# How SGD Selects the Global Minima in Over-parameterized Learning: A Dynamical Stability Perspective

Lei Wu[1], Chao Ma[2], Weinan E[2,3]

[1]School of Mathematical Science, Peking University, China [2]Program of Applied and Computational Mathematics, Princeton University, USA [3]Beijing Institute of Big Data Research, China
Contact: leiwu@pku.edu.cn, chaom@princeton.edu, weinan@math.princeton.edu

## Introduction

- In the over-parameterized setting, the empirical loss function has many global minima.
- Different algorithms choose different set of global minima, thus lead to different generalization performance.
- **Question:** What is the mechanism of global minima selecting for algorithms, especially SGD?

## The escape phenomenon

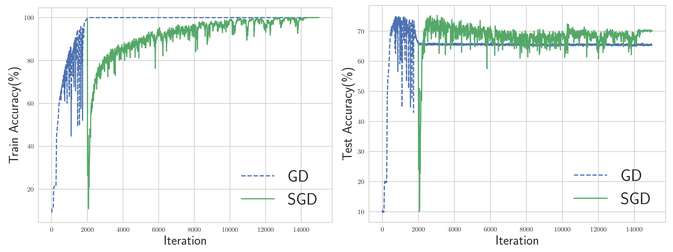
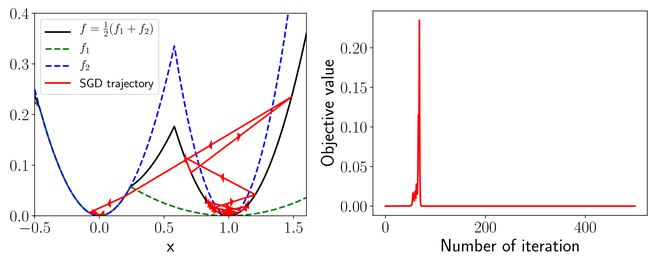



Figure 1: Fast escape phenomenon in fitting corrupted FashionMNIST.

- Though the GD iterations is already close to a global minimum, when the optimizer is switched from GD to SGD (with the same learning rate), the iterator escapes quickly from the global minimum and converges to another global minimum.
- By the right figure, the global minimum found by SGD generalizes better than that found by GD.

## Toy example




$$f_1(x) = \min\{x^2, 0.1(x-1)^2\}, \quad f_2(x) = \min\{x^2, 1.9(x-1)^2\}.$$

**SGD with learning rate $0.7$ escapes the right minimum because of unstability.**

## Linear Stability

- **Definition (Fixed point):**
  We say $x^*$ is a fixed point of stochastic dynamics
  $$x_{t+1} = x_t - G(x_t; \xi_t) \tag{1}$$
  if for any $\xi$, we have $G(x^*; \xi) = 0$.
- **Definition (Linear stability):**
  Let $x^*$ be a fixed point of stochastic dynamics (1). Consider the linearized dynamical system:
  $$\tilde{x}_{t+1} = \tilde{x}_t - A_{\xi_t}(\tilde{x}_t - x^*), \tag{2}$$
  where $A_{\xi_t} = \nabla_x G(x^*, \xi_t)$. We say that $x^*$ is *linearly stable* if there exists a constant $C$ such that,
  $$\mathbb{E}\|\tilde{x}_t\|^2] \le C\|\tilde{x}_0\|^2, \text{for all } t > 0. \tag{3}$$

## Theory for SGD

### Theorem

*Consider an approximation of $f(x)$ near global minimum $x^*$, $f(x) \approx \frac{1}{2n}\sum_{i=1}^n (x-x^*)^\top H_i (x-x^*)$. The global minimum $x^*$ is linearly stable for SGD with learning rate $\eta$ and batch size $B$ if the following condition is satisfied*

$$\lambda_{\max}\left\{(I - \eta H)^2 + \frac{\eta^2(n-B)}{B(n-1)}\Sigma\right\} \le 1. \tag{4}$$

*In (4), $H = \frac{1}{n}\sum_{i=1}^n H_i, \Sigma = \frac{1}{n}\sum_{i=1}^n H_i^2 - H^2$.*

The selection of global minima can be shown in a sharpness-non-uniformity diagram, with sharpness $a = \lambda_{\max}(H)$, non-uniformity $s = \lambda_{\max}(\Sigma)$.

