An Information Geometric Understanding of Deep Learning

Pratik Chaudhari
Email: pratikac@seas.upenn.edu

I am interested in understanding how intelligent agents represent the structure of the physical world as they learn to perform different kinds of tasks. I have utilized this understanding to build machine learning systems that can be trained using very few resources and that can learn from few labeled data. This is a cross-disciplinary effort and spans techniques from information theory, differential geometry and statistical physics, in addition to learning theory and optimization.

Most of my work is published in machine learning conferences: International Conference on Learning Representations (ICLR), Conference on Neural Information Processing Systems (NeurIPS) and International Conference on Machine Learning (ICML). The first two are among the top 10 venues across all fields of science and engineering in terms of the h5-index. I have also published in broad readership journals such as Proceedings of the National Academy of Sciences (PNAS).

I will next describe my scientific accomplishments at Penn and elaborate upon my future plans.[*]

1 The geometry of data and its role in learning

Deep neural networks have many more parameters than the number of training data. They could therefore overfit, i.e., make inaccurate predictions outside the training set. And yet, these networks make remarkably accurate predictions—defying accepted statistical wisdom. Training these networks involves solving a high-dimensional, large-scale and non-convex optimization problem and should be prohibitively hard. And yet, training is tractable—even easy. A satisfactory explanation to these stark paradoxes is largely missing. This is because existing theories, e.g., based on over-parameterization, inductive biases such as convolutional or attention layers, stochastic optimization, etc. overlook the role of data. My work has shown how it is really the properties of natural data that predominantly make deep networks so effective.

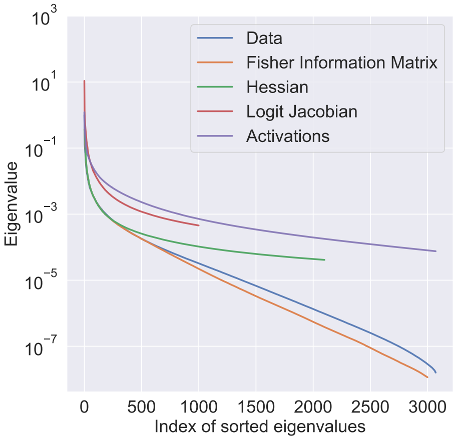

Figure 1: Sloppy models exhibit a geometric decay in the eigenspectra of the Fisher matrix (orange), the Hessian (green), and many other types of correlations. This mirrors the decay of the eigenvalues of the input correlation matrix (blue). This decay is reminiscent of the power law decay of energy in images, which is a consequence of there being objects in the physical world. The taller head and geometric decay are usually seen in curated datasets that typically have fewer nuisance variations.

Generalization in deep learning can be explained using the fact that natural data lies on low-dimensional manifolds[1–3]
I showed that the ability of deep networks to avoid overfitting can be explained by a certain characteristic structure called "sloppiness". The signature of sloppiness is a Fisher Information Matrix (FIM) with eigenvalues that decay geometrically over a very large range. This indicates a large degree of redundancy in the learned parameters. There is one set of parameters that is tightly constrained by the data, another which can vary twice as much without affecting predictions, and so on. See Fig. 1. I showed that deep networks exhibit sloppy structure because natural data has correlations that decay geometrically, i.e., data spans a very small fraction of the embedding space.

Existing techniques to characterize generalization break down—they become vacuous—when applied to deep networks. I obtained the first analytical non-vacuous PAC-Bayes bound on the generalization error of

[*] All research discussed here was conducted by students in my group at Penn, with me as the senior author.

general deep networks by exploiting the low dimensionality of input data.[1,2] Conversely, deep networks can overfit when input data is not low-dimensional. The structure in the data ensures that the effective dimensionality of the network is extremely small—less than 0.5% out of the millions of degrees of freedom (weights) govern predictions. This work shows that empirical phenomena such as wide/flat minima[4,5] and implicit regularization of stochastic training methods[6] identified in my prior work can be explained by the low-dimensionality of the input data. This work therefore contextualizes a large body of follow-up work in deep learning that investigated these phenomena. Based on these phenomena, I had also developed extremely efficient distributed algorithms to train deep networks.[7] In the next section (Fig. 2), I will describe how my perspective on such optimization-focused questions has evolved due to this connection with the geometry of the data.

