The **Anisotropic Noise** in Stochastic Gradient Descent: Its Behavior of Escaping from Sharp Minima and Regularization Effects

Zhanxing Zhu*, Jingfeng Wu*, Bing Yu, Lei Wu, Jinwen Ma.

Peking University  Beijing Institute of Big Data Research

June, 2019
The implicit bias of stochastic gradient descent

- Compared with gradient descent (GD), stochastic gradient descent (SGD) tends to generalize better.
- This is attributed to the noise in SGD.
- In this work we study the anisotropic structure of SGD noise and its importance for escaping and regularization.
Stochastic gradient descent and its variants

Loss function \( L(\theta) := \frac{1}{N} \sum_{i=1}^{N} \ell(x_i; \theta) \).

Gradient Langevin dynamic (GLD)
\[
\theta_{t+1} = \theta_t - \eta \nabla_{\theta} L(\theta_t) + \eta \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma_t^2 \mathbf{I}).
\]

Stochastic gradient descent (SGD)
\[
\theta_{t+1} = \theta_t - \eta \tilde{g}(\theta_t), \quad \tilde{g}(\theta_t) = \frac{1}{m} \sum_{x \in B_t} \nabla_{\theta} \ell(x; \theta_t).
\]

The structure of SGD noise
\[
\tilde{g}(\theta_t) \sim \mathcal{N} \left( \nabla L(\theta_t), \Sigma_{\text{sgd}}(\theta_t) \right), \quad \Sigma_{\text{sgd}}(\theta_t) \approx \frac{1}{m} \left[ \frac{1}{N} \sum_{i=1}^{N} \nabla \ell(x_i; \theta_t) \nabla \ell(x_i; \theta_t)^T - \nabla L(\theta_t) \nabla L(\theta_t)^T \right].
\]

SGD reformulation
\[
\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t) + \eta \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \Sigma_{\text{sgd}}(\theta_t)).
\]
GD with unbiased noise

\[ \theta_{t+1} = \theta_t - \eta \nabla_{\theta} L(\theta_t) + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \Sigma_t). \]  \hspace{1cm} (1)

Iteration (1) could be viewed as a discretization of the following continuous stochastic differential equation (SDE):

\[ d\theta_t = -\nabla_{\theta} L(\theta_t) \, dt + \sqrt{\Sigma_t} \, dW_t. \]  \hspace{1cm} (2)

Next we study the role of noise structure \( \Sigma_t \) by analyzing the continuous SDE (2).
Escaping efficiency

Definition (Escaping efficiency)

Suppose the SDE (2) is initialized at minimum \( \theta_0 \), then for a fixed time \( t \) small enough, the *escaping efficiency* is defined as the increase of loss potential:

\[
\mathbb{E}_{\theta_t}[L(\theta_t) - L(\theta_0)]
\]

(3)

Under suitable approximations, we could compute the escaping efficiency for SDE (2),

\[
\mathbb{E}[L(\theta_t) - L(\theta_0)] = - \int_0^t \mathbb{E} \left[ \nabla L^T \nabla L \right] + \int_0^t \frac{1}{2} \mathbb{E} \text{Tr}(H_t \Sigma_t) \, dt
\]

(4)

\[
\approx \frac{1}{4} \text{Tr} \left( \left( I - e^{-2Ht} \right) \Sigma \right) \approx \frac{t}{2} \text{Tr} (H \Sigma) .
\]

(5)

Thus \( \text{Tr} (H \Sigma) \) serves as an important indicator for measuring the escaping behavior of noises with different structures.
Factors affecting the escaping behavior

The noise scale For Gaussian noise $\epsilon_t \sim \mathcal{N}(0, \Sigma_t)$, we can measure its scale by $\|\epsilon_t\|_{\text{trace}} := \mathbb{E}[\epsilon_t^T \epsilon_t] = \cdots = \text{Tr}(\Sigma_t)$. Thus based on $\text{Tr}(H \Sigma)$, we see that the larger noise scale is, the faster the escaping happens.

