A DIFFUSION THEORY FOR DEEP LEARNING DYNAMICS:

Stochastic Gradient Descent Exponentially Favors Flat Minima

Article

Xie, Z., Sato, I., & Sugiyama, M. (2020). A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. arXiv preprint arXiv:2002.03495.

Authors
Zeke Xie, Issei Sato, & Masashi Sugiyama

Presented by
Aymeric Côme & Omar Iken

Master 2 Data Science

Université de Lille

November 25, 2021

Outline

1. Introduction
2. Behavior of SGD Noise Near Critical Points
3. SGD Escapes from Minima
4. Experimentation results and analysis

Introduction

Loss functions & Minima

A vast majority of Machine Learning problems come back to minimizing a loss function over some parameters. In Deep Learning especially, these can be very complex - how can we find a good minimum?

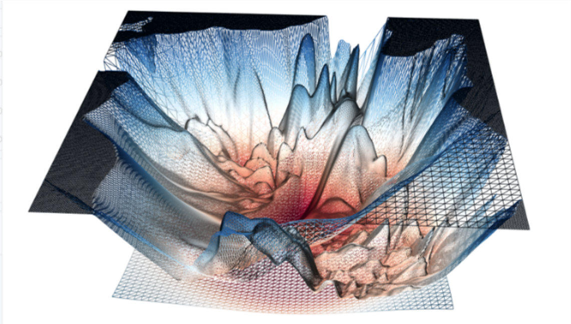


Figure 1: Visualization of the loss function of a Neural Network
[https://www.telesens.co/2019/01/16/neural-network-loss-visualization/]

From literature [Hardt et al.,2016]: flat minima are better for generalization

However, it is mathematically unclear how deep learning can select a flat minimum among so many minima

Diffusion Theory

Diffusion Theory

The diffusion theory is an important theoretical tool to understand how deep learning dynamics works. It helps us model the diffusion process of probability densities of parameters instead of model parameters themselves

Find a minimum of the loss ⇄ pour a drop of water on it and see where it goes

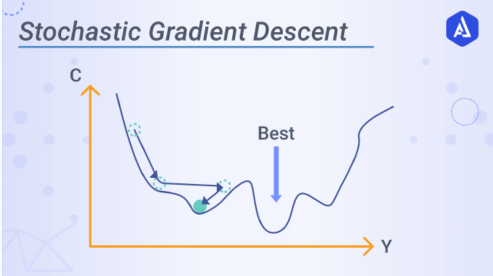


Figure 2: Illustration of SGD finding a local rather than the global minimum

[https://www.akira.ai/glossary/stochastic-gradient-descent]

Contributions

Main Contributions

- The proposed theory formulates the fundamental roles of gradient noise, batch size, the learning rate, and the Hessian in minima selection during SGD.
- The SGN covariance is approximately proportional to the Hessian and inverse to batch size.
- Either a small learning rate or large-batch training requires exponentially many iterations to escape minima in terms of ratio of batch size and learning rate.
- SGD favors flat minima exponentially more than sharp minima.

Behavior of SGD Noise Near Critical Points

Stochastic Gradient Noise Analysis

SGD Dynamics [Mandt et al.]

$$\theta_{t+1} = \theta_t - \eta \frac{\partial \hat{L}(\theta_t)}{\partial \theta_t} = \theta_t - \eta \frac{\partial L(\theta_t)}{\partial \theta_t} + \eta \underbrace{C(\theta_t)^{1/2}\zeta_t}_{\text{gradient noise}} \tag{1}$$

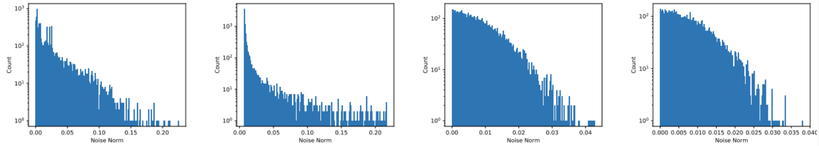
- $\hat{L}(\theta_t)$: loss of one minibatch
- $\zeta_t \sim \mathcal{N}(0, I)$
- $C(\theta_t)$: covariance matrix

$$C(\theta_t)^{1/2}\zeta_t = \frac{\partial L(\theta_t)}{\partial \theta_t} - \frac{\partial \hat{L}(\theta_t)}{\partial \theta_t}$$

According to Generalized Central Limit Theorem [Gnedenko et al., 1954], the Gaussian approximation of SGN is reasonable.

Experimental verification

Verify that the noise is Gaussian rather than Lévy:



(a) "SGN" across parameters (b) Lévy noise (c) SGN across mini-batches (d) Gaussian noise

Figure 3: The histogram of the norm of the gradient noises computed with the three-layer fully-connected network on MNIST

[Simsekli et al., 2019] suggested Lévi noise, but for isotropic cases (the same distribution for all dimensions of θ). The experiences show that anisotropic parameter-dependent SGN rather is Gaussian.

