1. Title: Relations Between Image Classification Adversarial Attacks And Finite-Time Lyapunov Exponents

2. Specific Aim: I will test the hypothesis that fast gradient sign attacks [1] that push an input across the decision boundary into a different attractor basin for images in the MNIST handwritten number dataset, will appear on a ridge where there is a rapid change in the Maximal Finite-Time Lyapunov Exponents (FTLE) from negative to positive.

3. Execution Plan & First Steps:

- **Software/Toolkit:** Python with the PyTorch framework for network construction.
- **Initial Milestone (Weeks 5-6):** Re-implement the method in [2] to compute the maximal-FTLE for classification of the MNIST dataset using a feed-forward fully connected deep neural network. The goal is to replicate the MNIST: maximal-FTLE field, test error, and mean/standard maximal-FTLE plots seen in [2].
- **Core Task (Weeks 7-9):**
  1. Replicate the image-based, fast gradient sign adversarial attacks seen on MNIST data in [1].
  2. Integrate the replicated adversarial attack code with the feed-forward network from [2].
  3. Visualize effective adversarial attacks on the maximal-FTLE field plot.
  4. Compute correlation metrics (r-values) between each attack's maximal-FTLE field location and the nearest ridge point to that attack for each of the numbers.
- **Plan B:** If little correlation is seen between the adversarial attack maximal-FTLE locations and the maximal-FTLE ridge locations, then I will pivot to a diagnostic analysis. I will systematically measure the gradient norm and network activity of the attack images to pinpoint why the correlation is poor.

4. "Minimum Viable Product": The key results will be a figure and a table. The figure will visualize the maximal-FTLE field positions of the adversarial attacks and the ridges on a 2D plane projected down from, effectively, the last hidden layer of the network. The table will show the corresponding correlation metrics between the maximal-FTLE field locations for each attack and all of the numbers' nearest ridge points to that attack.

5. "Stretch" Goal: I will first try to incorporate additional FTLE into ridge field location computations to see if the correlation metrics can be improved. After this, I will attempt to "flip the script" and generate adversarial attacks from knowledge of the FTLE field ridges. The classification error percentages will be compared between the FTLE and the fast gradient sign attacks for each of the numbers.

6. Initial Steps & Preliminary Results (Optional): I re-implemented both the maximal-FTLE computation in [2] and their toy example, where an x-y position was fed into a neural network that classified any given point as being inside or outside some circle. The visualization of my toy example's maximum FTLE field looks exceedingly similar to the one in [2].

References

[1] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[2] L Storm, Hampus Linander, J Bec, Kristian Gustavsson, and Bernhard Mehlig. Finite-time lyapunov exponents of deep neural networks. *Physical Review Letters*, 132(5):057301, 2024.