# The Inductive Biases of Stochastic Gradient Descent

Guillame Corlouer

July 11, 2025

## 1 Introduction

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 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).

## 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 during training relative to Bayes**, 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.

## 2 Implicit Biases of Gradient Descent

Consider a neural network $f(\theta; x, y)$ with parameters $\theta \in \mathbb{R}^d$ and an empirical loss function $L(\theta)$. Let the gradient of the loss be given by $g = -\nabla_\theta L(\theta)$. Gradient descent is given by the following update rule:

$$\theta_{k+1} = \theta_k - \eta g(\theta_k) \tag{1}$$

where $\eta$ is the learning rate. For theoretical analyses, it is often useful to consider the gradient flow (GF) which is the continuous time limit of gradient descent. Consider that GD is a discretisation of a continuous flow with discrete time steps $t = k\eta$ and let $\eta \to 0$ then we can write the GF equation:

$$d\theta(t) = -g(\theta(t))dt \tag{2}$$

The implicit biases of gradient descent have emerged as a fundamental topic in understanding why overparameterized neural networks generalize well despite classical learning theory predictions. This literature review examines the key theoretical advances in characterizing these implicit biases, particularly focusing on recent developments from 2020-2025.

### 2.1 Low-Rank Bias in Deep Linear Networks

The low-rank bias represents one of the most mathematically well-characterized implicit regularization effects in deep learning. For a deep linear network with weight matrices $W_1, W_2, \ldots, W_L$, the learned representation $W = W_1 W_2 \cdots W_L$ exhibits significantly lower effective rank than the ambient dimension would suggest.

**?** established foundational theoretical results for linearly separable data with exponential-tailed losses, proving that gradient descent converges to the max-margin solution with precise convergence rates: directional convergence at $O(\log\log(t)/\log(t))$ and loss convergence at $O(1/t)$. This max-margin bias was extended by **?** to homogeneous neural networks, demonstrating convergence to KKT points of parameter-space margin maximization.

The nuclear norm minimization conjecture, initially proposed by **?**, suggested that gradient flow with infinitesimal initialization converges to minimum nuclear norm solutions. However, this view was fundamentally challenged by subsequent work. **?** and **?** showed that matrix factorization exhibits implicit bias toward rank minimization rather than nuclear norm minimization. Specifically, **?** proved that gradient flow on depth-2 matrix factorization is mathematically equivalent to "Greedy Low-Rank Learning" – a sequential rank minimization algorithm that greedily adds rank-1 components.

Recent work by **?** provided rigorous dynamical analysis characterizing how the effective rank evolves during training in a distinctive "waterfall" pattern, where different eigenvalues converge at different rates. Their analysis reveals that deeper networks exhibit stronger low-rank bias, with the bias strength depending critically on initialization scale and network depth.

### 2.2 Edge of Stability Phenomenon

The edge of stability (EoS) phenomenon, discovered by **?**, represents a paradigm shift in understanding neural network optimization. During training, the sharpness (largest eigenvalue of the loss Hessian) progressively increases until reaching approximately $2/\eta$, where $\eta$ is the learning rate, then hovers at this critical threshold while loss continues decreasing.

**?** provided the first rigorous mathematical explanation through self-stabilization theory. The dynamics can be captured by cubic Taylor expansion: as iterates diverge along the top Hessian eigenvector, the cubic term causes curvature to decrease until stability is restored. This creates a self-stabilizing effect where gradient descent implicitly follows projected gradient descent under the constraint $S(\theta) \leq 2/\eta$.

**?** demonstrated that EoS represents a novel mechanism of implicit regularization occurring due to non-smooth loss landscapes. The mathematical analysis shows that gradient descent updates evolve along deterministic flows on the manifold of minimum loss, contrasting with previous results that relied on infinitesimal updates or noise.

**?** extended the theoretical understanding through bifurcation theory, proving that different gradient descent trajectories align on specific bifurcation diagrams independent of initialization. This trajectory alignment provides rigorous mathematical foundations for both progressive sharpening and EoS phenomena. Recent extensions to stochastic settings reveal that SGD operates in "Edge of Stochastic Stability," where batch sharpness rather than full-batch sharpness stabilizes at $2/\eta$.

### 2.2.1 Milestones: explicit EoS studies in DLNs

**2022 - Two-layer linear net** (**?**): Four-phase trajectory showing progressive sharpening, EoS oscillation, then eventual monotone descent.

**2022 - General GD theory on EoS** (**?**): GD follows a deterministic flow on the low-loss manifold even past the stability bound.

**2023 - Diagonal DLNs** (**?**): Characterises implicit regularisation at EoS; explains why SGD can benefit from slightly super-critical learning rates.

**2023 - Two-step GD beyond EoS** (**?**): Modified updates tame oscillations without shrinking the step-size.

**2024 - UV toy model** (**?**): Minimalist proof that progressive sharpening and EoS appear with a single training example.

**2025 - Deep matrix-factorisation** (**?**): First fine-grained analysis of loss oscillations in a subspace whose dimension is set by the learning rate.

**2025 - Single-neuron-per-layer chains** (**?**): Relates dataset difficulty and depth to how fast sharpness saturates at the $2/\eta$ threshold.