Theories of generalization often provide post hoc bounds, i.e., one uses a trained model to compute the theoretically predicted test error. This is somewhat unsatisfactory because one could also compute the cross-validation error using the trained model and ignore the theory. My recent work makes progress towards a more mechanistic theory.[3] The key idea is to characterize the difference between weight trajectories of general deep networks as they train on slightly different data, but along level sets in the energy landscape as opposed to the weight space. Much like expression of the test error of a kernel machine, this leads to an estimate of the test error that is a quadratic form of the residuals at initialization and an effective Gram matrix that is akin to a Rayleigh coefficient that depends on the Fisher information matrix integrated along the trajectory. In other words, this work pins down the mechanism of generalization in deep learning: the geometry of data restricts the volume of the hypothesis space that is explored during training.

The training process of many deep networks explores an extremely low-dimensional manifold because input data is low-dimensional[8,9] The biggest hurdle in understanding deep networks comes from the complexity of the map between the weights w and the underlying probabilistic model. I have developed new techniques in information geometry to study such maps. The key idea is as follows. Given N training samples $\{(x_i, y_i^*)\}_{i=1}^N$ and C classes, we can think of the network as a point in the product simplex $[0,1]^{C\times N} \ni \big(p_w(y=i \mid x_j)\ \forall i,j\big)$. This is a finite-dimensional slice of the manifold of infinite-dimensional quantities $p_w(y \mid x)$. If $N \sim 10^6$ and $C \sim 10^3$, it is a $\sim 10^9$-dimensional object. I have developed new techniques to recover geometry in such spaces in spite of such high dimensionality. The key idea is to isometrically embed the space of probability distributions into lower-dimensional Minkowski spaces. Using these techniques, one can calculate essentially any geometric quantity, e.g., geodesics, different types of curvatures or tubes formed by trajectories in the space of probability distributions, for deep networks in spite of their enormous dimensionality.

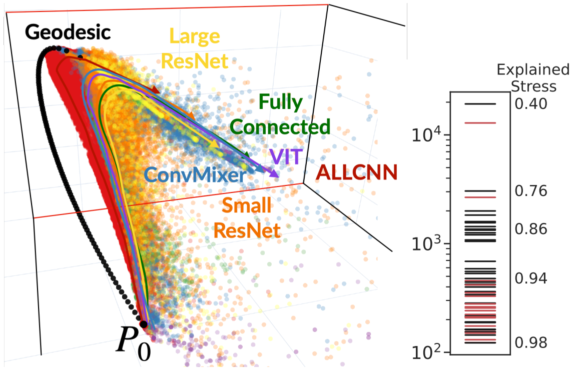

Figure 2: Models (points) with different architectures (colors), training methods (SGD, SGD with Nesterov's acceleration, Adam, different learning rates and batch sizes), and regularization (with and without batch-normalization, weight-decay, data augmentation)—there are about 150,000 different models trained on CIFAR-10 in this picture—lie on the same manifold in the space of probability distributions. Eigenvalues are spread over long ranges, so this manifold looks like a "hyper-ribbon". It is much longer than it is wider, and much wider than it is thicker. Top three dimensions visible here capture 76% of the pairwise distances between points, top 50 capture 98%.

I showed that the training process in deep learning explores an extremely low-dimensional manifold in the space of probability distributions.[8] Networks with a wide range of architectures, sizes, trained using different optimization methods, regularization techniques, data augmentation techniques, and weight initializations lie on the same manifold. See Fig. 2. Networks initialized at very different parts of the prediction space converge to the solution along a similar manifold. On this manifold, networks with different architectures follow distinguishable trajectories but other factors have a minimal influence. Larger networks train along a similar manifold as that of smaller networks, just faster. And therefore, it is not as if large deep networks work better because they predict differently than smaller ones; rather they are trained to a smaller loss using typical training recipes.

This is a very surprising finding. This work shows that the optimization problem in deep learning is extremely low-dimensional. It is proof that there exist neural architectures with extremely few parameters—as few as 3, 4, 5, . . . , 50—that can reproduce the predictions of networks with millions of weights. The basis in Fig. 2 is one such "architecture". . . except that was computed post hoc. If one can directly train and perform inference in such concise bases, then the amount of energy required for deep learning can be multiple orders of magnitude smaller than it is today. I showed, using tools in dynamical systems theory for characterizing the eigenspectrum of the reachability Gramian, that this stark low-dimensionality emerges because input data is sloppy.[9] So this is yet another demonstration of how the geometry of the data controls the performance of deep learning systems.

