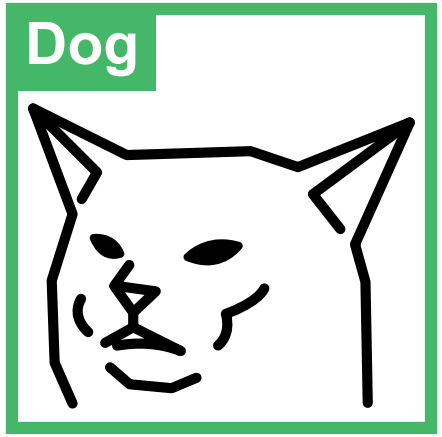


WHY DOES SGD PREFER FLAT MINIMA?

THROUGH THE LENS OF DYNAMICAL SYSTEMS

Hikaru Ibayashi, Masaaki Imaizumi

M. Mahdi Mostafaei

The Empirical Phenomenon & The Objective

The Big Question:

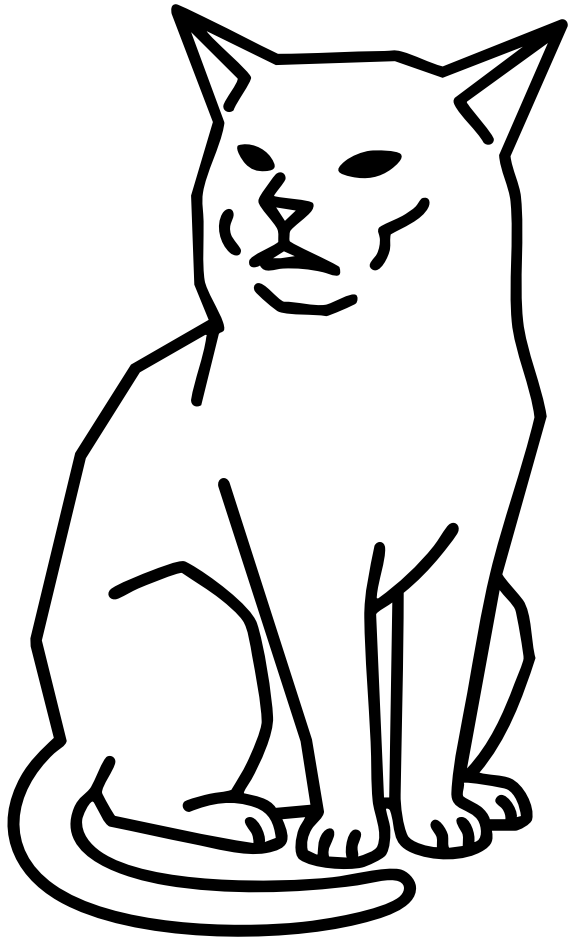
Neural networks that settle into "flat" minima empirically generalize better to unseen data than those stuck in "sharp" minima. How does the algorithm "know" to avoid sharp minima?

The Hypothesis:

The prevailing theory is that the inherent stochastic noise of mini-batch SGD allows it to "escape" sharp minima quickly.

The Paper's Objective:

To rigorously prove that SGD escapes sharp minima exponentially fast. Crucially, they do this by treating modern SGD as a **non-stationary** continuous dynamical system.



The Flaw in Previous Theories

The Old Approach:

Previous works (e.g., Xie et al.) modeled the optimizer state (a continuous probability cloud of parameters) reaching a **stationary distribution** (thermal equilibrium), where the probability density no longer changes over time. They used the Kramers escape rate.

The Dimensionality Problem:

It takes $\mathcal{O}(d)$steps (where $d$is the number of parameters) for a system to reach stationarity. Because modern deep learning models have billions of parameters, real-world SGD *never* actually reaches equilibrium.

The LDT Solution:

The authors abandon stationary probability densities and employ **Large Deviation Theory (LDT)**, which analyzes the deterministic "action" of escaping trajectories, bypassing the need for equilibrium entirely.

