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

by Lei Wu, Chao Ma, Weinan E

Dec 6, 2019

Reproduced By: Patrick Myers, Gaurav Jindal, Rishab Bamrara, Phillip Seaton

Team: Skyhawks
Motivation:

- In models with many parameters, such as deep learning, multiple global minima can exist.
- Although these global minima perform equally well on the training set, some generalize better than others.
- By better understanding how global minima are selected in such scenarios, it should be easier to select models which generalize better to testing data.
Though GD is close to a global minimum, switching to SGD causes the model to converge to a different global minimum which generalizes better than the GD minimum. However, SGD takes longer to converge.
Background - Escape Phenomenon (contd.):

- In this example, SGD will converge to $x = 0$ escaping from the right minima due to instability. Gradient descent behaved in a similar manner with the same learning rate.
- Curvature is more stable at $x = 0$ which causes the escape phenomenon to occur.

\[ f_1(x) = \min\{x^2, 0.1(x - 1)^2\}, \quad f_2(x) = \min\{x^2, 1.9(x - 1)^2\} \]
\[ f(x) = \frac{1}{2} (f_1(x) + f_2(x)) \]
The first equation is SGD with a batch size of 1. SGD can only pick minima where \( s \leq 1 / \eta \)

At \( x = 1 \), \( s = 1.8 > 1 / 0.7 \)
At \( x = 0 \), \( s = 0 < 1 / 0.7 \)
We see that at \( x = 0 \) the requirement is met, and \( x = 0 \) is the only valid minima.
Related Work:

- Hu et al. [5] examined the escape phenomena and concluded that it is generally easier to escape from sharper minimizers.
- Jastrzebski et al. [6] found that the noise factor (learning rate / batch size) affects the sharpness of the solution that SGD will reach.
- Wilson et al. [12] showed that adaptive gradient methods will generally converge to solutions which do not generalize as well as those which will be reached by standard SGD.
Background - Sharpness and Non-uniformity:

\[ H_i = \nabla^2 f_i(x^*). \]

\[ H = \frac{1}{n} \sum_{i=1}^{n} H_i, \Sigma = \frac{1}{n} \sum_{i=1}^{n} H_i^2 - H^2. \]

- **Sharpness \((a)\):** A measure of how quickly the slope of a loss function changes on average, represented mathematically by the second derivative of the loss function.

\[ a = \lambda_{\text{max}}(H) \]

- **Non-uniformity \((s)\):** A measure of smoothness across a loss function.

\[ s = \lambda_{\text{max}}(\Sigma^{1/2}) \]
For minimizing the following training error \( f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) \) by a general optimizer: \( x_{t+1} = x_t - G(x_t; \xi_t) \)

**Definition 1:** \( x^* \) is a **fixed point** in stochastic dynamics, if for any \( \varepsilon \), \( G(x^*; \varepsilon) = 0 \).

**Definition 2:** If \( x^* \) is a fixed point in stochastic dynamics, and there is a linearized dynamical system \( \tilde{x}_{t+1} = \tilde{x}_t - A_{\xi_t} (\tilde{x}_t - x^*) \) where, \( A_{\xi_t} = \nabla_x G(x^*, \xi_t) \). Then, \( x^* \) is **linearly stable** if there exists a \( C \) such that, \( \mathbb{E}[\|\tilde{x}_t\|^2] \leq C\|\tilde{x}_0\|^2 \) for all \( t > 0 \).

For SGD, \( G(x_t; \xi_t) = \eta \nabla f_{\xi_t}(x_t) \);
**Theorem 1**: The global minimum $x^*$ is **stable** for SGD with learning rate $\eta$, and batch size $B$ if the following condition is satisfied:

$$\lambda_{\text{max}} \left\{ (I - \eta H)^2 + \frac{\eta^2(n - B)}{B(n - 1)} \Sigma \right\} \leq 1$$
Both sharpness and non-uniformity have an effect on the selection of global minima by GD and SGD.

In general, SGD will prefer to select global minima with a lower degree of non-uniformity.