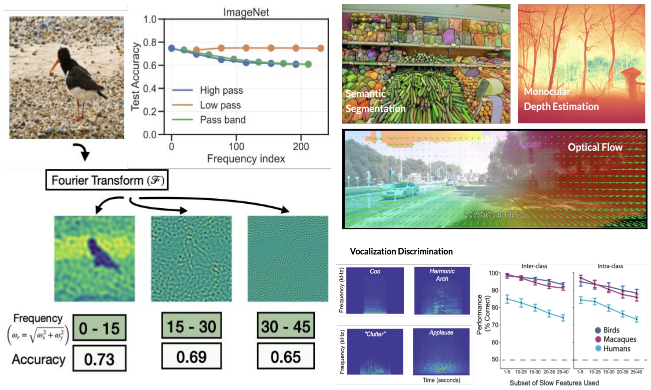

Figure 3: Left: The test accuracy of a network trained on Imagenet using images projected on Fourier frequencies with an explained variance as small as 0.06% (right-most band) is as high as 65%. Other frequency bands, and even random bands, can be used to predict categories with more than 60% accuracy. Right (top): This phenomenon is also seen for other tasks such as semantic segmentation, monocular depth, and optical-flow estimation. Right (bottom): For auditory tasks, "fast features", i.e., the ones that vary most over time, can support effective discrimination, counter to influential theories in neuroscience such as slow feature analysis.

Typical tasks are redundant functions of the input data[10] In addition to the structure in input data, there is important structure in the tasks that we perform upon the data. I showed that typical perception tasks (visual recognition, semantic segmentation, optical flow and depth estimation, or auditory discrimination) are highly redundant functions of their inputs.[10] See Fig. 3. These tasks can be predicted surprisingly accurately no matter which part of the input is used, whether inputs are projected in the principal subspace, e.g., of principal components analysis (PCA), Fourier, or wavelet bases, which contains salient variations, or whether they are projected in the tail subspace which consists of non-salient features. Even a random subspace works non-trivially well. I used a technique called partial information decomposition to show that this phenomenon is an intrinsic property of the input data and the task, not deep networks.

This finding forces us to rethink many long-standing ideas. For example, the prevalent perspective is that perception tasks are hard because solving them is like finding a needle (sufficient statistic) in a haystack (of nuisances). This finding shows that there are many near-sufficient statistics, many needles. Next, removing statistical redundancy in stimuli, e.g., by whitening, low-pass filtering, compression, sparse coding, has been argued to be fundamental in how the brain works. This cannot be the complete story because it does not account for the environment/task. If the task is redundant in the input, an organism, or a machine, need not invest in a particular feature set. Any feature set is near-sufficient for most tasks. Ontogenetic learning refines neural circuitry to improve this feature set. To crystallize this line of thinking, I am currently developing an argument that this redundancy could be—the serendipitous—consequence of the sensory apparatus adapting to a changing stream of ecological tasks.

Adaptive Riemannian optimization on the manifold of diffeomorphisms[11–13] Registration, i.e., identifying correspondences between images of, say, the brain taken across time, individuals, species etc., is the first step in almost every biomedical and life sciences pipeline. It is a good problem to showcase how optimization can exploit the underlying geometry of the problem. Like deep learning, this problem is also very ill-conditioned. But unlike deep learning, one needs to find near-global minima. Second-order methods are therefore necessary for registration. But the Hessian of an in vivo $\sim$1 mm^3 resolution brain MRI requires $\sim$15 PB to even store. The second key facet of this problem comes from the fact that Euclidean geometry is poorly suited for work on registration problems, one is usually searching for a diffeomorphism between images.

I have developed new adaptive Riemannian optimization algorithms that work on the group of diffeomorphisms.[11] They exploit the Lie group structure to take gradient steps directly on the manifold. The key idea is to use the "Eulerian" differential where the descent direction at each iteration lies in the Lie algebra at identity, not that of the current iterate. Therefore, there is no need to parallel transport the momentum and curvature information across iterations. This algorithm called Fire ANTs not only gives more accurate registrations but it is also 300× to 3000× faster than existing methods. Fire ANTs can work with extremely high-resolution ex vivo images, well below 1 µm^3. Typical methods break down below 1 mm^3 resolution. Fire ANTs can solve on one desktop GPU in minutes–hours, problems that require hours–days using current pipelines on servers with $\sim$1 TB RAM. This opens up totally new ways of analysis, tuning, as well as new applications such as building atlases of ex vivo mouse and monkey brains, registration of high-resolution histology slides, etc. Some of these would take multiple years with existing methods—which is near impossible. Fire Ants can solve these problems in 1–2 days—well within the reach of any clinician/researcher. I have also developed techniques to include this optimization-based formulation as a "layer" within a deep network.[12,13] This further improves robustness, speed and applicability to different domains.

