounds for linear multi-task learning. *The Journal of Machine Learning Research*, 7:117–139, 2006.

Miller, G. A. *WordNet: An Electronic Lexical Database*. 1998.

Nielsen, F. and Boltz, S. The burbea-rao and bhattacharyya centroids. *IEEE Transactions on Information Theory*, 57 (8):5455–5466, 2011.

Pennington, J., Socher, R., and Manning, C. D. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.

Qi, H., Brown, M., and Lowe, D. G. Low-shot learning with imprinted weights. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5822–5830, 2018.

Quinn, K. N. *Patterns of Structural Hierarchies in Complex Systems*. PhD thesis, Cornell University, 2019.

Quinn, K. N., Clement, C. B., De Bernardis, F., Niemack, M. D., and Sethna, J. P. Visualizing probabilistic models and data with intensive principal component analysis. *Proceedings of the National Academy of Sciences*, 116 (28):13762–13767, 2019.

Ramesh, R. and Chaudhari, P. Model Zoo: A Growing "Brain" That Learns Continually. In *Proc. of International Conference of Learning and Representations (ICLR)*, 2022.

Ruan, Y., Dubois, Y., and Maddison, C. J. Optimal representations for covariate shift. *arXiv preprint arXiv:2201.00057*, 2021.

Saxe, A. M., McClelland, J. L., and Ganguli, S. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23):11537–11546, 2019.

Shen, Z., Liu, Z., Liu, Z., Savvides, M., Darrell, T., and Xing, E. Un-mix: Rethinking image mixtures for unsupervised visual representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2216–2224, 2022.

Shi, X., Padhi, I., and Knight, K. Does string-based neural MT learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1526–1534, 2016.

Simoncelli, E. P. and Olshausen, B. A. Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216, 2001.

Snell, J., Swersky, K., and Zemel, R. Prototypical networks for few-shot learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 4080–4090, 2017a.

Snell, J., Swersky, K., and Zemel, R. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems*, pp. 4077–4087, 2017b.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., Cubuk, E. D., Kurakin, A., and Li, C.-L. FixMatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in Neural Information Processing Systems*, 33, 2020.

Sorscher, B., Ganguli, S., and Sompolinsky, H. The geometry of concept learning. *bioRxiv : the preprint server for biology*, 2021.

Tanaka, H. and Kunin, D. Noether's learning dynamics: The role of kinetic symmetry breaking in deep learning. *arXiv preprint arXiv:2105.02716*, 2021.

Tarvainen, A. and Valpola, H. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *arXiv preprint arXiv:1703.01780*, 2017.

Teoh, H. K., Quinn, K. N., Kent-Dobias, J., Clement, C. B., Xu, Q., and Sethna, J. P. Visualizing probabilistic models in Minkowski space with intensive symmetrized Kullback-Leibler embedding. *Physical Review Research*, 2(3):033221, August 2020. ISSN 2643-1564.

Thrun, S. and Pratt, L. *Learning to Learn*. 2012.

Tripuraneni, N., Jin, C., and Jordan, M. I. Provable meta-learning of linear representations. *arXiv preprint arXiv:2002.11684*, 2020.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. Matching networks for one shot learning. In *Advances in Neural Information Processing Systems*, pp. 3630–3638, 2016.

Vosniadou, S. and Brewer, W. F. Mental models of the earth: A study of conceptual change in childhood. *Cognitive psychology*, 24(4):535–585, 1992.

Wei, A., Hu, W., and Steinhardt, J. More than a toy: Random matrix models predict how real-world neural representations generalize. *arXiv preprint arXiv:2203.06176*, 2022.

Yang, R., Mao, J., and Chaudhari, P. Does the data induce capacity control in deep learning? In *Proc. of International Conference of Machine Learning (ICML)*, 2022.

Zagoruyko, S. and Komodakis, N. Wide residual networks. In *British Machine Vision Conference 2016*, 2016.

Zamir, A. R., Sax, A., Shen, W., Guibas, L. J., Malik, J., and Savarese, S. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3712–3722, 2018.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pp. 12310–12320, 2021.

Zhang, R. Making convolutional networks shift-invariant again. In *International conference on machine learning*, pp. 7324–7334. PMLR, 2019.

Zhong, Y., Tang, H., Chen, J., Peng, J., and Wang, Y.-X. Is self-supervised learning more robust than supervised learning? *arXiv preprint arXiv:2206.05259*, 2022.

A. Details of the experimental setup

Data

We performed experiments using two datasets.

1. CIFAR10 (Krizhevsky, 2009) has 10 classes (airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck) with RGB images of size 32×32, and
2. ImageNet (Deng et al., 2009) has 1000 classes each with about 1000 RGB images of size 224×224.

