# Inductive Biases of SGD: Research Proposal

Guillaume Corlouer

July 8, 2025

## Motivation

If SGD approximates Bayesian inference, we can leverage the rich literature on Bayesian statistics to obtain more rigorous asymptotic guarantees for AI safety. For instance the local learning coefficient from singular learning theory enables us to understand the asymptotic behavior of in-distribution generalization error in deep neural networks by measuring the model's effective complexity. Furthermore, we can derive higher-order statistical moments of the distribution using the model's free energy. Additionally, being able to sample from the distribution of SGD trajectories would prove valuable for assessing the robustness of some safety property. More specifically, we could sample models from the underlying distribution of parameters within some degenerate subspace of parameters that are compatible with minimizing the loss and evaluate their safety properties. If a significant proportion of sampled models exhibit unsafe behaviors (such as deception), this would provide evidence that the corresponding training checkpoint is not robuslty safe.

The assumption that SGD approximates Bayesian inference holds when we can model SGD as Langevin dynamics and take the long time limit of SGD (Mandt 2018). This regime requires several conditions: the data must be i.i.d., the batch size must be large enough to invoke the central limit theorem, the learning rate must be small, and the noise covariance must be isotropic, finite, and non degenerate. Under these conditions, we can solve the Fokker-Planck equation of the Langevin dynamics to show that the long-run distribution of SGD converges to the tempered posterior (Gibbs distribution). The corresponding Bayesian posterior is the special case where temperature $\mathrm{T} = 1$. However, this regime does not accurately describe actual deep neural network training. Specifically, standard assumptions for the Langevin dynamics that are not satisfied:

- **Gradient Noise Anisotropy**: Gradient noise covariance depends on the geometry of the loss surface. In particular, it is sensitive to degeneracies in the loss surface which add a stickiness effect to SGD dynamics. Furthermore, the gradient noise covariance is not always invertible due to degenerate components in the loss landscape.
- **Loss landscape is degenerate** and the gradient noise covariance is non invertible at a degenerate critical point.
- **Heavy-tailed Gradient Noise**: Empirically violated due to observed heavy-tailed gradient noise distributions. The Wiener process in the SDE model of SGD could be replaced by some $\alpha$-stable process.
- **Detailed Balance (stationary distribution at equilibrium with Curl-Free Probability Current)**: In general we expect a non-zero curl in the probability currents for the stationary distribution.

- **Small Learning Rate Approximation**: Empirically having a finite non infinitesimal learning rate is important as can be seen by edge-of-stability phenomena with oscillatory dynamics.

These violations suggest that the distribution of SGD trajectories deviates from the tempered posterior resulting from Langevin models of SGD.

To address these shortcomings, we propose investigating refined SDE models of SGD that incorporate key empirical phenomena:

- **Anisotropic Covariance**: Models the anisotropic gradient noise empirically observed during training.
- **Heavy-tailed Noise generated by $\alpha$-stable Distributions**: Captures the heavy-tailed gradient noise distributions observed empirically during training.
- **Central Flow Term**: Accounts for oscillatory dynamics and sharpness-induced flows arising from edge-of-stability phenomena (see the Central Flow paper).
- **Fractional Time Fokker-Planck Derivative**: Describes the subdiffusive behavior of SGD (see Max Hennick's paper on fractal dynamics).
- **Variant of SGD with momentum and RMSprop (e.g. ADAM)**: Adam optimizers are used in LLMs training, not SGD. Langevin must be modified into an underdamped dynamics with momentum (only take momentum into account).

## Outline of the project

We want to list all the conditions under which SGD deviates from the Langevin regime. Most of these conditions have already been mentioned in the motivation section. We will prioritize the assumptions that seem most important and tractable to investigate. We will then relax some of these assumptions (e.g., noise anisotropy) to develop a refined SDE model of SGD. Using a toy setup, we will compare the dynamics of SGD, our SDE model, and Langevin dynamics during training. For a good control experimental setup, we choose deep linear networks. Deep linear networks are particularly interesting because they are more analytically and empirically tractable to study than deep neural networks with non-linear activations. For example, they have been studied extensively in the literature (see Saxe et al.) and, under appropriate initialization, they display rich saddle-to-saddle dynamics that we expect to be relevant for understanding the training dynamics of deep neural networks. Furthermore, we have some characterization of the zero set of DLNs (see Simon's paper) and some of their critical points.

For each SDE model of SGD, it will be particularly interesting to track statistics that enable us to compare it with SGD and Langevin dynamics (LD) around non-strict saddle points (saddle points with at least one flat direction). Non-strict saddle points are typical during the training of deep neural networks, so understanding SDE models of SGD around them seems particularly promising for transferring our results to non-linear deep neural networks. We want to examine various statistics to compare SGD, SDE, and LD. At a given saddle point, we can analyze the dynamics of SGD, SDE, and LD. More specifically, for each process we can examine and compare:

- The losses behaviors during training of DLNs when using each of these algorithms.
- The exit time from the saddle point via negatively curved directions

