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

Zhanying Zhu, Jingfeng Wu, Bing Ya, Lei Wu, Jinwen Ma
{zhanying.zhu, pkuwjf}@pku.edu.cn

Peking University & Beijing Institute of Big Data Research

Abstract
We study the anisotropic noise of stochastic gradient descent (SGD) and its benefits on helping the dynamic escaping from minima. Concretely, we show that:
1. Compared with the isotropic noise, the curvature-aware anisotropic noise benefits to escape from sharp minima.
2. The noise of SGD is indeed aligned with the Hessian of loss surface in neural network settings.

Thus we conclude that SGD could efficiently escape from sharp minima, towards flatter ones that typically generalize well, and partly explain the implicit regularization of SGD.

The Continuous Approximation of SGD

Loss function: $L(\theta) = \frac{1}{d} \sum_{i=1}^{d} f(x_i, \theta)$, $\theta \in \mathbb{R}^d$.
SGD: $\theta_{t+1} = \theta_t - \frac{\eta}{d} \sum_{i=1}^{d} \nabla f(x_i, \theta_t)$, where $B_t$ is a randomly selected mini-batch.

A general form: gradient descent with unbiased noise
$\theta_{t+1} = \theta_t - \frac{\eta}{d} \nabla f\left(x_t, \theta_t + \epsilon_t \right) + \epsilon_t$. $\epsilon_t \sim N(0, \Sigma_t)$.

Continuous approximation with stochastic differential equation
$dt = -\nabla f\left(x_t, \theta_t \right) dt + \sqrt{\Sigma_t} dW_t$.

Fig. 1: 2D demo example. Compared dynamics are initialized at the sharp minima. See Fig. 2 for the definition of each legend: Left. The trajectory of each compared dynamics for escaping from the sharp minimum in one run. Right. Success rate of reaching the flat solution in 100 repeated runs.

Escaping Efficiency
We define the escaping efficiency as the expected increase of the potential or the loss.

Definition 1 (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
$E_0\left[L(\theta_t) - L(\theta_0)\right] = -\frac{1}{d} \sum_{i=1}^{d} \mathbb{E}\left[\nabla f(x_i, \theta_0) \right] dt + \frac{1}{2} \mathbb{E}[\text{Tr}(H_t \Sigma_t)] dt$

Suitable approximations could be shown for the above SDE (2).
$\mathbb{E}[L(\theta_t) - L(\theta_0)] = -\frac{1}{d} \sum_{i=1}^{d} \mathbb{E}\left[\nabla f(x_i, \theta_0) \right] dt + \frac{1}{2} \mathbb{E}[\text{Tr}(H_t \Sigma_t)] dt$

Then for such anisotropic $\Sigma$ and its isotropic equivalence $\Sigma_\parallel$ = $\frac{\Sigma}{d}$ under constraint (6), we have the follow ratio describing their difference in terms of escaping efficiency,
$\frac{\text{Tr}(H_t \Sigma_t)}{\text{Tr}(H_t \Sigma_\parallel)} = O\left(\frac{d}{d-1}\right)$.

Fig. 3: FashionMNIST (tweaked) experiments. Left. The first 4 eigenvalues of Hessian at $\theta_0$, the sharp minima found by SGD after 3000 iterations. Right. The projection coefficient estimation $\hat{a}$, as shown in Proposition 1.

SGD and the Curvature of Loss Surface

Proposition 2. Consider a binary classification problem with data $(x_i, y_i)_{i \in [1, n]}$, and mean square loss,
$L(\theta) = \mathbb{E}_{(x,y) \sim \mathcal{D}}[\ell(\theta; x, y) - y^2]$, where $\ell$ denotes the network and $\delta$ is a threshold activation function,
$\ell(x) = \min\{\max\{f(x), 1 - \delta\}, 0\}$.\footnote{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 $\alpha > 0$, for $\theta$ close enough to minima $\theta^*$, $\hat{a}\theta \Sigma_\parallel(\theta) \geq \alpha \lambda_1 \frac{\mathbb{E}[\text{Tr}(H_t \Sigma_\parallel)]}{\text{Tr}(H_t \Sigma_\parallel)}$.

holds almost everywhere, for $\lambda(\theta)$ and $\hat{a}$ being the maximal eigenvalue and its corresponding eigenvector of Hessian $H(\theta)$.

Fig. 4: The escape indicator of SGD noise and its isotropic equivalence. Left. One hidden layer neural networks. The solid and the dashed lines represent the value of $Tr(H_t \Sigma_\parallel)$ and $Tr(H_t \Sigma_t)$, respectively. The number of hidden nodes varies in $[2, 128, 1024]$. Right. FashionMNIST (tweaked) experiments.

Proposition 2 and 1 together illustrate that the anisotropic noise of SGD helps it escape faster from sharp minima, compared with its isotropic equivalence, which partly explains the implicit bias of SGD.

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.

References