2 The geometry of the space of learnable tasks

Almost all applications of deep learning today rely on the fact that a deep network trained on one task can be adapted easily to new tasks. In spite of such widespread usage, key questions such as "when are tasks similar to each other" or "how to best transfer a model from one to the other" are left unanswered. These are theoretical questions. But addressing them is also critical for the industry because the cost of training/updating models has become prohibitive. The central theme of my work in this direction is as follows. I argue that the flexibility of representations must be a consequence of the data—of a shared structure in the space of tasks. Because the training process does not explicitly encourage flexible representations. Characterizing this structure, can therefore help us build deep learning systems that can address a wide diversity of tasks and learn new tasks with few data.

An information-geometric distance on the space of tasks[14,15] The Information Bottleneck (IB) principle does not capture the flexibility of neural representations. Indeed, information discarded while learning the minimal sufficient statistic on one task is precisely what will be useful for a different task.[14] In order for the representation to be useful on a downstream task, the IB must be modified to force redundancy in the representation, e.g., by reconstruction of the input. This is important: masked auto-encoders (MAEs) and generative pre-trained transformers (GPTs) which power almost all applications today employ this idea. The modified IB leads to a thermodynamics of representation learning, e.g., the "first law" characterizes trade-offs between redundancy (D), sufficiency (C), and the complexity (R) of the representation. The "second law" corresponds to monotonic progress towards a convex "RDC" manifold that connects these quantities, where the free-energy is minimal.

One can define analogs of adiabatic and iso-thermal processes on the RDC manifold, e.g., an iso-classification process that keeps the classification accuracy constant when a model is adapted from one task to another. This idea can be used to define the "optimal" distance between two tasks.[15] This distance is the Riemann length of the shortest trajectory that a model trained on the source task needs to take in order to predict well on the target task. I showed how simply fitting the model on the target task (as is typically done) is sub-optimal. One must also modify the task gradually from the source to the target, e.g., using displacement interpolation in optimal transport. This distance is consistent with our intuitive notions of the ease of transfer, e.g., distance between the same pair of tasks is small if the model class is larger. It also leads to an algorithm that can optimally adapt a model to a new task.

It is widely accepted that if source and target tasks are "close" then we can learn them together or transfer across them. The merit of my work is that it formulates a general and computable definition of what "close" means.

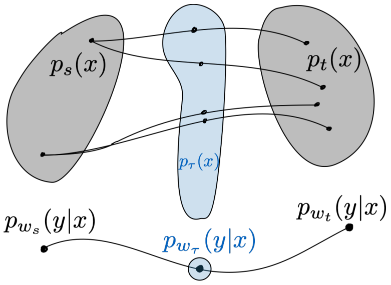

Figure 4: The optimal distance between two tasks $p_s(x, y)$ and $p_t(x, y)$ is formulated as an optimization problem that transports the input marginal $p_s(x)$ from the source task to that of the target task $p_t(x)$ with the ground-metric given by the Riemannian length of the model's trajectory from the source task $p_{w_s}(y \mid x)$ to the target task $p_{w_t}(y \mid x)$ as the model takes gradient steps on the interpolated task $p_\tau(x)$.

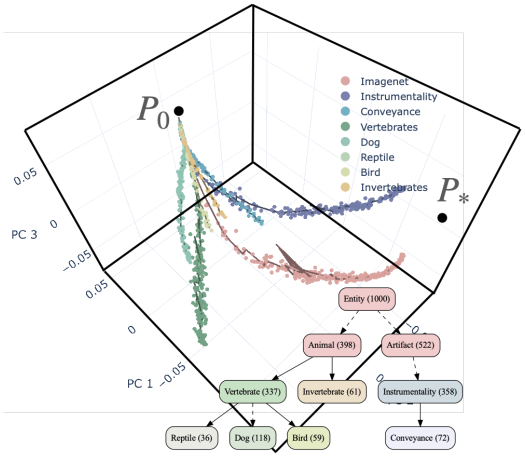

Figure 5: The manifold of models on trajectories of networks being trained on different subsets of Imagenet is also effectively low-dimensional, it looks like a hyper-ribbon. The bifurcation of training trajectories is consistent with the semantic structure in Wordnet (inset). This points to an intrinsic shared structure between images and natural language.

Typical tasks are supported on low-dimensional manifolds in the space of probability distributions[16–19] A learning task is a probability distribution on inputs and outputs. The global geometry of the space of such distributions is rather generic. But we should expect that typical tasks one encounters in practice possess a more defined geometry. The global geometry of the space of tasks is relevant for questions like "where do typical tasks lie" or "what are good priors for learning new tasks".