ImageNet classes are derived from the leaves of the Wordnet hierarchy (Miller, 1998) which is visualized by Bostock (2018). We use this hierarchy to create tasks using different subsets of ImageNet; We use all classes under a node to create a task. The tasks that we consider are: Dogs, Vertebrates, Invertebrates, Instrumentality, Reptile and Birds. We also consider a task with 333 randomly selected classes and unlike other tasks, it spans many different phyla of ImageNet.

Architectures We use a Wide-Resnet (Zagoruyko & Komodakis, 2016) architecture for supervised learning experiments on CIFAR-10 (WRN-16-4 with depth 16 and widening factor of 4) and a Resnet-18 (He et al., 2016) to train a model using SimCLR. All experiments on ImageNet use the Resnet-50 architecture.

All convolutional layers are initialized using the Kaiming-Normal initialization. For the Wide-Resnet, the final pooling layer is replaced with an adaptive pooling layer in order to handle input images of different sizes.

We make three modifications to these architectures.

1. We remove the bias from the final classification layer; this helps keep the logits of the different tasks on a similar scale.
2. In the experiments for Result 3 (episodic meta-learning) and Result 6 (fine-tuning), we replace batch normalization with layer norm in the Wide-Resnet. This is because we found in preliminary experiments that batch-normalization parameters make training meta-learning models very sensitive to choices of hyper-parameters (e.g., the support or query shot), and that the learned representations of new tasks were quite different in terms of their predictions (and thereby the Bhattacharyya distance) but all the difference was coming from modifications to the BN parameters.
3. In the Resnet-50, we replace the pooling layers with BlurPool (Zhang, 2019). The bias parameter in batch normalization is set to zero with the affine scaling term set to one.

Training procedure All models are trained in mixed-precision (32-bit weights, 16-bit gradients) using stochastic gradient descent (SGD) with Nesterov's acceleration with momentum coefficient set to 0.9 and cosine annealing of the learning rate schedule. Batch-normalization parameters are excluded from weight decay.

CIFAR10 datasets use padding (4 pixels) with random cropping to an image of size 28×28 or 32×32 respectively for data augmentation. CIFAR10 images additionally have random left/right flips for data augmentation. Images are finally normalized to have mean 0.5 and standard deviation 0.25.

Supervised learning models (including fine-tuning) for CIFAR10 are trained for 100 epochs with a batch-size of 64 and weight decay of 10^{-5} using the Wide-Resnet.

Episodic meta-learners are trained using a Wide-Resnet and with the prototypical loss (Snell et al., 2017a). For the 2-way meta-learner, each episode contains 20 query samples and 10 support samples. For the 5-way meta-learner, each episode contains 50 query samples and 10 support samples. We found (Result 4) to hold across different choices of these hyper-parameters in small-scale experiments. Models are trained for around 750 epochs and the episodic learner is about 5 times slower to train with respect to wall-clock time.

We train models using SimCLR on CIFAR10 and on tasks created from CIFAR10. For the augmentations, we use random horizontal flips, random grayscale, random resized crop and color jitter. Models are trained for 200 epochs for 2-way classification problems and for 500 epochs when trained on the entirety of CIFAR10 with the Adam optimizer and an initial learning rate of 0.001.

A.1. Experiments on ImageNet

We make use of FFCV (Leclerc et al., 2022). which is a data-loading library that replaces the pytorch Dataloader. FFCV reduces the training time on ImageNet to a few hours, which allows us to train 100s of models on ImageNet, or on tasks created from it. Our implementation of ImageNet training builds on the FFCV repository [1].

ImageNet models are trained for 40 epochs with progressive resizing – the image size is increased from 160 to 224 between the epochs 29 and 34. Models are trained on 4 GPUs with a batch-size of 512. The training uses two types of

[1] https://github.com/libffcv/ffcv-imagenet/tree/main

augmentations – random-resized crop and random horizontal flips. Additionally, we use label smoothing with the smoothing parameter set to 0.1.

A.2. Implementing InPCA in very high dimensions

We calculate an InPCA embedding of models along multiple trajectories, e.g., a typical experiment has about 25 trajectories (multiple random seeds, tasks, or representation learning methods) and about 50 models (checkpoints) along each trajectory. Each model is a very high-dimensional object (with dimensionality NC where $N \sim 10^5$ and $C \sim$ 10-10^3). Even if the matrix D in (9) is relatively manageable with $n \sim 1250$, each entry of D is $\mathrm{d_B}(P_u, P_v)$ and therefore requires $\sim 10^8$ operations to compute. Implementing InPCA—or even PCA—for such large matrices requires a large amount of RAM. We reduced the severity of this issue to an extent using Numpy's memmap functionality https://numpy.org/doc/stable/reference/generated/numpy.memmap.html. Also note that calculating only the top few eigenvectors of (9) suffices to visualize the models, we do not need to calculate all.

The formula (2) is an effective summary of the discrepancies between how the predictions made by two probabilistic models differ; even small differences in two models, e.g., even if both P_u and P_v make mistakes on exactly the same input samples, if $p_u^n(c)$ is slightly different than $p_v^n(c)$ for even one of n or c, the divergence is non-zero. InPCA is capable of capturing the differences between two such models (9). However, when the number of classes is extremely large, the number of terms in the summation is prohibitively large and analyzing the discrepancies or calculating the embedding becomes rather difficult.