The noise seems to be more Gaussian on some layers/networks than others.

The empirical analysis in Figure 3 holds well at least when the batch size B is larger than 16

SGD Dynamics

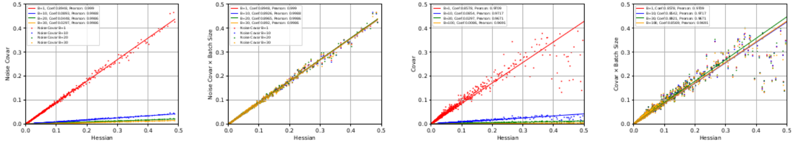
Continuous-time Dynamics of SGD

$$d\theta = -\frac{\partial L(\theta)}{\partial \theta} dt + 2\left[2D(\theta)\right]^{1/2} dW_t \qquad (\eta = dt) \tag{2}$$

Based on [Smith & Le (2018)], we can express the SGN covariance as:

$$C(\theta) \approx \frac{1}{Bn} \sum_{j=1}^{n} \nabla L(\theta, x_j) \nabla L(\theta, x_j)^T = \frac{1}{B} FIM(\theta) \approx \frac{1}{B} H(\theta) \tag{3}$$

near minima (gradient noise variance dominates gradient mean), with B the batch size and $H(\theta)$ the Hessian of the loss at θ

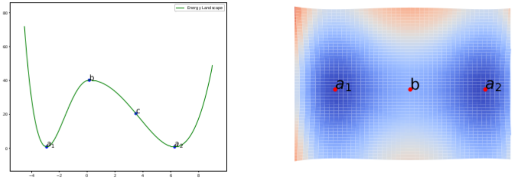


(a) Pretrained Model (b) Pretrained Model (c) Random Model (d) Random Model

Figure 4: Empirical verification of (3) by using three-layer fully-connected network on MNIST

SGD Escapes from Minima

Kramers Escape Problem



(a) 1-Dimensional Escape (b) High-Dimensional Escape

Figure 5: Kramers Escape Problem.

What is the mean escape time for a particle governed by Equation (2) to escape from Sharp Valley a_1 to Flat Valley a_2 ?

Mean Escape Time

$$\tau = \frac{1}{\gamma} = \frac{\mathcal{P}(\theta \in V_a)}{\int_{S_a} J.dS} \tag{4}$$

with $\mathcal{P}(\theta \in V_a)$ the probability inside Valley a, J the probability current and S_a the surface boundary of the valley

Classical Assumptions

To place themselves in the framework of density diffusion theory, they made three classical assumptions:

1. The Second Order Taylor Approximation:

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

2. Quasi-Equilibrium Approximation: The system is in quasi-equilibrium near minima.
3. Low Temperature Approximation: The gradient noise is small (low temperature)

Assumptions 2 and 3 both mean that the diffusion theory can better describe the "slow" (flat minima) escape processes.

Escape Paths

Escape Paths

A path is critical if:

- The gradient perpendicular to the path direction must be zero
- The second order derivatives perpendicular to the path must be non-negative

Most Possible Paths (MPPs) for escaping must be critical(the probability density is concentrated on MPPs).

Multiple-Path Escape

If there are multiple MPPs between the start valley and the end valley, then

$$\gamma_{total} = \sum_{\text{MPP } p} \gamma_p$$

⋆ Higher dimension may increase the number of escape paths.

⋆ Multiple-valley escape can be seen as multiple two-valley escape.

Minima Selection

$$\mathbb{P}(\theta \in \text{Valley } a) = \frac{\tau_a}{\sum_v \tau_v} \tag{5}$$

SGD Diffusion Theory

SGLD diffusion

- *SG Langevin Dynamics*: SG with injected white noise (temperature).
- Hypothesis: in final epochs, white noise dominates SGN ($\eta \approx 0$)
- Mean escape time for SGLD:

$$\tau = \frac{1}{\gamma} = \frac{2\pi}{|H_{be}|} \sqrt{\frac{-det(H_b)}{det(H_a)}} \exp\left(\frac{\Delta L}{D}\right) \tag{6}$$

  - H_a and H_b the Hessians at the minimum a and the saddle point b.
  - H_{be} the eigenvalue of H_b corresponding to the escape direction.
  - ΔL the loss barrier height and D the diffusion coefficient ($\approx$ temperature)