SGD to SDE (then Discrete to Continuous)

The Discrete Algorithm:

$$\theta_{k+1} = \theta_k - \eta \nabla L^B(\theta_k)$$

Decomposing the Noise:
Separating the true gradient from the noise, and applying the Central Limit Theorem:

$$\theta_{k+1} = \theta_k - \eta \nabla L(\theta_k) + \sqrt{\frac{\eta}{B}} W_k$$

$$W_k \sim \mathcal{N}(0, \eta C(\theta_k))$$

The Continuous SDE:
Approximating this via the Euler scheme gives the Stochastic Differential Equation (SDE)

$$\dot{\theta}_t = -\nabla L(\theta_t) + \sqrt{\frac{\eta}{B}} C(\theta_t)^{1/2} \dot{w}_t$$

Core Assumptions

Assumption 1 (Locally Quadratic):
Near the minimum, the loss surface is a quadratic bowl:

$$L(\theta) = L(\theta^*) + \frac{1}{2}(\theta - \theta^*)^\top H^* (\theta - \theta^*)$$

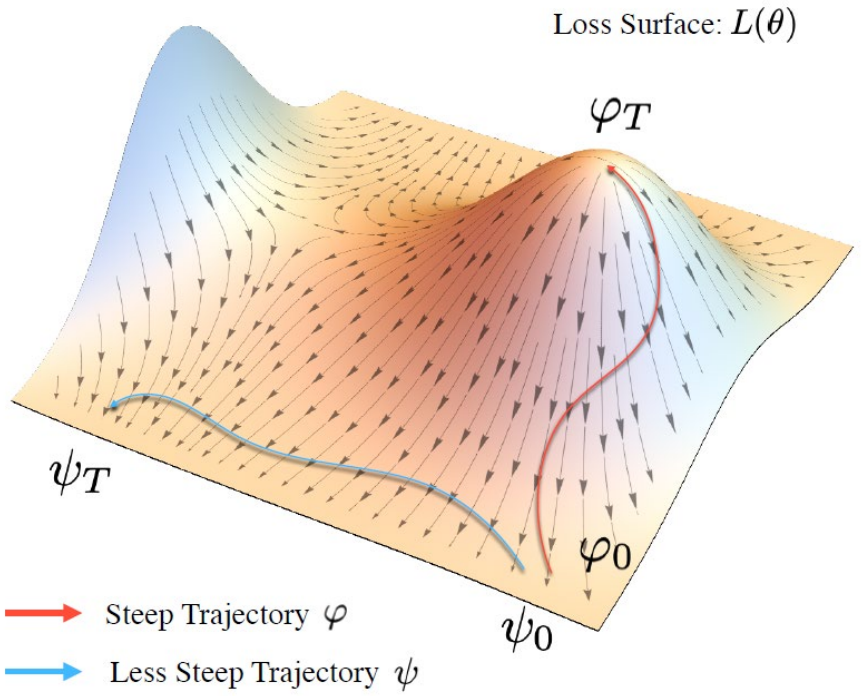
Assumption 2 (The Golden Assumption):
The SGD noise covariance exactly equals the Hessian of the loss landscape:

$$C(\theta^*) = H^*$$

This links the algorithm's anisotropic (direction-dependent) noise directly to the physical sharpness (λ_{max}) of the landscape.

LDT and the Action Functional

The Action Functional:
LDT analyzes trajectories instead of steady-state probabilities. For a trajectory φ over time T, they define "steepness", measuring the effort required to move against the gradient:



$$S_T(\varphi) := \frac{1}{2}\int_0^T (\dot{\varphi}_t + \nabla L(\varphi_t))^\top C(\varphi_t)^{-1}(\dot{\varphi}_t + \nabla L(\varphi_t))\, dt$$

The Quasi–Potential

The quasi-potential $V(\theta)$ is the absolute minimum steepness required to travel from the minimum $\boldsymbol{\theta}^*$ to a point $\boldsymbol{\theta}$ on the boundary.

$$V(\theta) := \inf_{T>0} \inf_{\varphi:(\varphi_0,\varphi_T)=(\theta^*,\theta)} S_T(\varphi)$$

The Intractable Problem:

Calculating the true $V(\theta)$ is mathematically impossible because the noise matrix $C(\varphi_t)^{-1}$ keeps changing at every microscopic step.