We also developed a method to work around this issue. We can use a random stochastic matrix (whose columns sum up to 1) to project the outputs for each sample $\{p_u^n(c)\}_{c=1,\ldots,C}$ into a smaller space before calculating (2). This amounts to pretending as if the model predicts not the actual classes but a random linear combination of the classes (even if the model is trained on the actual classes). This is a practical trick that is necessary only when we are embedding a very large number of very high-dimensional probabilistic models. We checked in our Imagenet experiments that using this trick gives the same embeddings.

In this paper, we did not need to use this projection trick. However, we found that this tricks makes it computationally faster to compute the embeddings and we have seen it to work well in practice. We have shared the code for this procedure, since it allows other people to reproduce the results using fewer computational resources.

B. Additional Result

Result 7: Contrastive learning methods trained on different datasets learn similar representations We compared representations learned using contrastive learning with those from supervised learning to understand some aspects of why the former are so effective.

We used SimCLR (Chen et al., 2020a) to perform contrastive learning on images from four sets of classes (airplane-automobile, bird-cat, ship-truck and all of CIFAR-10). We compared the learned representation to that from supervised learning on two tasks (airplane-automobile and all of CIFAR-10) in Figure A1. Models trained using contrastive learning on two-class datasets learn very different representations from models trained on the same task but using supervised learning. Models trained using contrastive learning on different datasets learning similar representations (trajectories of all three two-class datasets are very close to each other). This is reasonable because contrastive learning does not use any information from the labels. It is surprising however that the trajectory of models from contrastive learning on these two-class datasets is similar to trajectories of models from contrastive learning on the entire CIFAR-10.

Let us elaborate upon this a bit more. We have color-matched the lines in Figure A1b with those in Figure A1a. The black curve is the trajectory of supervised learning on the entire CIFAR-10; red is the trajectory of SimCLR trained on the entire CIFAR-10. Figure A1b compares the distances of trajectories in Figure A1a from the red one "contrastive"; this is why there is no red trajectory in Figure A1b.

- The first thing to note here is that the black and red trajectories are quite close to each other; the black line in Figure A1b is only about 20 times far away from red as compared to their corresponding tube radii.
- Next observe that the trajectory of SimCLR on Task 1 (light blue), SimCLR on Task 2 (green) and SimCLR on Task 3 (yellow) are very similar to each other; this is seen in both Figure A1a and in Figure A1b.
- Third, they are closer to SimCLR on all of CIFAR-10 than any supervised learning trajectories (this is seen in Figure A1b because their curves are below everyone else). Thus, contrastive learning on datasets with different classes learns similar representations.
- The learned representation of two-class SimCLR models is similar to the one obtained using data from all classes (red) (in this experiment this occurs up to about $t_w = 0.4$ progress) but they do not go all the way to the truth (i.e., the end