*Historical note:* classical work by **?** already captured the "progressive sharpening" of singular values in DLNs, but did not frame it in EoS terms.

### 2.2.2 Insights distilled from the linear literature

- **Sharpness self-regulation.** Exact or bounded spectra show $\lambda_{\max}$ equilibrates near $2/\eta$ once conservation laws linked to layer-wise products break at the threshold (**?**).

- **Low-dimensional oscillations.** Beyond the threshold, loss oscillates inside a subspace of dimension $\leq \lceil 2/\eta \rceil$—rank 1 for depth-2 nets (**?**).

- **Bifurcation picture.** Fixed $\eta$ triggers a period-doubling cascade and even chaos in deep linear chains, mirroring nonlinear ReLU behaviour (**??**).

- **Implicit-bias shift.** At EoS, usual max-margin or minimum-norm biases weaken; diagonal-network analysis shows large $\eta$ tilts GD and SGD towards different sparsity profiles (**?**).

- **Depth–data interplay.** More depth lowers the $\eta$ needed to hit EoS, while *harder* datasets delay sharpness saturation (**?**).

#### 2.2.3 Open directions

1. Extending proofs to *non-square* or convolutional linear networks.
2. Linking EoS oscillations to *generalisation*; most current theory still uses synthetic data.
3. Developing a full EoS theory for *mini-batch SGD*; first steps exist but are incomplete (**?**).

**Bottom line.** Multiple independent lines of evidence now confirm that DLNs *do* exhibit edge-of-stability dynamics. Because their algebra is explicit, they have become the canonical sandbox for explaining why sharpness hovers near $2/\eta$, how oscillations arise, and what this implies for implicit regularisation.

### 2.3 Additional Implicit Biases

Beyond low-rank bias and edge of stability, neural networks exhibit multiple implicit biases that collectively shape learned representations.

#### 2.3.1 Feature Averaging Bias

**?** rigorously characterized feature averaging bias, showing that two-layer ReLU networks trained with gradient descent on multi-cluster data converge to averaged combinations of discriminative features rather than learning individual features. This bias creates fundamental adversarial vulnerability, as networks become susceptible to perturbations aligned with negative directions of averaged features.

#### 2.3.2 Spectral Bias

**?** established that neural networks exhibit spectral bias, learning from low to high frequencies during training. The mathematical analysis through Neural Tangent Kernel theory reveals that for harmonic functions $f(x) = \sum a_k \sin(kx)$, low-frequency components (small $k$) are learned first, with convergence time scaling as $O(\kappa^d/p(x))$ where $p(x)$ is local data density.

#### 2.3.3 Simplicity Bias

**?** demonstrated that networks exhibit simplicity bias, preferring simple over complex features even when complex features are more predictive. They showed extreme cases where networks become invariant to all predictive complex features, relying exclusively on the simplest available features. **?** proved that 1-hidden layer networks depend primarily on linearly separable subspaces, providing mathematical characterization of this bias.

#### 2.3.4 Sparsity Bias

**?** showed that for diagonal networks, initialization scale controls implicit bias: as initialization scale $\alpha \to 0$, the implicit regularizer approaches $R(w) = \|w\|_1$, while $\alpha \to \infty$ gives $R(w) = \|w\|_2^2$, creating automatic feature selection without explicit regularization.

### 2.4 Discrete versus Continuous Optimization

The relationship between gradient descent (discrete updates) and gradient flow (continuous time limit) reveals important distinctions in implicit bias. While gradient flow provides clean theoretical analysis, finite step sizes introduce discretization effects that can significantly alter implicit biases.

For linear models, both formulations generally converge to similar solutions – minimum $\ell_2$-norm interpolators for squared loss and max-margin solutions for exponential losses. However, discrete gradient descent with finite learning rates can exhibit different implicit regularization patterns not captured by continuous analysis (**?**).

In neural networks, the differences become more pronounced. Homogeneous networks with favorable curvature properties along gradient flow trajectories allow discrete gradient descent to closely approximate continuous flow. However, finite step sizes can induce different implicit biases, particularly in stochastic variants where different eigenvalue directions are explored (**?**).

## 2.5 Recent Theoretical Advances (2020-2025)

The period 2020-2025 has witnessed several breakthrough theoretical advances. The discovery of Edge of Stability fundamentally changed understanding of neural network optimization, revealing that large learning rates create beneficial implicit bias toward flatter minima through self-stabilization mechanisms.

Feature averaging identification by **?** provided the first rigorous explanation for adversarial vulnerability in gradient descent, showing that networks trained with standard procedures systematically learn vulnerable representations. This breakthrough suggested that granular supervision (classifying individual features rather than original classes) can achieve optimal adversarial robustness.

Matrix factorization theory experienced major advances with **?** definitively resolving the nuclear norm conjecture, proving that gradient flow implements greedy rank minimization rather than nuclear norm minimization. **?** provided precise dynamical analysis of how effective rank evolves during training.

Extensions to non-homogeneous networks by **?** represent significant theoretical generalization, providing asymptotic convergence characterization for generic deep networks beyond the homogeneous case. **?** generalized foundational binary classification results to realistic multiclass settings through the framework of Permutation Equivariant and Relative Margin-based (PERM) losses.