Power-Law Escape Rate of SGD

Takashi Mori (Riken)

joint work with Liu Ziyin, Kangqiao Liu, and Masahito Ueda

Motivation

stochastic gradient descent (SGD) is an efficient optimization method behind the success of deep learning

SGD is beneficial to generalization: SGD prefers flat minima

N. S. Keskar et al., ICLR 2017

Understanding SGD is crucial for our understanding of deep learning

Many previous studies (if not all) assume that the SGD noise is uniform

It has been pointed out that anisotropy of the SGD noise is important

Z. Zhu, J. Wu, B. Yu, L. Wu, and J. Ma, ICML 2019

Z. Xie, I. Sato, and M. Sugiyama, ICLR 2021

Our work: SGD noise strength depends on the position in the parameter space, which gives rise to the power-law escape rate from local minima

Setup

supervised learning input $x^{(\mu)}$ label $y^{(\mu)}$ $\mu = 1,2,\ldots,N$

network output $f(\theta, x^{(\mu)})$ θ: the set of network parameters

mean-squared error $\ell_\mu = \frac{1}{2}(f(\theta, x^{(\mu)}) - y^{(\mu)})^2$

Setup

supervised learning input $x^{(\mu)}$ label $y^{(\mu)}$ $\mu = 1,2,\ldots,N$

network output $f(\theta, x^{(\mu)})$ θ: the set of network parameters

mean-squared error $\ell_\mu = \frac{1}{2}(f(\theta, x^{(\mu)}) - y^{(\mu)})^2$

SGD iteration $\theta_{k+1} = \theta_k - \eta \nabla L_{B_k}(\theta_k)$ $L_{B_k}(\theta) = \frac{1}{B}\sum_{\mu \in B_k} \ell_\mu(\theta)$

η: learning rate

B: mini-batch size

Setup

supervised learning input $x^{(\mu)}$ label $y^{(\mu)}$ $\mu = 1,2,\ldots,N$

network output $f(\theta, x^{(\mu)})$ θ: the set of network parameters

mean-squared error $\ell_\mu = \frac{1}{2}(f(\theta, x^{(\mu)}) - y^{(\mu)})^2$

SGD iteration $\theta_{k+1} = \theta_k - \eta \nabla L_{B_k}(\theta_k)$ $L_{B_k}(\theta) = \frac{1}{B}\sum_{\mu \in B_k} \ell_\mu(\theta)$

η: learning rate

B: mini-batch size

continuous time $k\eta \to t \in \mathbb{R}$

Itô stochastic differential equation (SDE)

$$d\theta_t = -\nabla L(\theta_t)dt + \sqrt{\Sigma(\theta_t)} \cdot dW_t$$

Q. Li, C. Tai, and E. Weinan, ICML 2017

S. L. Smith and Q. V. Le, ICLR 2018

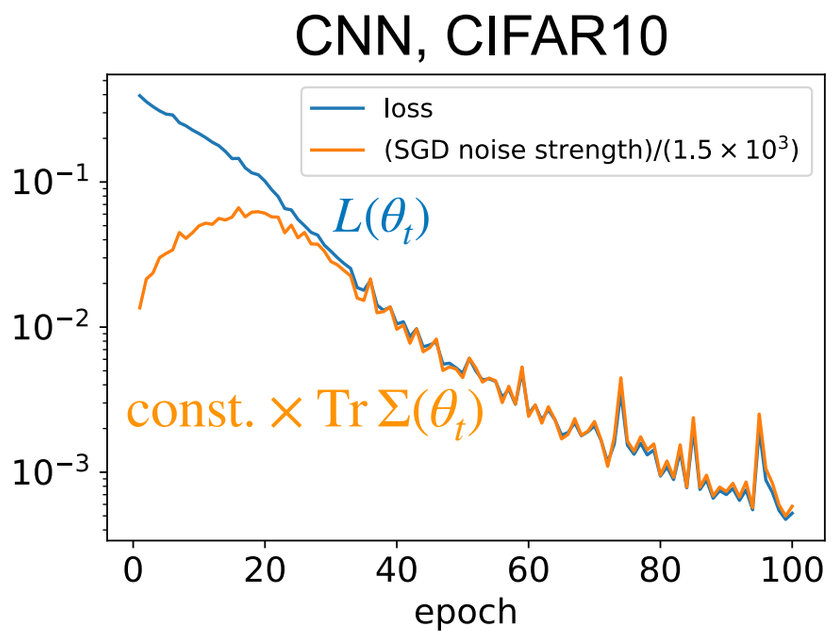
$\Sigma(\theta)$: SGD noise covariance matrix