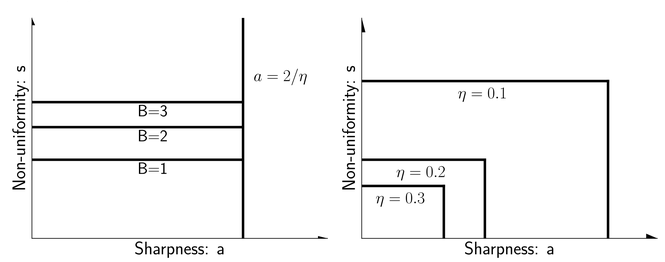



Figure 2: The sharpness-non-uniformity diagram, showing the rectangular region that is linearly stable for SGD. The left and right figure shows the influence of batch size $B$ and learning rate $\eta$, respectively.

## Numerical experiments

- **Learning rate and sharpness**

Table 1: Sharpness of the solutions found by GD with different learning rates. Dashes indicate that GD blows up with that learning rate.

| $\eta$ | 0.01 | 0.05 | 0.1 | 0.5 |
|---|---|---|---|---|
| FashionMNIST | $53.5 \pm 4.3$ | $39.3 \pm 0.5$ | $19.6 \pm 0.15$ | $3.9 \pm 0.0$ |
| CIFAR10 | $198.9 \pm 0.6$ | $39.8 \pm 0.2$ | $19.8 \pm 0.1$ | $3.6 \pm 0.4$ |
| prediction $2/\eta$ | 200 | 40 | 20 | 4 |

- **GD tends to select the sharpest possible minima.**

- **Minima selected by SGD**

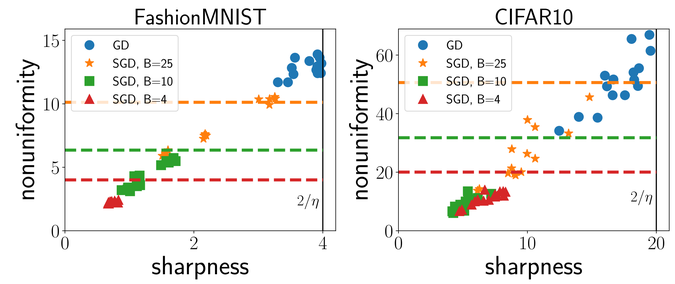



Figure 3: The sharpness-non-uniformity diagram for the minima selected by SGD. Different colors correspond to different set of hyper-parameters. The dash line shows the predicted upperbound for the non-uniformity.

- **The points for different set of hyper-parameters lie below the corresponding dash line.**
- **The sharpness and non-uniformity are correlated.**

- **The correlation of sharpness and non-uniformity**

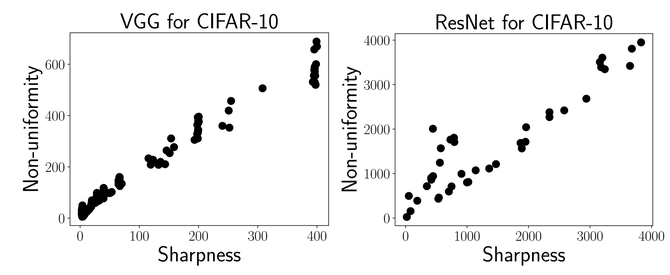



Figure 4: Scatter plot of sharpness and non-uniformity, suggesting that the non-uniformity and sharpness are roughly proportional to each other for these models.