Using variants of the information-geometric techniques that led to Fig. 2, I showed that the manifold of probabilistic models trained on different tasks, e.g., disjoint subsets of Imagenet, is effectively low-dimensional.[16] See Fig. 5. This dimensionality is extremely small—for Imagenet, 3 dimensions of the embedding out of ∼50M dimensions capture 78.72% of pairwise distances. This points to striking shared structure among these seemingly unrelated tasks. I showed that representation learning methods, like supervised, transfer, multi-task, meta, or contrastive learning, learn very similar representations at the end of training. This makes the ideas developed above using thermodynamics and the RDC manifold more concrete, e.g., distance between tasks is just the length of the geodesic with the optimal transfer trajectory being the geodesic, the ideal pre-training objective for transfer is the bifurcation point in the tree of trajectories in Fig. 5, etc.

The most important implication of the above work is as follows. When a single network is trained on multiple tasks, one sees that synergistic/competing tasks aid/hurt each other's learning.[17] This is particularly problematic because the set of synergistic and competing tasks is different for each task. And therefore, no matter how large a network is and how it is trained, it cannot predict optimally on every task. I have developed a technique called Model Zoo to address these issues. See Fig. 6. Much like the AdaBoost algorithm which sub-samples the train set, Model Zoo builds an ensemble by selecting synergistic tasks to train a given task with. Model Zoo is the state of the art algorithm for continual learning and multi-task learning. It has spurred a large body of follow up work, including some of my own.[18,19] Today, almost all implementations of large language models (LLMs) use a mixture-of-experts architecture to fit large and diverse data corpora, which is exactly what Model Zoo argued for.

Self-supervised learning as building priors on the space of tasks[20,21] The ideal representation that can be adapted to downstream tasks easily cannot be a single network. I have argued that the ideal foundation model is a Bayesian prior over the space of tasks—a “foundation prior”.[20] I showed how one can compute such priors by maximizing the mutual information between the weights and the unlabeled data. This prior, called a reference prior, automatically selects a finite hypothesis class with complexity commensurate to the amount of labeled data. Computing the prior involves running the Blahut-Arimoto algorithm on the weight space of a deep network. This is only tractable if the space of tasks has a hyper-ribbon geometry, i.e., if data is sloppy.

Reference priors are a theory for semi/self-supervised learning. This theory is important because existing algorithms—even if they perform extremely well—are a zoo of heuristics and ad hoc choices, e.g., some enforce feature invariance to augmentations, some minimize the entropy of predictions, others use pseudo-labeling/disagreement-based losses etc. I showed how almost all of these heuristics can be derived from a single objective—that of the reference prior. Many of the practical difficulties in using existing algorithms, e.g., representation collapse during training, are mitigated with the reference prior objective.

I have developed these ideas further to show that foundation models like masked auto-encoders (MAEs) are effective because they instantiate a nonlinear basis that captures the scale of spatial correlations in the data.[21] This observation sheds light on the remarkable efficacy of MAEs, e.g., almost all implementations prefer MAEs over contrastive approaches. It also leads to effective heuristics for selecting hyper-parameters, which is important because training MAEs is extremely expensive.

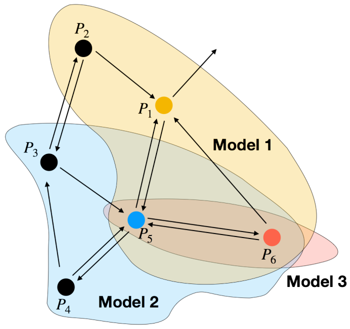

Figure 6: A single network cannot predict optimally on multiple tasks. Model Zoo fits an ensemble of models where each member is trained on synergistic tasks, e.g., Model 1 for task P_1 uses data from tasks P_2, P_5 and P_6.

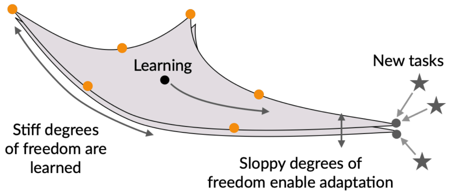

Figure 7: Reference priors are discrete priors (orange points) that are maximally spread out on the hyper-ribbon-like manifold of the space of tasks. Jeffreys prior is uniform on this manifold. This is why reference priors can be adapted with very few labeled samples to new tasks which lie near corners of the manifold.

Learning at scale with very few labeled data[22–25] The relevant limit in machine learning is when the number of samples $N \to 0$ not $N \to \infty$. As a litmus test of the theoretical understanding of the space of tasks discussed above, I have focused on such “few-shot learning” problems.