W_t: Wiener process (Brownian motion)

Summary of the result

- For the mean-squared error, the SGD noise is proportional to the Hessian and the loss value near a (local or global) minimum at θ^*

$$\Sigma(\theta) \approx \frac{2\eta L(\theta)}{B} H(\theta^*)$$



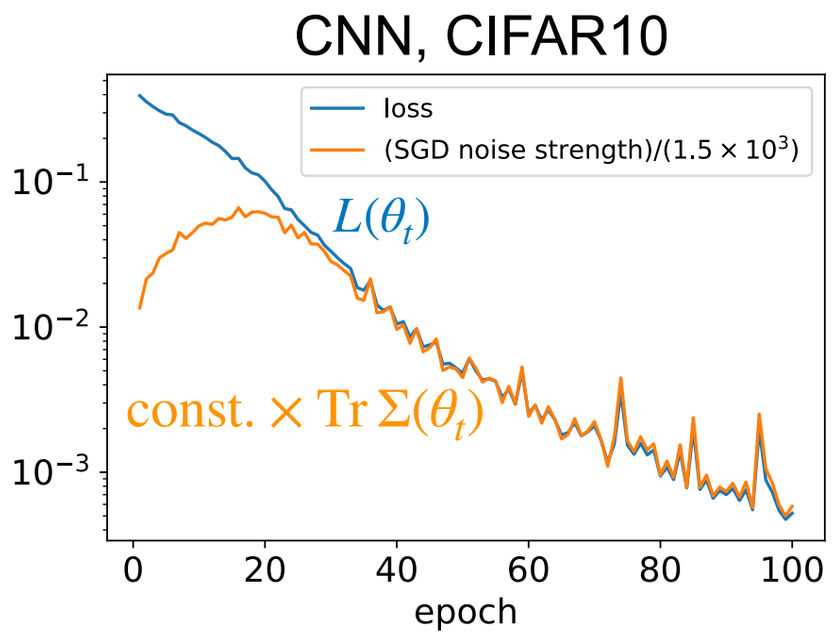
a key assumption: *decoupling approximation*

$$\frac{1}{N}\sum_{\mu=1}^{N} \ell_\mu \nabla f_\mu \nabla f_\mu^\top \approx \left(\frac{1}{N}\sum_{\mu=1}^{N} \ell_\mu\right)\left(\frac{1}{N}\sum_{\mu=1}^{N} \nabla f_\mu \nabla f_\mu^\top\right) \approx L(\theta) H(\theta^*)$$

Summary of the result

- For the mean-squared error, the SGD noise is proportional to the Hessian and the loss value near a (local or global) minimum at θ^*

$$\Sigma(\theta) \approx \frac{2\eta L(\theta)}{B} H(\theta^*)$$



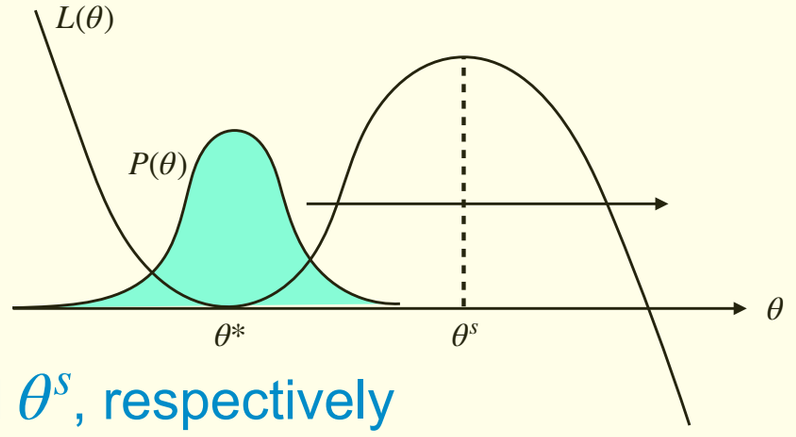
a key assumption: *decoupling approximation*

$$\frac{1}{N}\sum_{\mu=1}^{N} \ell_\mu \nabla f_\mu \nabla f_\mu^\top \approx \left(\frac{1}{N}\sum_{\mu=1}^{N} \ell_\mu\right)\left(\frac{1}{N}\sum_{\mu=1}^{N} \nabla f_\mu \nabla f_\mu^\top\right) \approx L(\theta) H(\theta^*)$$