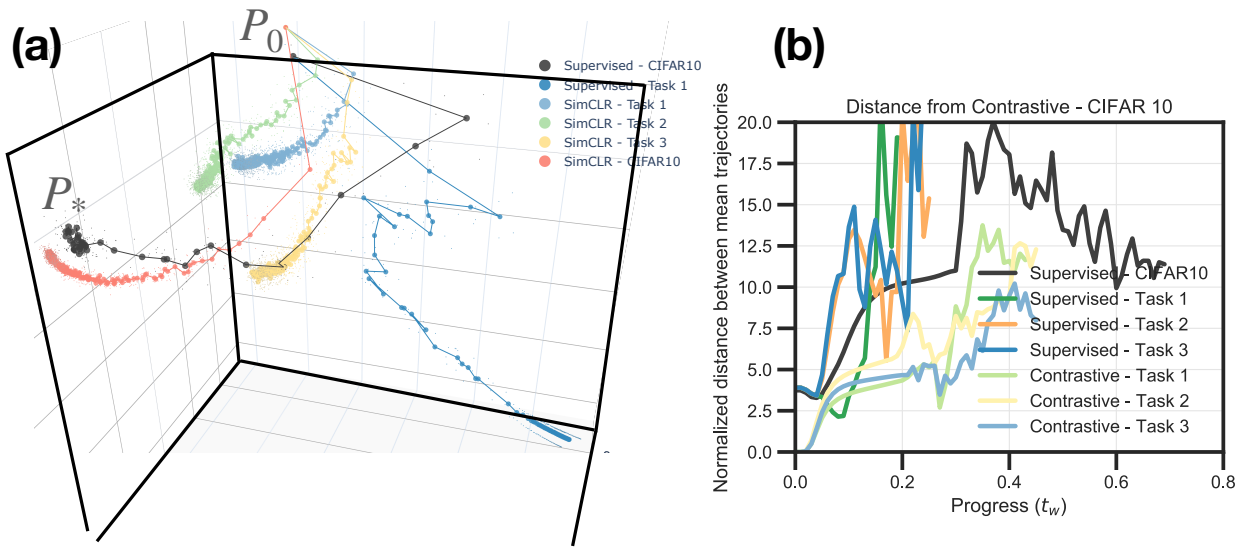


Figure A1. **(a)** Trajectories of contrastive learning (SimCLR) on 3 datasets (two classes each) and entire CIFAR-10 compared to those of supervised learning. SimCLR on entire CIFAR-10 learns a similar representation as that of the supervised learned model P_* (which fits the training data perfectly). SimCLR trajectories are close to each other even if different datasets were used to train them. It may seem from the embedding that SimCLR trajectories are similar to that supervised learning, which would be very surprising because the former does not use any labels, but see below.

(b) Bhattacharyya distance between the mean trajectories of all models and the mean trajectory of SimCLR on all CIFAR-10. This distance is normalized by the average of the tube radii (like Figure 7b). SimCLR trajectories of two-class datasets are indeed very close to each other (mean distance is $\sim 5\times$ more than their tube radii for about 45% of the way ($t_w \approx 0.2$)). This plot indicates that two-class SimCLR trajectory (light blue) is close to SimCLR on all of CIFAR-10. But two-class supervised learning trajectory (darker blue) is much farther away from SimCLR on all of CIFAR-10.

point of black line). This shows the benefit of having data from many classes during contrastive learning.

Also see Figure A3 for distances computed with respect to other trajectories which can be used to further investigate these claims.

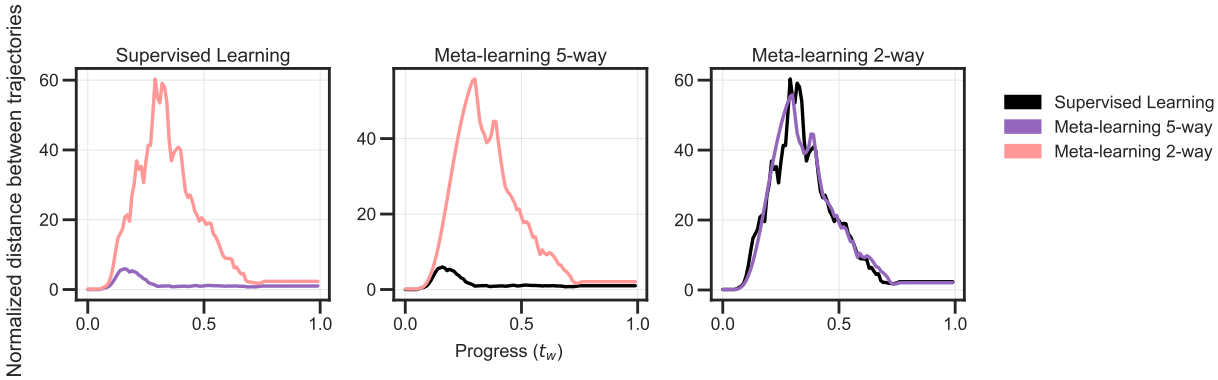
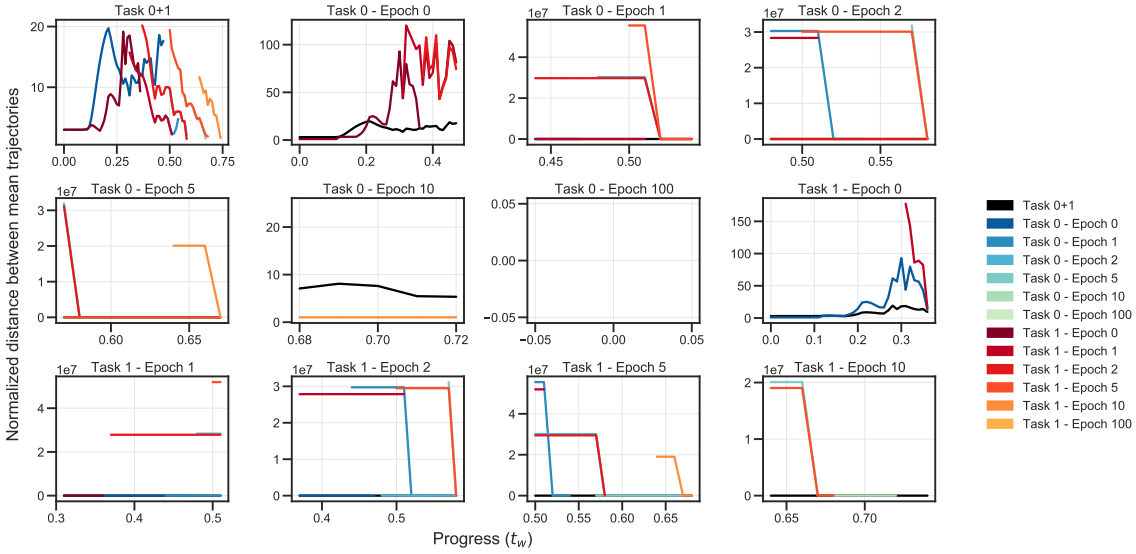