- Mixing time to quasi-stationary distributions or to stationary distributions along degenerate directions (for example, by examining the spectral gap between the first eigenvalues of the Fokker-Planck operator)
- The moments of the quasi-stationary distribution (or stationary distribution on degenerate subspaces) of each process
- The moments of the underlying processes, such as comparing the gradient noise covariance of SDE models with those of SGD and Langevin dynamics (e.g., examining singular values)
- Distributional divergences using distance metrics (e.g., Wasserstein distance or KL divergence) between the process distributions, when estimable
- Autocorrelations of the processes to measure non-Markovianity
- Entropy production rate and rotational probability current to quantify irreversibility and test detailed balance assumptions
- Mean squared displacement of parameters to characterize diffusion and identify sub-diffusive or super-diffusive regimes
- Stochastic Lyapunov exponents to measure edge-of-stability effects
- Heavy-tailedness indicators, such as the Hill estimator

Whenever possible, we will obtain both empirical results and analytical results (e.g., asymptotics) for the statistics of these processes on DLNs. In particular, we want to understand how these models differ during training by tracking these statistics across different saddle points throughout the training trajectory. It will also be interesting to vary the network depth to understand how the differences between SGD and SDEs evolve as model depth increases.

## Outcome

We want to write a report scoping different research directions to better understand the inductive biases of SGD and how they differ from the Bayesian inference regime (by examining differences with Langevin dynamics). The goal of this report is to submit it to grant-makers to secure funding for hiring experts to investigate each direction in detail. In the report, we will explain in detail each SDE model of SGD that results from relaxing one assumption underlying LD (e.g., anisotropic noise). To demonstrate that a given assumption matters, we will compute statistics comparing the dynamics of SGD, SDE, and LD both empirically and analytically on DLNs during saddle-to-saddle dynamics.

First, we will compare Langevin dynamics with anisotropic noise, which we call anisotropic Langevin dynamics (ALD). We should expect anisotropy to be important: it has been shown that the gradient noise covariance is proportional to the Hessian of the loss function at critical points, which we know to be singular (containing zero eigenvalues). Furthermore, anisotropic noise represents one of the simplest relaxations of the LD model, making it a natural starting point—especially given recent work I have conducted on anisotropic noise in low-dimensional loss landscapes. By examining various statistics, we will also better calibrate our understanding of the difficulty of this research agenda before exploring other directions. This work may lead to publishable results. For other SDE models, we can explain the model and demonstrate simple statistics around saddle points to investigate the importance of their underlying assumptions.

## Connection to AI safety

Another important direction would be to connect the study of the inductive biases of SGD with AI alignment. In particular, we want to get a clearer idea of what theoretical results would be particularly useful for scalable oversight (e.g. debate) or other safety agendas (e.g. ARC, interpretability). At the moment there are some broad directions that seem promising but I want to get a clearer picture of the safety relevance and time sensitivity of applications to AI alignment.

- **Robustness of some safety property**: With an accurate sampler of the distribution of parameters selected by SGD, we could sample models compatible with minimizing the loss and evaluate their safety propertie to test their robustness (e.g. by sampling quasi-stationary distribution).
- **Defining new observables to track during training that are analogous to SLT** but adapted to SGD and its variants. For example, we could compute a Free energy of the (out of equilibrium) distribution of parameters selected by SGD during training and derive a measure of effective complexity and other statistical moments. A particularly interesting result here would be computing the in-distribution, and out-of-distribution generalization error in terms of these statistical quantities.
- **Understanding better how SGD matters for learning algorithms**, in particular relating the inductive biases of SGD to learning algorithms step by step relative to Bayes posterior which directly "teleports" to a solution.
- **Broad safety applications**: How we can use this to include or exclude some data to make the training process more likely to converge to a safe solution.

In terms of the theory of change there are two important failure modes to avoid in this project:

- Doing work that is better to delegate to future AI as I expect they'll accelerate theory work a lot. So we want to focus on theory work that is time-sensitive.
- Doing work that is fundamental but difficult to relate to alignment for the next couple of years and that is better suited to academics already working on theory of deep learning.

## Research Plan

Tentative plan. It is possible that we focus only on one direction, depending on how quickly we can get empirical results.

| **Week** | Dates | Main Goals | Key Tasks |
|---|---|---|---|
| **1** | June 30–July 6 | Prepare experiments | - Scope experiments<br>- Start Coding DLN setup |
| **2** | July 7–July 13 | SGD and ALD on DLNs | - Running SGD and ALD on DLNs<br>- Derive test statistics: escape time and mixing time of quasi-stationary distribution |

| | | | |
|---|---|---|---|
| **3** | July 14–July 20 | SGD and ALD on DLNs | - Compare distributions between SGD and ALD models<br>- Make progress on general paper on inductive biases of SGD |
| **4** | July 21–July 27 | Heavy tailness on DLNs | - Explore test statistics to monitor effect of heavy tailness |
| **5** | July 22–July 28 | Heavy tailness on DLNs | - Mid-point review |
| **6** | July 29–Aug 4 | Rotational current and detailed balance | - Qualitative results |
| **7** | Aug 5–Aug 11 | Rotational current and detailed balance | - Writing up longer report |
| **8** | Aug 12–Aug 18 | Central flow term | - More experiments |
| **9** | Aug 19–Aug 25 | Central flow term | - More experiments<br>- Prepare talk for Illiad conference |
| **10** | Aug 26–Sep 1 | Illiad conference | |

Early on in the project (next week) I will want to spend a couple of days thinking about the connection to AI alignment.