Some Tricks …

The Proxy System Trick (Lemma 1)

Instead of solving the real system, you can evaluate a "proxy" isotropic system where the noise is uniform. We effectively pretend that the noise matrix is just the identity matrix $C(\theta) = I$.

For this simplified proxy system, the boundary quasi-potential $\hat{V}_0$ evaluates to exactly twice the depth of the minimum:

$$\hat{V}_0 = 2\Delta L$$

Still Some Tricks …

Bringing it Back to Reality (Lemma 2)

Because the noise variance is dictated by the Hessian, the noise is strongest in the direction of maximum curvature. This maximum curvature is the maximum eigenvalue of the Hessian, known as the sharpness, λ_{max}

We simply scale the proxy quasi-potential down by this sharpness to find the true quasi-potential V_0

$$V_0 \approx \frac{2\Delta L}{\lambda_{max}}$$

Escape Time

Theorem 1 (Fundamental Limit of LDT):

LDT states the expected exit time $\mathbb{E}[\tau]$ is exponentially dominated by the lowest quasi-potential on the boundary:

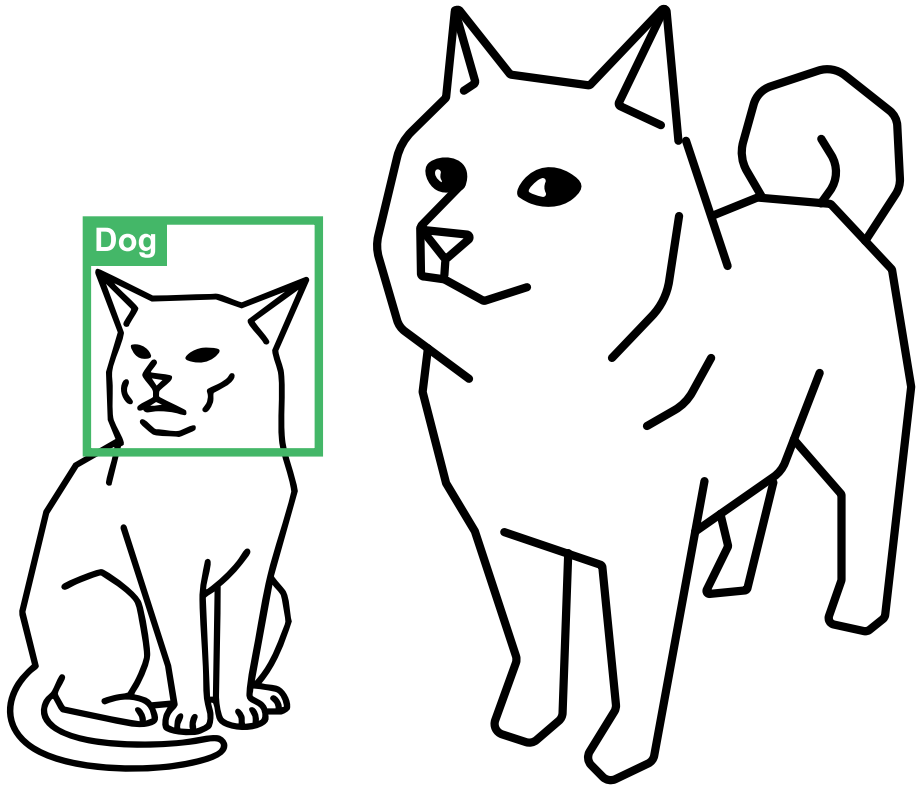
$$\lim_{\eta \to 0} \frac{\eta}{B} \ln \mathbb{E}[\tau] = V_0$$

Theorem 2 (Continuous SDE Exit Time):

Substituting our derived V_0:

$$\lim_{\eta \to 0} \frac{\eta}{B} \ln \mathbb{E}[\tau] \approx 2 \frac{\Delta L}{\lambda_{max}}$$

$$\mathbb{E}[\tau] \approx \exp\left(\frac{B}{\eta} \frac{2\Delta L}{\lambda_{max}} \right)$$


Dog