Inductive Biases of SGD: Research Proposal

Motivation

Stochastic Gradient Descent (SGD) (and its variants) is a central algorithm driving deep learning. SGD is often modeled mathematically using simplified stochastic differential equation (SDE) approximations. A common assumption is to model SGD as some Langevin dynamics but it is empirically invalid during actual deep neural network training. Specifically, standard assumptions for Langevin that are not satisfied:

- **Gradient Noise Anisotropy**: Noise covariance depends on the geometry of the loss surface.
- **Heavy-tailed Gradient Noise**: Empirically violated due to observed heavy-tailed gradient distributions.
- **Detailed Balance (Curl-Free Probability Current)**: In general we expect a non-zero curl in probability currents.
- **Small Learning Rate Approximation**: Empirically violated by observed edge-of-stability phenomena with oscillatory dynamics.

These violations imply that the distribution of SGD trajectories significantly deviates from the tempered posterior (Gibbs distribution) resulting from Langevin models of SGD.

To address these shortcomings, we propose to investigate refined SDE models incorporating key phenomena:

- **Levy Noise with Anisotropic Covariance**: Capturing heavy-tailed and anisotropic gradient noise empirically observed during training.
- **Central Flow Term**: Capturing oscillatory dynamics and sharpness-induced flows from edge-of-stability phenomena.
- **Fractional time Fokker Planck derivative** for subdiffusive behaviour.

Theory project: It would be useful to write a clear paper on SDE models of SGD and the deviations from the Langevin regime. The paper would introduce a more general model of SGD that includes central flow, Levy noise, anisotropic covariance, fractional time Fokker Planck derivative and rotational probability current. The paper could also review the literature on how these different assumptions matter for SGD.

Generally speaking understanding the inductive biases of SGD is likely to be important for AI alignment. More specifically, accurately sampling the SGD trajectory distribution would be helpful for understanding the robustness of safety properties. Such an approach will aid in determining the likelihood of model trajectories remaining aligned and safe upon further training, as opposed to bifurcating toward deceptive or misaligned models (or being already deceptive if most models sampled on the degenerate set of loss minimizers contain many deceptive models).

Empirical directions

Objective: Validate our refined SDE model by comparing the distributions it generates against actual SGD trajectory distributions and those induced by a Langevin SDE approximation of SGD.

It seems particularly interesting to investigate the effect of noise anisotropy and heavy-tailed gradient noise on the saddle to saddle dynamics in DLNs. The saddle-to-saddle dynamics in DLN is a well studied phenomnon that has sufficiently rich dynamics close to more realistic deep learning models (e.g. feature learning).

The importance of stochasticity in the dynamics of SGD for learning is particularly interesting. So I suggest to let the central flow factor aside for some later study. Additionally, it seems plausible that mixing time are too long for SGD to reach the stationary distribution so we could also let the curl-free probability current aside for some later study. Although keep in mind that stationary distribution might be relevant on subspaces of the parameter space over which mixing time are long enough (e.g. maybe because SGD explores along degenerate directions). We could also look at the effect of the central flow term on the saddle to saddle dynamics.

Setup: Deep Linear Networks (DLNs) We could focus on DLNs, as they offer more analytically tractable analyses and known characterizations of degenerate minima (see Simon paper). It is possible to have interesting saddle-to-saddle dynamics in DLNs which resemble feature learning in deep neural networks. Furthermore, it seems that SGD favours low rank solutions during training of deep linear networks. We want to understand deviation from the Langevin regime in the SDE model of SGD on the saddle to saddle dynamics by investigating the effect of noise anisotropy and heavy-tailed gradient noise around Saddle points.

Experimental Protocol.
Initialization:

- Saddle to saddle initialization: Initialize parameters with a small covariance such that the initial parameters are close to a saddle point.
- More lazy NTK regime: Initialize parameters with a large covariance such that initial parameters can be close to a global minimum.

Data generation:

- Generate synthetic data as random matrices: can try normal or heavy tailed random matrices.