To eliminate the impact of noise scale, assume that

\[
\text{given time } t, \text{ Tr}(\Sigma_t) \text{ is constant.} \quad (6)
\]

The ill-conditioning of minima For the minima with Hessian as scalar matrix $H_t = \lambda I$, the noises in same magnitude make no difference since $\text{Tr}(H_t \Sigma_t) = \lambda \text{Tr}\Sigma_t$.

The structure of noise For the ill-conditioned minima, the structure of noise plays an important role on the escaping!
The impact of noise structure

Proposition

Let $H_{D \times D}$ and $\Sigma_{D \times D}$ be semi-positive definite. If

1. $H$ is ill-conditioned. Let $\lambda_1, \lambda_2 \ldots \lambda_D$ be the eigenvalues of $H$ in descent order, and for some constant $k \ll D$ and $d > \frac{1}{2}$, the eigenvalues satisfy

$$\lambda_1 > 0, \lambda_{k+1}, \lambda_{k+2}, \ldots, \lambda_D < \lambda_1 D^{-d};$$ (7)

2. $\Sigma$ is “aligned” with $H$. Let $u_i$ be the corresponding unit eigenvector of eigenvalue $\lambda_i$, for some projection coefficient $a > 0$, we have

$$u_1^T \Sigma u_1 \geq a\lambda_1 \frac{Tr \Sigma}{Tr H}.$$ (8)

Then for such anisotropic $\Sigma$ and its isotropic equivalence $\bar{\Sigma} = \frac{Tr \Sigma}{D} I$ under constraint (6), we have the follow ratio describing their difference in term of escaping efficiency,

$$\frac{Tr(H \Sigma)}{Tr(H \bar{\Sigma})} = O \left( aD^{(2d-1)} \right), \quad d > \frac{1}{2}. $$ (9)
Analyze the noise of SGD via Proposition 1

By Proposition 1, The anisotropic noises satisfying the two conditions indeed help escape from the ill-conditioned minima. Thus to see the importance of SGD noise, we only need to show it meets the two conditions.

- Condition 1 is naturally hold for neural networks, thanks to their over-parameterization!
- See the following Proposition 2 for the second condition.
SGD noise and Hessian

Proposition

Consider a binary classification problem with data \( \{(x_i, y_i)\}_{i \in I}, y \in \{0, 1\} \), and mean square loss, \( L(\theta) = \mathbb{E}_{(x,y)} \| \phi \circ f(x; \theta) - y \|_2^2 \), where \( f \) denotes the network and \( \phi \) is a threshold activation function,

\[
\phi(f) = \min \{ \max \{ f, \delta \}, 1 - \delta \},
\]

where \( \delta \) is a small positive constant.

Suppose the network \( f \) satisfies:

1. it has one hidden layer and piece-wise linear activation;
2. the parameters of its output layer are fixed during training.

Then there is a constant \( a > 0 \), for \( \theta \) close enough to minima \( \theta^* \),

\[
u(\theta)^T \Sigma(\theta) u(\theta) \geq a \lambda(\theta) \frac{\text{Tr} \Sigma(\theta)}{\text{Tr} H(\theta)}
\]

(11)

holds almost everywhere, for \( \lambda(\theta) \) and \( u(\theta) \) being the maximal eigenvalue and its corresponding eigenvector of Hessian \( H(\theta) \).
Examples of different noise structures

Table: Compared dynamics defined in Eq. (1).