SGD diffusion

- Anisotropic, parameter-dependent and not Gibs-Boltzmann like
- Mean escape time for SGD:

$$\tau = \frac{1}{\gamma} = \frac{2\pi}{|H_{be}|} \exp\left[\frac{2B\Delta L}{\eta}\left(\frac{s}{H_{ae}} + \frac{1-s}{|H_{be}|}\right)\right] \tag{7}$$

for some s depending on the path

Experimentation results and analysis

Experimentation

- Run escape processes for various **gradient noise scales, batch sizes, learning rates, sharpness.**
- *Sharpness:* multiply parameters by $\sqrt{k} \implies$ multiply Hessian by k
- *Experimental setup:* 10D Styblinski-Tang function and fully-connected networks on 4 real-world datasets, escape time calculated from needed number of iterations to go out of a valley

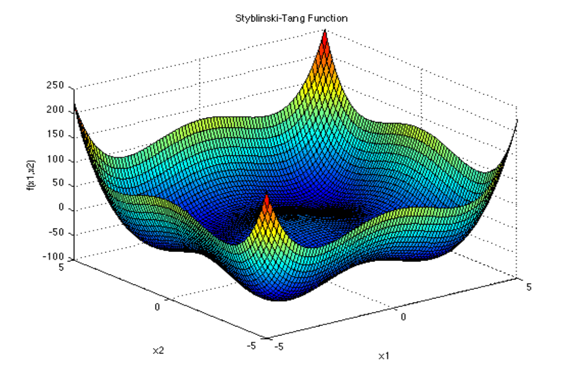
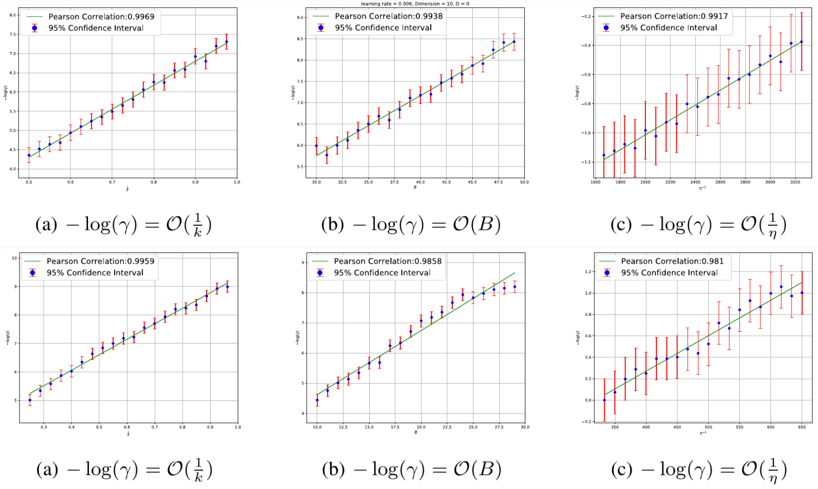


Figure 6: 2D Styblinski-Tang function [https://www.sfu.ca/~ssurjano/stybtang.html]

Results - SGD

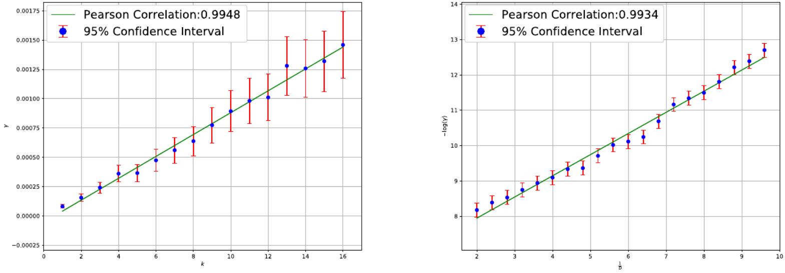


(a) $-\log(\gamma) = \mathcal{O}(\frac{1}{k})$ (b) $-\log(\gamma) = \mathcal{O}(B)$ (c) $-\log(\gamma) = \mathcal{O}(\frac{1}{\eta})$

(a) $-\log(\gamma) = \mathcal{O}(\frac{1}{k})$ (b) $-\log(\gamma) = \mathcal{O}(B)$ (c) $-\log(\gamma) = \mathcal{O}(\frac{1}{\eta})$

Figure 7: Mean escape time analysis for SGD, on Styblinski-Tang function (top) and Avila dataset (bottom), depending on $\frac{1}{k}$, B and $\frac{1}{\eta}$

Results - SGLD



(a) $\gamma = \mathcal{O}(k)$

(b) $-\log(\gamma) = \mathcal{O}(\frac{1}{D})$

Figure 8: Mean escape time analysis for SGLD, on Styblinski-Tang function, depending on $\frac{1}{k}$ and B

SGD favors flat minima exponentially more than sharp minima, exponential influence of batch size and learning rate

Conclusion

To wrap up:

- Formulation of the SGN behavior as a diffusion problem
- Demonstrate that one essential advantage of SGD is selecting flat minima with an exponentially higher probability than sharp minima.
- Use diffusion theory results to estimate the escape time for valleys for anisotropic and parameter-dependent SGN
- Validate the approach with experiments.