Trajectory Generation:

- **SGD Trajectories**: Simulate true SGD with mini-batches of synthetic data.
- **Langevin SDE**: Standard Brownian-motion-driven Langevin dynamics with isotropic Gaussian noise and no central flow.
- **Refined SDE**: Proposed SDE with Levy-stable noise (heavy tails), and anisotropic covariance (rotational current). Potentially look at central flow and fractional time Fokker Planck derivative.

Measuring distance between SGD and SDE distributions

- Estimate some distance between distributions obtained from SGD and the various SDE models. **Question: What notion of distance would be tractable and relevant?** e.g. KL divergence might be difficult to estimate.
- Quantify the shape of distributions (e.g. covariance, skewness and higher moments).

Suggestions of Distance Metrics for Comparison: This is something to investigate in more details.

- **Wasserstein-2**: Measure of distributional differences (connections with optimal transport?)
- **Maximum Mean Discrepancy (MMD)**: Non-parametric kernel-based measure.
- **Approximate KL Divergence**: can we estimate some empirical version?

Tracking escape dynamics from saddle points. We could test wether noise anisotropy and heavy tailness of SDE with Levy process (using some alpha stable distribution) helps to predict escaping from saddle points better compared with a Gaussian SDE. Process:

- Detect saddle points (for at loss plateaus, monitor gradient norm and hessian eigenvalues as one should be negative).
- Collect gradient noise around saddle points from SGD and SDE models (naive vs refined)
- Track tail index indicator e.g. Hill estimator. Hypothesis: Heavy tailness correlates with faster escape from saddle points.
- We could look at escape time from saddle point with anistoropic heavy tail noise vs Langevin SDE and SGD.

Broad Research Plan

Early draft for now. Some tasks and goals that we might do:

- Read literature on the relevance of noise anisotropy and heavy tailness of gradient noise for deep learning
- Validate that DLNs are interesting toy models for our project. Otherwise explore other toy models (simple low dimensional loss landscapes, toy models of superpositions, toy models of computations e.g. modular addition, parity learning, etc.)
- Read literature on DLNs
- Understand how to generate processes with alpha stable distribution
- Scope experiments (especially with Avi)
- Code DLN setup
- Being able to run SGD and SDE training dynamics on DLNs (**Is it computationally difficult to run many SGD and SDE trajectories on DLNs?**)
- Being able to detect saddle points in DLNs
- Measuring distance between distributions of SGD and SDE models
- Tracking escape dynamics from saddle points
- Computing escape time from saddle points
- Figuring out good test statistics to monitor effect of heavy tailness and noise anisotropy on escape dynamics from saddle points
- Figuring out test statistics to measure moments of the SGD and SDE distributions
- Running many experiments to get robust statistics and compare distributions between SGD and SDE models

- Interpreting results
- Writing up general paper on SDE models of SGD: deviations from the Langevin regime and SLT
- Writing up a paper toward the end of the fellowship from our empirical results
- Presenting our work at the Illiad conference

Week	Dates	Main Goals	Key Tasks
1	June 30–July 6	Prepare experiments	- Read Deep Linear Networks literature - Read relevant literature on noise anisotropy and heavy tailness of gradient noise - Scope experiments with Avi - Validate that DLNs are interesting toy models for our project - Start Coding DLN setup
2	July 7–July 13	Run SGD and SDE experiments and DLNs	- Running SGD and SDEs on DLNs - Reading more literature on DLNs and influence of SGD noise on saddle to saddle dynamics - Figuring out test statistics to monitor effect of heavy tailness and noise anisotropy on escape dynamics from saddle points (literature will be useful here)
3	July 14–July 20	Get some qualitative results from experiments	- Get some summary statistics to compare distributions between SGD and SDE models - Make progress on general paper on inductive biases of SGD
4	July 21–July 27	Experiments	- Estimate test statistics to monitor effect of heavy tailness and noise anisotropy
5	July 22–July 28	Experiments	- Mid-point review
6	July 29–Aug 4	Experiments	- More experiments - Writing up methods and background for paper - General paper on inductive biases of SGD

7	Aug 5–Aug 11	Interpret results	- More experiments - Writing up paper
8	Aug 12–Aug 18	Experiments	- More experiments
9	Aug 19–Aug 25	Experiments	- More experiments - Prepare talk for Illiad conference
10	Aug 26–Sep 1	- Illiad conference	