<table>
<thead>
<tr>
<th>Dynamics</th>
<th>Noise $\epsilon_t$</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>SGD</td>
<td>$\epsilon_t \sim \mathcal{N}\left(0, \Sigma_{t}^{\text{sgd}}\right)$</td>
<td>$\Sigma_{t}^{\text{sgd}}$ is the gradient covariance matrix.</td>
</tr>
<tr>
<td>GLD constant</td>
<td>$\epsilon_t \sim \mathcal{N}\left(0, \varrho_t^2 I\right)$</td>
<td>$\varrho_t$ is a tunable constant.</td>
</tr>
<tr>
<td>GLD dynamic</td>
<td>$\epsilon_t \sim \mathcal{N}\left(0, \sigma_t^2 I\right)$</td>
<td>$\sigma_t$ is adjusted to force the noise share the same magnitude with SGD noise, similarly hereinafter.</td>
</tr>
<tr>
<td>GLD diagonal</td>
<td>$\epsilon_t \sim \mathcal{N}\left(0, \text{diag}(\Sigma_{t}^{\text{sgd}})\right)$</td>
<td>$\text{diag}(\Sigma_{t}^{\text{sgd}})$ is the diagonal of the covariance of SGD noise $\Sigma_{t}^{\text{sgd}}$.</td>
</tr>
<tr>
<td>GLD leading</td>
<td>$\epsilon_t \sim \mathcal{N}\left(0, \sigma_t \tilde{\Sigma}_t\right)$</td>
<td>$\tilde{\Sigma}<em>t$ is the best low rank approximation of $\Sigma</em>{t}^{\text{sgd}}$.</td>
</tr>
<tr>
<td>GLD Hessian</td>
<td>$\epsilon_t \sim \mathcal{N}\left(0, \sigma_t \tilde{H}_t\right)$</td>
<td>$\tilde{H}_t$ is the best low rank approximation of the Hessian.</td>
</tr>
<tr>
<td>GLD 1st eigven($H$)</td>
<td>$\epsilon_t \sim \mathcal{N}\left(0, \sigma_t \lambda_1 u_1 u_1^T\right)$</td>
<td>$\lambda_1, u_1$ are the maximal eigenvalue and its corresponding unit eigenvector of the Hessian.</td>
</tr>
</tbody>
</table>
2-D toy example

Figure: 2-D toy example. Compared dynamics are initialized at the sharp minima. **Left:** The trajectory of each compared dynamics for escaping from the sharp minimum in one run. **Right:** Success rate of arriving the flat solution in 100 repeated runs.
One hidden layer network

Figure: One hidden layer neural networks. The solid and the dotted lines represent the value of $\text{Tr}(H\Sigma)$ and $\text{Tr}(H\bar{\Sigma})$, respectively. The number of hidden nodes varies in $\{32, 128, 512\}$. 
FashionMNIST experiments

**Figure**: FashionMNIST experiments. **Left**: The first 400 eigenvalues of Hessian at $\theta^*_{GD}$, the sharp minima found by GD after 3000 iterations. **Middle**: The projection coefficient estimation $\hat{a} = \frac{u_1^T \Sigma u_1 \text{Tr} H}{\lambda_1 \text{Tr} \Sigma}$ in Proposition 1. **Right**: $\text{Tr}(H_t \Sigma_t)$ versus $\text{Tr}(H_t \bar{\Sigma}_t)$ during SGD optimization initialized from $\theta^*_{GD}$, $\bar{\Sigma}_t = \frac{\text{Tr} \Sigma_t}{D} I$ denotes the isotropic equivalence of SGD noise.
FashionMNIST experiments

Figure: FashionMNIST experiments. Compared dynamics are initialized at $\theta^*_GD$ found by GD, marked by the vertical dashed line in iteration 3000. **Left:** Test accuracy versus iteration. **Right:** Expected sharpness versus iteration. Expected sharpness (the higher the sharper) is measured as $\mathbb{E}_{\nu \sim \mathcal{N}(0, \delta^2 I)} [L(\theta + \nu)] - L(\theta)$, and $\delta = 0.01$, the expectation is computed by average on 1000 times sampling.
Conclusion

- We explore the escaping behavior of SGD-like processes through analyzing their continuous approximation.
- We show that thanks to the anisotropic noise, SGD could escape from sharp minima efficiently, which leads to implicit regularization effects.
- Our work raises concerns over studying the structure of SGD noise and its effect.
- Experiments support our understanding.

Poster: Wed Jun 12th
06 : 30 ~ 09 : 00 PM @ Pacific Ballroom #97