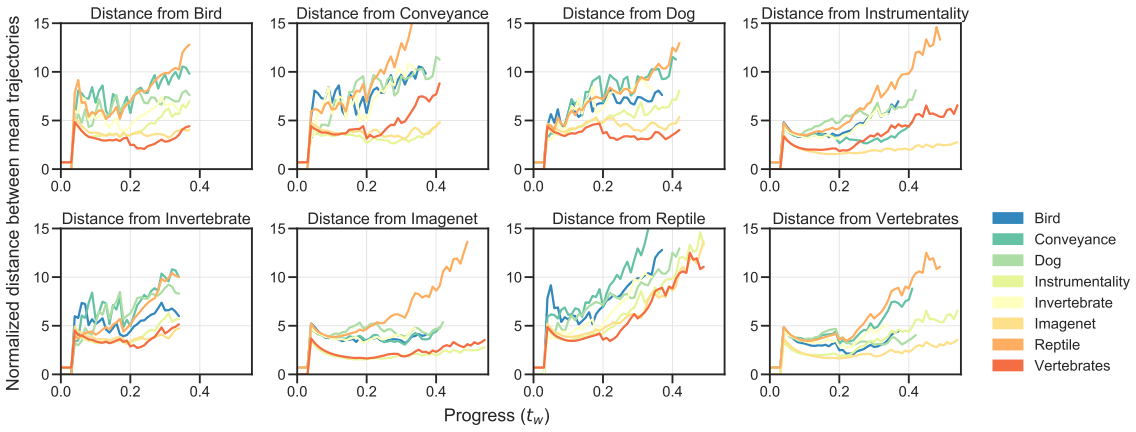
Figure A2. **Distance between trajectories of supervised and meta-learning at different values of progress.** Distances between the average trajectories of different algorithms (e.g., 2-way episodic learning and supervised learning, and 5-way episodic learning and supervised learning in the leftmost panel) are normalized by the average of the radii of the tubes corresponding to each trajectory. We find that trajectories of 2-way meta-learning deviate significantly from those of supervised learning for a large fraction of the trajectory. On the other hand, 5-way meta-learning is similar to the supervised learning trajectory for almost the entirety of the trajectory.

C. Bhattacharyya Distance

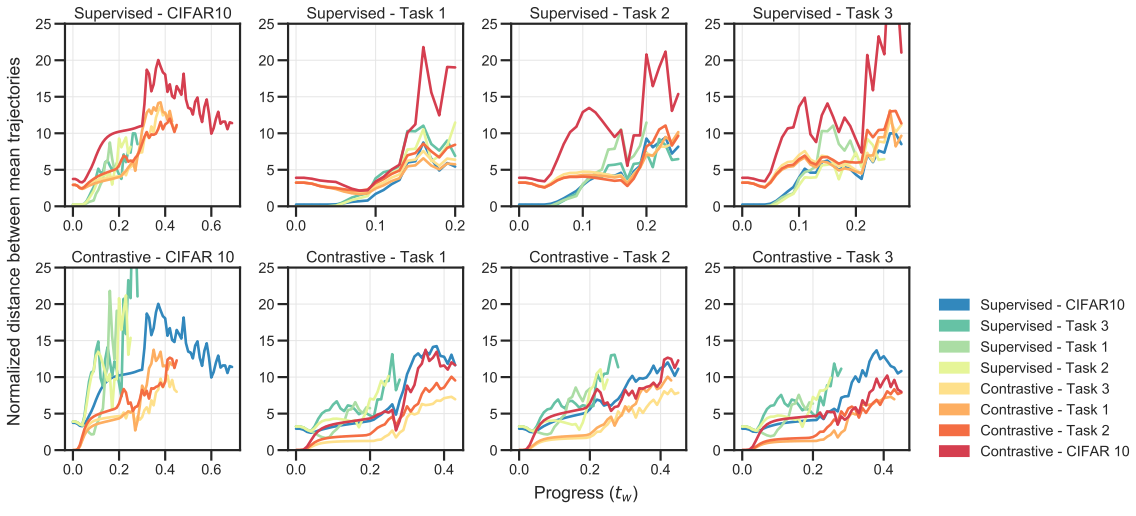
We provide additional details regarding (2). Let $\bar{y} = (y_1, \cdots y_N)$, denote the labels assigned to each of the N samples. Since there are C classes in total, $\vec{y}$ can take a total of C^N different values denoted by the set Y^N. Given, two models P_u



(a)



(b)



(c)

Figure A3. This figure shows the extended version of the distances between trajectories of probabilistic models; two of them are identical to the ones in Figure 7b and Figure A1b.

