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

{zhanxing.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. Concisely, 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: $L(\theta) := \frac{1}{N}\sum_{i=1}^{N} \ell(x_i;\theta), \theta \in \mathbb{R}^D$.

SGD: $\theta_{t+1} = \theta_t - \eta\frac{1}{m}\sum_{x\in B_t}\nabla_\theta \ell(x;\theta_t)$, where B_t is a randomly selected mini-batch.

A general form: gradient descent with unbiased noise

$$\theta_{t+1} = \theta_t - \eta\nabla_\theta L(\theta_t) + \epsilon_t,\ \epsilon_t \sim \mathcal{N}(0,\Sigma_t)\,. \tag{1}$$

Continuous approximation with stochastic differential equation

$$\mathrm{d}\theta_t = -\nabla_\theta L(\theta_t)\,\mathrm{d}t + \sqrt{\Sigma_t}\,\mathrm{d}W_t. \tag{2}$$

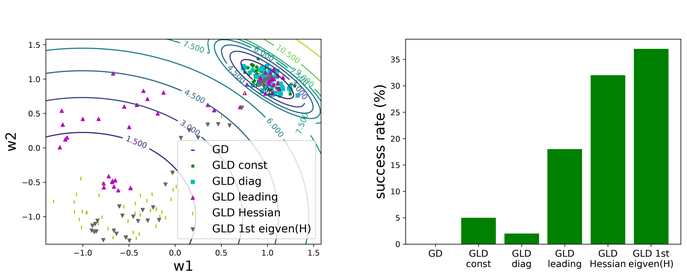

Figure 1: 2-D toy example. Compared dynamics are initialized at the sharp minima. See Figure 2 for the definition of each legend. **Left**: The trajectory of each compared dynamics for escaping from the sharp minimum in one run. **Right**: Success rate of arriving the flat solution in 100 repeated runs.

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 θ_0, then for a fixed time t small enough, the* escaping efficiency *is defined as*

$$\mathbb{E}_{\theta_t}[L(\theta_t) - L(\theta_0)] \tag{3}$$

Under suitable approximations, it could be shown that for SDE (2),

$$\mathbb{E}[L(\theta_t) - L(\theta_0)] = -\int_0^t \mathbb{E}\left[\nabla L^T \nabla L\right] + \int_0^t \frac{1}{2}\mathbb{E}\mathrm{Tr}(H_t\Sigma_t)\,\mathrm{d}t \tag{4}$$

$$\approx \frac{1}{4}\mathrm{Tr}\left(\left(I - e^{-2Ht}\right)\Sigma\right) \approx \frac{t}{2}\mathrm{Tr}\left(H\Sigma\right). \tag{5}$$

Therefore $\mathrm{Tr}\left(\mathbf{H\Sigma}\right)$serves as an important indicator for measuring the escaping behavior of noises with different structures.

To eliminate the impact of noise scale and focus on exploring the effect of noise structure, assume that

$$\textbf{given time } \mathbf{t}, \mathrm{Tr}(\mathbf{\Sigma_t}) \textbf{ is constant}. \tag{6}$$

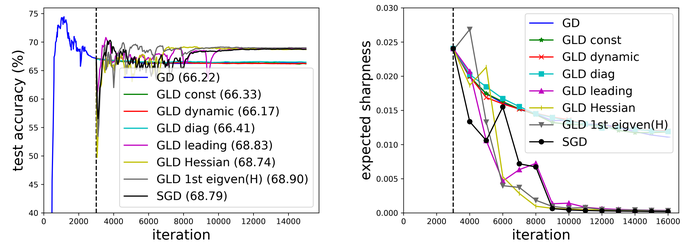

Figure 2: FashionMNIST (tweaked) experiments. Compared dynamics are initialized at θ^*_{GD} found by GD, marked by the vertical dashed line in iteration 3000. *GLD const*: constant noise; *GLD dynamic*: the isotropic equvalence of SGD noise; *GLD diag*: the diagnoal approximation of SGD noise; *GLD leading*: the low rank approximation of SGD noise; *GLD Hessian*: the noise with Hessian as covariance; *GLD 1st eigven(H)*: the rank-1 approximation of the Hessian noise. **Left**: Test accuracy versus iteration. **Right**: Expected sharpness versus iteration. Expected sharpness (the higher the sharper) is measured as $\mathbb{E}_{\nu\sim\mathcal{N}(0,\delta^2 I)}\left[L(\theta+\nu)\right] - L(\theta)$, and $\delta = 0.01$, the expectation is computed by average on 1000 times sampling.