The Imagenet-21K dataset contains 14.2M images from ∼22,000 categories. About a third of the categories have less than 100 images each, a tenth have less than 10 each. Machine learning systems that need to make predictions on these rare categories often use data from frequently occurring categories to learn a representation. But such “meta”-learning methods break down at this scale. I showed that a dead-simple approach: supervised learning followed by fine-tuning to the new categories, beats any meta-learning method.[22†] It is also extremely scalable and therefore the ideal approach for building industry-scale visual recognition systems. This was a milestone result in the field. Almost all applications today implement this paradigm.

There is an important technical reason as to why this simple approach works well. My work rectified a major misunderstanding then prevalent in the field. I argued that a predictor that is good on average on all possible tasks (which is what meta-learning seeks) is ideal for none of them—certainly not for the specific test task. In other words, meta-learning is not the solution for such problems. We need to perform transductive

[†] I worked with a fresh-undergraduate researcher on this paper when I was a postdoc at Caltech/Amazon Web Services.

inference—not inductive inference. Transduction in deep learning applications was pioneered in my work and it is picking steam today. Techniques such as in-context learning used in LLMs implement transductive inference.

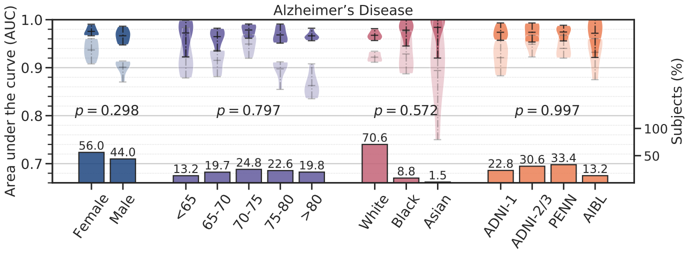

Figure 8: Models (solid violin plots) trained using structural MRI features with demographic, clinical, genetic factors and cognitive scores, on data from 5 different clinical studies are unbiased across sub-populations stratified by sex, age, race and cohort in spite of strong data imbalance (bar plots denote the sample size (%)). Results for schizophrenia and autism spectrum disorder classification are similar.

On the opposite end of the spectrum, small data is the harsh truth in medical sciences. Clinical data is highly heterogeneous with diverse anatomies, phenotypes, genomic traits, demographic/racial disease profiles, as well as operational aspects such as data acquisition protocols. One cannot predict well across the population for such heterogeneous data.[23] And that is why models can be biased. I offered a positive datapoint in this debate.[24] See Fig. 8. For Alzheimer's disease, schizophrenia and autism spectrum disorder, I showed that the bias of diagnostic models can be essentially removed if simple and well-established safeguards such as good data pre-processing, hyper-parameter tuning and rigorous model selection are utilized. Perhaps the reason for overlooking these basics is the prevalent dogma that deep networks do not overfit. This is not always true—clinical data is quite different from natural data.

In general, one has to embrace heterogeneity and explicitly adapt models to the target population.[25] I again showed that a dead-simple approach—weighted empirical risk minimization (ERM) on data from the source group and small amount of data from the target group—works substantially better than any existing approach. This is yet another instance of transductive inference—weighted ERM can be thought of as a universal procedure for domain adaptation. On 20 neuroimaging studies across the world, across two diagnostic problems: Alzheimer's disease (AD) and schizophrenia, prognostic problems (brain age prediction, early diagnosis and progression of symptoms) and across target age, sex, race and clinical cohorts, I showed that models trained with this approach are extremely performant (more than 0.95 AUC for AD diagnosis, and less than 5 years error for brain age prediction). This is the largest study of its kind and is the definitive state of the art. It makes a strong argument for machine learning-based diagnostic models of neurological disorders being ready for evaluation in clinical settings.

3 Ongoing and future research directions

I will next describe two broad research directions that I plan to focus on in the future.

Smaller is better: new neural architectures and fast methods to train them GPT-4 was trained using ∼2.5M GPU days at a cost of more than $100M, each inference call to this model runs on ∼100 GPUs. In contrast, the results in Figs. 2 and 5 suggest that ultra-small neural architectures exist that are as effective as large deep networks. I am working on discovering such architectures. I am also developing second-order optimization methods to train large transformers and multi-modal self-supervised learning models. I believe we can reduce the energy required to train and run deep learning systems by multiple orders of magnitude. This is perhaps the most important problem today. Should we expect success? I believe so. Multi-layer networks with affine-ReLU neurons are universal function approximators, but they are highly inefficient bases. Tasks in machine learning are not arbitrary distributions between inputs and outputs. They possess important structure that makes them learnable, e.g., in Fig. 3. This is why the basis in Figs. 2 and 5 is so concise.