and P_v, the Bhattacharyya distance averaged over the samples is

$$\begin{aligned}
\mathrm{d_B}(P_u,P_v) &:= -N^{-1}\log\left(\sum_{\bar{y}\in Y^N}\sqrt{P_u(\bar{y})P_v(\bar{y})}\right)\\
&= -N^{-1}\log\left(\sum_{\vec{y}\in Y^N}\prod_{n=1}^{N}\sqrt{p_u(y_n)\,p_v(y_n)}\right)\\
&= -N^{-1}\log\left(\sum_{y_1=1}^{C}\sum_{y_2=1}^{C}\cdots\sum_{y_N=1}^{C}\left(\prod_{n=1}^{N}\sqrt{p_u(y_n)\,p_v(y_n)}\right)\right)\\
&= -N^{-1}\log\left(\prod_{i=1}^{N}\left(\sum_{y_i=1}^{C}\sqrt{p_u(y_i)\,p_v(y_i)}\right)\right)\\
&= -N^{-1}\sum_{i=1}^{N}\log\left(\sum_{y_i=1}^{C}\sqrt{p_u(y_i)\,p_v(y_i)}\right).
\end{aligned}$$

Uncovering the structure of high-dimensional probabilistic models is difficult because most distances between probability distributions saturate with the dimensionality, e.g., the Hellinger distance which is a metric, is essentially equal to 2 in high-dimensions. Quinn et al. (2019, Figure 1) illustrates how a high-dimensional model benefits from using the Bhattacharyya distance compared to using the Hellinger distance in uncovering the intrinsic structure of the manifold. We believe that the logarithm in the Bhattacharyya distance keeps it well-behaved. We actually know of one other distance that gives meaningful results and that is the symmetric KL-divergence (Teoh et al., 2020), for the same reason: due to the logarithm. All analysis in our paper can therefore be done with the symmetric-KL divergence (which is also not a metric) and the results do look similar.

There are a couple more reasons that motivated us to use the Bhattacharyya distance. First, the Bhattacharyya distance to the truth P_* is equal to one half of the cross-entropy loss. Second, it is reassuring that both the Bhattacharyya distance locally gives the Fisher Information Matrix, which is positive semi-definite and therefore induces a local metric.

Bhattacharyya distance violates the triangle inequality and we speculate that this is necessary in order to uncover the low-dimensional structure in high-dimensional data. Understanding why it is important to violate the triangle inequality is a deep question and we do not know how to answer it yet. We do not use the Bhattacharyya distance itself to say things like "task A is close to task B" and as a result, the conclusions do not suffer from the violation of the triangle inequality. We only say things like "training on task A is equivalent to training for 80% progress on task B, or "training using contrastive learning is equivalent to training using supervised learning for 25% of the progress".

D. Imprinting as an alternative to training the final layer

Consider a total of C classes. We would like to find weights $\{w_c\}_{c=1}^{C}$ that maximize the log-probability of the samples, under the constraint that for all $c \in C$, the norm of the weights $||w_c||$ is 1. Let $\varphi(x)$ denote a internal representation of sample x. The log-probability

$$\sum_{x:y_x=c}\log p(y=c\mid x) = \sum_{x:y_x=c} w_c\cdot\varphi(x) - \sum_{x:y_x=c}\log\left(\sum_{j=c}^{C}\exp\left(w_c\cdot\varphi(x)\right)\right), \tag{12}$$

is proportional to the inner-product $w_c\cdot\sum_{x:y_x=c}\varphi(x)$. Maximizing just this term under the norm constraint, we get the imprinted weights $\sum_i\phi(x_i^c)/||\sum_i\phi(x_i^c)||$ as the solution. Deriving an analytical expression for the optimal value of $\{w_c\}_{i=1}^{n_c}$ is difficult and hence we use the imprinted weights as an approximate solution. In our experiments, we found that the imprinted weights achieve an accuracy close to the optimal weights while being significantly easier to compute.

E. Invariant transformations of the internal representation

The internal representations are invariant to orthogonal transformations provided that we use imprinting to define a probabilistic model. This is because the internal representations define the same probabilistic model ever after an orthogonal transformation. Consider two internal representation ϕ and $U\cdot\phi$ where U is an orthogonal matrix. We note that the

probabilistic model for $U \cdot \phi$ after imprinting is

$$\log p_2(y = c \mid x_i) = \frac{U \cdot \sum_{y_x=c} \phi(x)}{||U \cdot \sum_{y_x=c} \phi(x)||} \cdot (U \cdot \phi(x_i)) - \log \left(\sum_{c=1}^{C} \exp \left(\frac{U \cdot \sum_{y_x=c} \phi(x)}{||U \cdot \sum_{y_x=c} \phi(x)||} \cdot (U \cdot \phi(x_i)) \right) \right)$$
$$= \frac{\sum_{y_x=c} \phi(x)}{||\sum_{y_x=c} \phi(x)||} \cdot \phi(x_i) - \log \left(\sum_{c=1}^{C} \exp \left(\frac{\sum_{y_x=c} \phi(x)}{||\sum_{y_x=c} \phi(x)||} \cdot \phi(x_i) \right) \right).$$

The probabilistic model for the representation $U \cdot \phi$ is identical to the probabilistic model for representation ϕ since norms and angles are preserved under orthogonal transformations. Hence the Bhattacharyya distance between ϕ and $U \cdot \phi$ is zero.