Both sharpness (a) and non-uniformity (s) are bounded by the ranges in the following expressions where \( \eta \) is learning rate and B is batch size:

\[
0 \leq a \leq \frac{2}{\eta}, \quad 0 \leq s \leq \frac{1}{\eta} \sqrt{\frac{B(n - 1)}{n - B}}.
\]
Sharpness-non-uniformity diagram of SGD:

- If we increase learning rate, then SGD is forced to choose a global minima closer to the origin (i.e. smaller sharpness and smaller non-uniformity)
- Decreasing the batch size only forces SGD to choose global minima with smaller non-uniformity, but does not affect sharpness
Experimental Setup:

- Examine the relationship between sharpness and non-uniformity on the convergence of GD and SGD using two different datasets with various batch sizes.
- Two classification problems, FashionMNIST and CIFAR10, will be used to verify this relationship within the context of deep learning.

<table>
<thead>
<tr>
<th>Network type</th>
<th># of parameters</th>
<th>Dataset</th>
<th># of training examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>FNN</td>
<td>898,510</td>
<td>FashionMNIST</td>
<td>1000</td>
</tr>
<tr>
<td>VGG</td>
<td>71,410</td>
<td>CIFAR10</td>
<td>1000</td>
</tr>
</tbody>
</table>
Data Summary:

- **FashionMNIST**: A classification dataset consisting of 28x28 grayscale images of 10 different types of clothing.

- **CIFAR10**: A classification dataset consisting of 32x32 color images across 10 different categories; only images in the “airplane” and “automobile” categories were used.
Our Results
Results

<table>
<thead>
<tr>
<th></th>
<th>Train accuracy</th>
<th>Test accuracy</th>
<th>Sharpness</th>
<th>Non uniformity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fashion MNIST</strong></td>
<td>99.9</td>
<td>80.4</td>
<td>19.9</td>
<td>40.4</td>
</tr>
<tr>
<td><strong>CIFAR 10</strong></td>
<td>100</td>
<td>88.9</td>
<td>19.2</td>
<td>53</td>
</tr>
</tbody>
</table>

- Trained the simple FNN model on Fashion MNIST and VGG11 model on CIFAR10 dataset.
Non-uniformity and Sharpness vs Batch Size

- Because we have 1000 samples, the rightmost points (batch size = 1000) corresponds to GD.
- As batch size increases, sharpness and non-uniformity tend to increase as well. Smaller batch sizes lead to flatter solutions.
- Lower learning rates tend to result in higher sharpness and non-uniformity.
Non-uniformity vs Sharpness

- There is a positive relationship between sharpness and nonuniformity.
- Models with higher sharpness will result in higher nonuniformity.
- Larger batch sizes cause the non-uniformity to be close to the upper bounds.
To better show the escape process, we only show the first 3000 iterations.

- Lower sharpness results on more stable global minima and higher test accuracy.
- The trials with higher sharpness take longer to escape an unstable minima.
Comparing GD and SGD Performance

- Higher sharpness causes less stability
- We see GD perform much better in the training accuracy but is only marginally better on the test accuracy.
Experimental Analysis:

• The positive correlation between sharpness and non-uniformity may explain why SGD tends to converge to flatter minima:
  – Flatter minima will have lower non-uniformity
  – It is easier to escape from areas that are non-uniform, particularly with SGD, making areas with low non-uniformity better candidates for convergence
Conclusion:

- Both sharpness and non-uniformity have important impacts on the selection of global minima by GD and SGD.
- In neural networks, non-uniformity is approximately proportional to sharpness.
- In general, SGD can more easily converge to a more uniform global minima than GD, resulting in better generalization and higher test accuracy.
  - However, this is a phenomena that still needs to be looked into further in future work.
Paper References:

## Team Member Contributions

<table>
<thead>
<tr>
<th>Name</th>
<th>Contribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patrick Myers</td>
<td>I helped organize code in the .ipynb and wrote code to run some experiments and visualize them. I also helped display and discuss our results in the powerpoint.</td>
</tr>
<tr>
<td>Gaurav Jindal</td>
<td>Worked on data loader and data preprocessing part. I also helped in plotting the results and discuss them in the slide.</td>
</tr>
<tr>
<td>Rishab Bamrara</td>
<td>Worked on the linear algebra library and understood the functions which compute sharpness and nonuniformity. Also coded for visualization of results.</td>
</tr>
<tr>
<td>Phillip Seaton</td>
<td>I worked with 3 methods: compute_minibatch, training the model and accuracy. I also helped with commenting code, creating the powerpoint slides and discussing our results.</td>
</tr>
</tbody>
</table>