More broadly, over the past decade, the focus has increasingly shifted away from modeling data and towards directly using raw observations. As a result, deep networks have to rediscover a lot of the structure in the data—structure that we often know a priori and could have designed for. We train models on essentially uncurated data without knowing the eventual task. We run the same model for vision, text, audio, and motor control by representing all of them as tokens. We pay for this flexibility in energy. Another undesirable consequence of this flexibility is that these systems are extremely fragile. Updating models is so risky, that even the largest models are trained from scratch. This is the second decade of deep learning. Our focus must shift to engineering deep learning systems that are performant, efficient, and reliable. And towards using these marvelous new scientific instruments to discover new fundamental facts of nature.

Long-range autonomy in complex, high-stakes, outdoor environments In my early PhD career, my work on safe urban driving at nuTonomy (now Hyundai-Aptiv Motional) helped launch the world's first autonomous taxi service in Singapore in 2016. There is a spate of urgent problems that require this skill-set. The January 2025 wildfires in Los Angeles—believed to be caused by drought across multiple years and exceptional wind patterns, with power and utility equipment supplying the spark—led to about 200,000 people being evacuated and about $50 billion in losses. Fires, floods, and storms rage ever more often. Critical infrastructure for power, transport, and human habitation that lies in their path is vulnerable. It is clear that our growing demands of nature need to be more sensitive to its unalterable laws. As nature fights back, we need to fortify ourselves.

Figure 9: My future work will focus on two key areas. First, in deep networks, we have built a great microscope. We must now focus on what lies underneath, on new facts of nature that this scientific instrument allows us to study. Second, I want to marshal this understanding for building long-range autonomous systems that can operate in outdoor environments. Measuring, restoring and sustaining the vital balance between nature and human development is perhaps the most important robotics problem facing us today.

I argue that these are actually robotics problems. First, we need fine-grained measurements to understand spatiotemporal patterns in ecosystems—and how they change—at the submesoscale, i.e., over months and years. These measurements are essential to complement coarse satellite data, and physics-based models that typically operate at larger spatial and temporal scales. Gleaning these measurements from vast terrains such as forests, farms, and rivers requires that robots can operate autonomously for very long periods of time in unstructured environments. Second, we need robots for preventative maintenance on infrastructure (e.g., power, pipelines, dams), pollution mitigation on land and water and ecological restoration tasks like revegetation.

I have begun work in this direction over the past two years. These problems demand fundamental advances in robot autonomy for outdoor environments, which will be the core focus of my work. They also require progress in sensing, actuation and communication technologies. Robots are largely built for and tested in controlled indoor settings today. Historically, when this barrier has been breached, robotics has taken quantum leaps, e.g., DARPA challenges on driving, humanoids, subterranean exploration etc.

4 Research group and funding sources

I currently advise 13 students on their doctoral research (5 ESE, 6 CIS, 1 AMCS and 1 Physics). Four students from my group have graduated with a Ph.D so far (1 ESE, 2 AMCS and 1 CIS). My research is funded by grants from the National Science Foundation (NSF) via programs such as CAREER, MoDL (Mathematics of Deep Learning), AI Institute on Artificial and Natural Intelligence (ARNI), National Robotics Initiative (NRI) and Cyber-Physical Systems (CPS). I have also received funding from the Office of Naval Research (ONR), Amazon Machine Learning Research Award, Intel Rising Star Faculty Award, and gifts from external sources. Altogether, I have received federal grants totaling $5.2M as PI, $36.5M as co-PI, and gifts/internal grants totaling $0.8M.