The imprinting procedure can be thought of as removing information from the representation that is not relevant to prediction on a task. While this is true for all datasets in general, there could exist some additional structure in the data that results in more invariances (e.g., more than invariances to orthogonal transformations $O(n)$).

F. Measuring goodness-of-fit of an InPCA embedding using explained stress

We would like to measure if a k-dimensional sub-space accurately preserves the true distances. For this purpose, we define a quantity called the "explained stress" that estimates the fraction of pairwise distances in the original space that are preserved in the k-dimensional embedding. This is analogous to the explained variance in principal component analysis (PCA); but explained variance is a measure of the how well the original points are preserved in the embedding whereas explained stress approximates how well pairwise Bhattacharyya distances are preserved. If we consider the embedding to be given by first k eigen-vectors, then the explained stress (χ_k) is

$$\chi_k = 1 - \frac{\left\| W - \sum_{i=1}^{k} \Sigma_{ii} \, U_i U_i^\top \right\|_{\mathrm{F}}}{\|W\|_{\mathrm{F}}} = 1 - \sqrt{\frac{\sum_{i=k+1}^{m} \Sigma_{ii}^2}{\sum_{i=1}^{m} \Sigma_{ii}^2}}. \tag{13}$$

Note that InPCA finds an embedding that *exactly* maximizes χ_k.

G. Calculating mean trajectories

We defined the distance between two trajectories to be $\mathrm{d}_{\text{traj}}(\tau_u^{1 \to \mathrm{U}}, \tau_v^{2 \to \mathrm{U}})$, i.e., the integral of the Bhattacharyya distance between the trajectories after mapping them to the same task and re-indexing them using the geodesic. Say we wish to compare a model trained on two tasks from CIFAR-10: Cats vs. Dogs and Airplane vs. Truck. We initialize multiple models for each of these two supervised learning problems (and we do so for every experiment in this paper) and train these 10 models. We can now calculate the mean trajectory of models on a task

$$\underset{\tau_\mu^1}{\operatorname{argmin}} \frac{1}{K} \sum_{k=1}^{K} \mathrm{d}_{\text{traj}}(\tau_{u_k}^1, \tau_\mu^1).$$

This optimization problem is very challenging because the variable is a trajectory of probabilistic models in a high-dimensional space. Even if we were to split this minimization and do it independently across time, this is still difficult because the solution is the so-called Bhattacharyya centroid on the product manifold defined in (1) and cannot be computed in closed form. See (Nielsen & Boltz, 2011) for an iterative formula. We therefore simply take the arithmetic mean of the probability distributions, i.e., $P_{\mu(t)} = \frac{1}{K} \sum_{k=1}^{K} P_{w_i(t)}$. This is similar to ensembling. We use the radius of the tube around the mean trajectory, i.e.,

$$r_u = \max_k \mathrm{d}_{\text{traj}}(\tau_{u_k}^1, \tau_\mu^1)$$

to normalize distances (more precisely, we normalize using the average of the radii of the two trajectories being compared). Note that this radius depends upon time (and is computed after mapping and reindexing the trajectories). If the distance between the means of two sets of trajectories is smaller than their individual average radii, then the tubes around the means intersect each other. In such cases, one can say that the representations learned (at that time point) are not distinguishable. We next show all distances between reindexed points along the trajectories discussed in Figures 1, 7 and A1. Note that each curve gives the integrands in (5), not the integral.

H. Frequently Asked Questions

1. **These results are all intuitive and inline with literature. It is intuitive that trajectories of trained networks explore a small part of the prediction space; it is intuitive that training on one task makes progress on another task; it is intuitive that episodic meta-learning reaches the same solution as that of supervised learning, etc.**

   Researchers working in the many sub-fields of machine learning discussed in this paper have various intuitions as to why their algorithms work. There are also numerous empirical results and theoretical models in these fields that are consistent with our findings. But this does not necessarily mean that we understand the underlying phenomena well. We must precisely quantify these intuitions and folklore results.

   For example, it may be intuitive that episodic learning is "similar" to supervised learning because while the former uses a clustering loss over different sets of classes in each mini-batch, the latter uses a cross-entropy loss, which as we discussed in the main paper, will also lead to a good clustering of the features. Such intuition has been borne out in empirical results as well: fine-tuning-based few-shot learning methods have similar accuracy as that of episodic meta-learning-based ones. One may even argue intuitively that since these methods give a similar accuracy, the representations learned by these methods must be similar.

