Research Statement

The goal of my research is to develop a better understanding of how and when machine learning succeeds and the broader societal implications of data-driven systems. My work lies at the intersection of machine learning, optimization, and statistics, including topics such as high dimensional learning and algorithmic explainability.

The remarkable and continuing empirical success of modern machine learning, especially deep learning, has opened up several foundational questions about our understanding of why learning models succeed. Answering these questions have the potential to expand the theoretical and algorithmic frontiers of learning and optimization. At the same time, as machine learning proliferates into the real world, it is of profound importance to critically evaluate the broad societal impacts of these systems. I am interested in developing a theoretical understanding of learning models that will lead to principled methods for building systems with a positive social impact.

My recent research spans two lines of work that contribute to these broad themes. My primary line of work develops new theories towards understanding the role of optimization in the success of modern machine learning models, especially deep neural networks. My second line of work is on formulating algorithms that aid from theoretical understanding. Hence paves the way for more theoretically grounded techniques that could potentially help solve the barriers of current deep learning algorithms, to name a few - Generalization and Catastrophic forgetting.

My recent work includes understanding the significance of layer-wise loss landscape and the criticality of their distances from the model's overall loss landscape.

An open-ended research question that I tried addressing recently is the problem of catastrophic forgetting in Continual learning algorithms and deep learning in general. Catastrophic forgetting is the problem of the models forgetting the previously trained data, which is very unlikely of a human brain. In other words, human brains are very good at effectively forgetting unnecessary data. As straight forward it may seem, it is not very easy to translate to machine learning models, where catastrophic forgetting can be extremely severe to a point it almost forgets everything.

After an in-depth survey of existing literature, it was very prominent that the entire continual learning community was taking the freeze the important neurons or regularize path. These methods of tackling the problem of catastrophic forgetting showed a negligible performance boost compared to just using samples of old data. Thus it became necessary to understand why things behave the way they do and check if there is something fundamentally wrong in the community's approach towards the continual learning setting.

I began by establishing a way of looking at the loss landscapes in a continual learning setting, which I later figured out that cannot be viewed in the same way as normal learning's loss landscape. In Incremental learning, at any given point, 3 loss landscapes should be considered - a landscape of old data (L1), new data (L2), and ideal loss landscape of a model trained with both old and new data together (L3). If a model is trained on old data, the optimizer traverses on L1 and settles at local minima. When the same model is incrementally trained on new data, catastrophic forgetting is imminent, and the optimizer starts to traverse on L2 and settles at local minima. Now, the reason for the poor performance is that local minima at L2 may not be a minima at L1 at all. But there is a region in the loss landscape where the minimas of L1, L2, and L3 coincides. This is the minima that we strive for our optimizer to settle down at. All the regularization techniques that exist in the literature which aim to preserve the synaptic connections, basically take a well-distinguished path towards this region of intersection of L1, L2, and L3 "implicitly" without getting stuck at local minima of other regions. This perspective and understanding is a community first, and I believe, is a step in the right direction. Further details on this work and several other insights I arrived at using various theoretical frameworks (to name a sample - Eigen Spectrum of Hessian Decomposition) will be published in my upcoming work.

As I have already mentioned an open-ended problem that I have worked on, I am restricting myself from summarizing/critiquing a paper and just citing a few works that I found interesting recently (in no particular order) -

- Emergent properties of the local geometry of neural loss landscapes - Stanislav Fort
- Large Scale Structure of Neural Network Loss Landscapes - Stanislav Fort
- Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs - Timur Garipov
- Deep Ensembles: A Loss Landscape Perspective - Stanislav Fort
- Fluctuation-dissipation relations for stochastic gradient descent - Sho Yaida
- Kernel and Deep Regimes in Overparametrized Models - Suriya Gunashekar
- The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks - Jonathan Frankle
- Deconstructing Lottery Tickets: Zeros, Signs, and the Supermask - Jason Yosinski
- Supermasks in Superposition - Mitchell Wortsman
- LCA: Loss Change Allocation for Neural Network Training - Jason Yosinski
- Shared Representational Geometry Across Neural Networks - Qihong Lu
- On the Expressive Power of Deep Neural Networks - Maithra Raghu
- Reconciling modern machine learning practice and the bias-variance trade-off - Mikhail Belkin