Anisotropic Noise Helps Escape from Minima

Proposition 1 shows the anisotropic noise is superior to its isotropic equivalence, in terms of escaping from minima.

Proposition 1. *Assume $H_{D\times D}$ and $\Sigma_{D\times D}$ are both semi-positive definite. Suppose that*

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*

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

2. **Σ is "aligned" with H.** *Let u_i be the corresponding unit eigenvector of eigenvalue λ_i, for some projection coefficient $a > 0$, we have*

$$u_1^T\Sigma u_1 \geq a\lambda_1\frac{Tr\Sigma}{TrH}. \tag{8}$$

Then for such anisotropic Σ 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,*

$$\frac{Tr\left(H\Sigma\right)}{Tr(H\bar{\Sigma})} = \mathcal{O}\left(aD^{(2d-1)}\right), \quad d > \frac{1}{2}. \tag{9}$$

- Thanks to the over-parameterization of neural networks, the first condition holds naturally.
- Specifically for the noise of SGD, Proposition 2 guarantees the second condition.

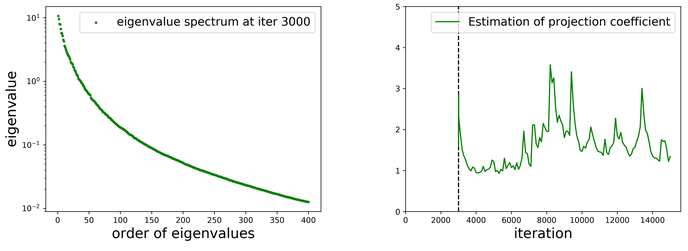

Figure 3: FashionMNIST (tweaked) experiments. **Left**: The first 400 eigenvalues of Hessian at θ^*_{GD}, the sharp minima found by GD after 3000 iterations. **Right**: The projection coefficient estimation $\hat{a} = \frac{u_1^T\Sigma u_1 \mathrm{Tr}H}{\lambda_1 \mathrm{Tr}\Sigma}$, as shown in Proposition 1.

SGD Noise and the Curvature of Loss Surface

Proposition 2. *Consider a binary classification problem with data $\{(x_i, y_i)\}_{i\in I}, y \in \{0, 1\}$, and mean square loss,*

$$L(\theta) = \mathbb{E}_{(x,y)}\|\phi \circ f(x;\theta) - y\|^2, \tag{10}$$

where f denotes the network and ϕ is a threshold activation function,

$$\phi(f) = \min\{\max\{f, \delta\}, 1-\delta\}, \tag{11}$$

δ 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 θ close enough to minima θ^*,

$$u(\theta)^T\Sigma(\theta)u(\theta) \geq a\lambda(\theta)\frac{Tr\Sigma(\theta)}{TrH(\theta)} \tag{12}$$

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

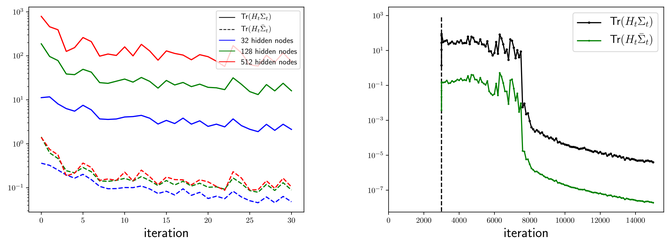

Figure 4: The escape indicator of SGD noise and its isotropic equivalence. **Left**: One hidden layer neural networks. The solid and the dotted lines represent the value of $\mathrm{Tr}(H\Sigma)$ and $\mathrm{Tr}(H\bar{\Sigma})$, respectively. The number of hidden nodes varies in $\{32, 128, 512\}$. **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

[1] Alon Brutzkus, Amir Globerson, Eran Malach, and Shai Shalev-Shwartz. Sgd learns over-parameterized networks that provably generalize on linearly separable data. *arXiv preprint arXiv:1710.10174*, 2017.

[2] Hadi Daneshmand, Jonas Kohler, Aurelien Lucchi, and Thomas Hofmann. Escaping saddles with stochastic gradients. *arXiv preprint arXiv:1803.05999*, 2018.

[3] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang. On large-batch training for deep learning: Generalization gap and sharp minima. In *In International Conference on Learning Representations (ICLR)*, 2017.

[4] Qianxiao Li, Cheng Tai, and E Weinan. Stochastic modified equations and adaptive stochastic gradient algorithms. In *International Conference on Machine Learning*, pages 2101–2110, 2017.