   But intuition is a double-edged sword. For example, intuition also suggests that deep networks with millions of weights fitted on a non-convex energy landscape on seemingly dissimilar tasks are unlikely to learn similar representations. But since transfer learning is so remarkably effective, it is also intuitive that the representations of different tasks are similar; after all the network trains on the target task quickly after being pre-trained on the source task. Since networks are initialized randomly at different locations in the weight space, there is no reason to expect that the trajectory in prediction space will be low-dimensional. And yet, one might reason that since all these networks are trained on the same dataset, their training trajectories have to be similar to each other...

   And this is why—to ground such intuition—we need precise quantitative studies. We have developed sophisticated techniques using information geometry to bear upon this problem. In some cases, we find surprising results—this opens up new avenues for theoretical and empirical investigation. In some cases our results may be consistent with the intuition of the practitioners—this lends credence to these techniques.

   We are not aware of existing results in the literature which point out the phenomena identified in our paper.

2. **Do these findings translate to other tasks and other algorithms?**

   We have used nine different tasks from ImageNet with very dissimilar classes and many sub-tasks of CIFAR-10, and five different representation learning methods (supervised learning, fine-tuning, episodic meta-learning, contrastive and semi-supervised learning).

   The technical tools developed in this paper can indeed be used to study many other problem domains, algorithms and research questions. Each of these come with their own challenges, e.g., for tasks in natural language processing, although the probabilistic model is well-defined, the output space and the number of "samples" (words, sentences etc) is extremely large. As a rough estimate, for a typical book with $N \approx 10^5$ words, the output space has $C \approx 10^{20}$ if the model predicts the next 5 words. Therefore, although techniques developed here are well-defined, implementing and using them to answer new questions will require new research ideas. We hope that the research community will use these techniques to understand learned representations in the future. To aid this, we commit to releasing all code, data (model checkpoints, model embeddings), and interactive visualizations with the final version of the paper.

3. **How faithful are the visualizations of the trajectories in the prediction space?**

   We are using a dimensionality reduction technique (InPCA) which has the specific property that if one uses all the eigenvalues (i.e., the dimensionality of the embedding space is equal to the number of probabilistic models being embedded) then the pairwise Bhattacharyya distances between points are preserved exactly. When we visualize the top three dimensions of the embedding, we only see a partial picture of the manifold and pairwise distances between points are no longer preserved. We can use the expression for explained stress (11) and (13) to estimate how faithful our visualizations are. In our experience, the explained stress of the top three dimensions is always extremely high and this number is similar for all experiments in the paper. For example, the 2430 models of ImageNet in Figure 1 lie in a 10^7-dimensional embedding space and yet the top 3 dimensions have an explained stress of about 80%; the explained stress of the top 2430 dimensions of the embedding would be 100%.

   The visualizations in our paper are only used for providing intuition and interpretability; all findings in the paper are made quantitatively on the basis of the integral of the Bhattacharyya distances between points along the trajectories.

4. **Can you explain the time re-parameterization using the geodesic? It may not be a good normalization in all situations.**

   Different models train at very different speeds, in particular at the beginning of training models move rapidly after each mini-batch update in the space of probability distributions (e.g., as measured by the Bhattacharyya distance). We can only record the trajectory at specific checkpoints (e.g., after each epoch or, at best, after each mini-batch). Our time re-parameterization technique allows us to interpolate two successive checkpoints using the geodesic that connects them. We discretize "progress" t_w into 100 equidistant intervals and interpolate the entire trajectory at these 100 points by calculating the appropriate points on the geodesic between two successive checkpoints; this is described in the narrative after (4). For initial parts of the trajectory, there are fewer checkpoints per unit progress because the network trains quickly. But the time re-parameterization still allows us to discretize the entire trajectory evenly. When we do the analysis using the distances between trajectories, this technique ensures equal importance to early and late times of the training process irrespective of how fast the network learns in these phases. Our techniques also allow us to quantitatively study the differences in how quickly different networks train, although this is not the focus of this paper.

5. **CIFAR10 is a very small dataset to apply semi, self-supervised learning.**

   Self-supervised learning models have also been trained on CIFAR-10 by a number of papers (Huang et al., 2021; Zhong et al., 2022; Shen et al., 2022). Contrastive learning has also been applied to smaller datasets in the context of unsupervised domain-adaptation (Ruan et al., 2021). In our experiments, SimCLR trains to an accuracy of 87.36% on CIFAR10 (after imprinting the final model), which is only about 7% worse than a typical run of supervised learning in spite of not using any labels whatsoever.

   For Result 5 on contrastive and semi-supervised learning, we chose CIFAR-10 for contrastive learning because it is incredibly expensive to train on ImageNet. Zbontar et al. (2021) report that their method requires 32 V100 GPUs over 124 hours. For all experiments in our paper, we train on 5 random seeds (to create trajectories from different initializations, interpolate them using geodesics, compute averages etc.). Running SimCLR on ImageNet to perform an analysis of its trajectories would be enormously expensive (about $60,000 on AWS for 5 trajectories of one algorithm...).