References

1. Yang, R., Mao, J., **Chaudhari, P.,** Does the Data Induce Capacity Control in Deep Learning? in Proc. of International Conference of Machine Learning (ICML) (2022).
2. Chen, D., Chang, W., **Chaudhari, P.,** Learning Capacity: A Measure of the Effective Dimensionality of a Model. arXiv preprint arXiv:2305.17332 (2023).
3. Yang, R., **Chaudhari, P.,** An Effective Gram Matrix Characterizes Generalization in Deep Networks. arXiv preprint arXiv:2504.16450 (under review at NeurIPS) (2025).
4. **Chaudhari, P.,** Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L., Zecchina, R., Entropy-SGD: Biasing Gradient Descent into Wide Valleys in Proc. of International Conference of Learning and Representations (ICLR) (2017).
5. **Chaudhari, P.,** Oberman, A., Osher, S., Soatto, S., Carlier, G., Deep Relaxation: Partial Differential Equations for Optimizing Deep Neural Networks. Journal of Research in the Mathematical Sciences (RMS) **5,** 1–30 (2018).
6. **Chaudhari, P.,** 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).
7. **Chaudhari, P.,** Baldassi, C., Zecchina, R., Soatto, S., Talwalkar, A., Oberman, A., Parle: Parallelizing Stochastic Gradient Descent in Conference on Machine Learning and Systems (MLSys) (2018).
8. Mao, J., Griniasty, I., Teoh, H. K., Ramesh, R., Yang, R., Transtrum, M., Sethna, J. P., **Chaudhari, P.,** The Training Process of Many Deep Networks Explores the Same Low-Dimensional Manifold. Proceedings of the National Academy of Sciences (PNAS) **121,** e2310002121 (2024).
9. Mao, J., Griniasty, I., Sun, Y., Transtrum, M. K., Sethna, J. P., **Chaudhari, P.,** An Analytical Characterization of Sloppiness in Neural Networks: Insights from Linear Models. arXiv preprint arXiv:2505.08915 (under review at Physics Review E) (2025).
10. Ramesh, R., Bisulco, A., DiTullio, R. W., Wei, L., Balasubramanian, V., Daniilidis, K., **Chaudhari, P.,** Many Perception Tasks Are Highly Redundant Functions of Their Input Data in Computational and Systems Neuroscience (COSYNE) (2025).
11. Jena, R., ***Chaudhari, P.,** *Gee, J. C., FireANTs: Adaptive Riemannian Optimization for Multi-Scale Diffeomorphic Registration. arXiv preprint arXiv:2404.01249 (under review at Nature Comm.) (2024).
12. Jena, R., Sethi, D., ***Chaudhari, P.,** *Gee, J. C., Deep Learning in Medical Image Registration: Magic or Mirage? in Proc. of Neural Information and Processing Systems (NeurIPS) (2024).
13. Jena, R., ***Chaudhari, P.,** *Gee, J. C., Deep Implicit Optimization for Robust and Flexible Image Registration. Medical Image Analysis (2025).
14. Gao, Y., **Chaudhari, P.,** A Free-Energy Principle for Representation Learning in Proc. of International Conference of Machine Learning (ICML) (2020).
15. Gao, Y., **Chaudhari, P.,** An Information-Geometric Distance on the Space of Tasks in Proc. of International Conference of Machine Learning (ICML) (2021).
16. Ramesh, R., Mao, J., Griniasty, I., Yang, R., Teoh, H. K., Transtrum, M., Sethna, J., **Chaudhari, P.,** A Picture of the Space of Typical Learnable Tasks in Proc. of International Conference of Machine Learning (ICML) (2023).
17. Ramesh, R., **Chaudhari, P.,** Model Zoo: A Growing "Brain" That Learns Continually in Proc. of International Conference of Learning and Representations (ICLR) (2022).
18. *De Silva, A., *Ramesh, R., Priebe, C. E., ****Chaudhari, P.,** **Vogelstein, J. T., The Value of Out-of-Distribution Data in Proc. of International Conference of Machine Learning (ICML) (2023).

19. *De Silva, A., *Ramesh, R., *Yang, R., **Vogelstein, J. T., __**Chaudhari, P.,__ Prospective Learning: Learning for a Dynamic Future in Proc. of Neural Information and Processing Systems (NeurIPS) (2024).

20. *Gao, Y., *Ramesh, R., **Chaudhari, P.,** Deep Reference Priors: What Is the Best Way to Pretrain a Model? in Proc. of International Conference of Machine Learning (ICML) (2022).

21. Bisulco, A., Ramesh, R., Balliestro, R., **Chaudhari, P.,** From Linearity to Non-Linearity: How Masked Autoencoders Capture Spatial Correlations in Proc. of International Conference on Computer Vision (ICCV) (2025).

22. Dhillon, G. S., **Chaudhari, P.,** Ravichandran, A., Soatto, S., A Baseline for Few-Shot Image Classification in Proc. of International Conference of Learning and Representations (ICLR) (2020).

23. Wang, R., **Chaudhari, P.,** Davatzikos, C., Embracing the Disharmony in Medical Imaging: A Simple and Effective Framework for Domain Adaptation. Medical Image Analysis (2021).

24. Wang, R., __*Chaudhari, P.,__ *Davatzikos, C., Bias in Machine Learning Models Can Be Significantly Mitigated by Careful Training: Evidence from Neuroimaging Studies. Proceedings of the National Academy of Sciences (PNAS) **120,** e2211613120 (2023).

25. Wang, R., Erus, G., __*Chaudhari, P.,__ *Davatzikos, C., Adapting Machine Learning Diagnostic Models to New Populations Using a Small Amount of Data: Results from Clinical Neuroscience. arXiv preprint arXiv:2308.03175 (2023).