- **[main result]** escape rate from a local minimum

(escape probability per time)

$$\kappa = \frac{\sqrt{h_e^* h_e^s}}{2\pi} \left[\frac{L(\theta^s)}{L(\theta^*)}\right]^{-\left(\frac{B}{\eta h_e^*} + 1 - \frac{n}{2}\right)}$$



h_e^*, h_e^s: Hessian eigenvalue along the escape direction at θ^* and θ^s, respectively

n: the effective dimension of a minimum (# of outlier eigenvalues of the Hessian)

Method: random time change

$$d\theta_t = -\nabla L(\theta_t)dt + \sqrt{\Sigma(\theta_t)} \cdot dW_t \qquad \Sigma(\theta) = \frac{2\eta L(\theta)}{B} H(\theta^*)$$

complicated multiplicative noise (hard to dal with) → transform to a simple additive noise

Method: random time change

$$d\theta_t = -\nabla L(\theta_t)dt + \sqrt{\Sigma(\theta_t)} \cdot dW_t \qquad \Sigma(\theta) = \frac{2\eta L(\theta)}{B} H(\theta^*)$$

complicated multiplicative noise →transform to a simple additive noise
(hard to dal with)

change of time variable $t \to \tau = \int_0^t L(\theta_s)ds$

$$\theta_t = \tilde{\theta}_\tau \qquad d\tau = L(\theta_t)dt \quad d\tilde{W}_\tau = \sqrt{L(\theta_t)}dW_t$$

$$d\tilde{\theta}_\tau = -\nabla U(\tilde{\theta}_\tau)d\tau + \sqrt{\frac{2\eta}{B}H(\theta^*)}d\tilde{W}_\tau \qquad U(\theta) = \log L(\theta)$$

SDE with simple *additive* noise on the *logarithmic loss landscape*

Method: random time change

$$d\theta_t = -\nabla L(\theta_t)dt + \sqrt{\Sigma(\theta_t)}\cdot dW_t \qquad \Sigma(\theta) = \frac{2\eta L(\theta)}{B}H(\theta^*)$$

complicated multiplicative noise (hard to dal with) → transform to a simple additive noise

change of time variable $t \to \tau = \int_0^t L(\theta_s)ds$

$$\theta_t = \tilde{\theta}_\tau \qquad d\tau = L(\theta_t)dt \quad d\tilde{W}_\tau = \sqrt{L(\theta_t)}dW_t$$

$$d\tilde{\theta}_\tau = -\nabla U(\tilde{\theta}_\tau)d\tau + \sqrt{\frac{2\eta}{B}H(\theta^*)}d\tilde{W}_\tau \qquad U(\theta) = \log L(\theta)$$

SDE with simple *additive* noise on the *logarithmic loss landscape*

Arrhenius law $\kappa \sim e^{-\Delta U/T}$

Eyring (1935), Kramers (1940), …

"temperature" $T = \eta h_e^*/B$

barrier height $\Delta U = U(\theta^s) - U(\theta^*) = \log[L(\theta^s)/L(\theta^*)]$

remark. Dependence on the effective dimension is not explained in this rough argument

Method: random time change

$$d\theta_t = -\nabla L(\theta_t)dt + \sqrt{\Sigma(\theta_t)}\cdot dW_t \qquad \Sigma(\theta) = \frac{2\eta L(\theta)}{B}H(\theta^*)$$

complicated multiplicative noise (hard to dal with) → transform to a simple additive noise

change of time variable $t \to \tau = \int_0^t L(\theta_s)ds$

$$\theta_t = \tilde{\theta}_\tau \qquad d\tau = L(\theta_t)dt \quad d\tilde{W}_\tau = \sqrt{L(\theta_t)}dW_t$$

$$d\tilde{\theta}_\tau = -\nabla U(\tilde{\theta}_\tau)d\tau + \sqrt{\frac{2\eta}{B}H(\theta^*)}d\tilde{W}_\tau \qquad U(\theta) = \log L(\theta)$$

SDE with simple *additive* noise on the *logarithmic loss landscape*

Arrhenius law $\kappa \sim e^{-\Delta U/T} = [L(\theta^s)/L(\theta^*)]^{-1/T}$

Eyring (1935), Kramers (1940), …

"temperature" $T = \eta h_e^*/B$

barrier height $\Delta U = U(\theta^s) - U(\theta^*) = \log[L(\theta^s)/L(\theta^*)]$

remark. Dependence on the effective dimension is not explained in this rough argument

Implications

$$\kappa = \frac{\sqrt{h_e^* h_e^s}}{2\pi}\left[\frac{L(\theta^s)}{L(\theta^*)}\right]^{-\left(\frac{B}{\eta h_e^*}+1-\frac{n}{2}\right)}$$

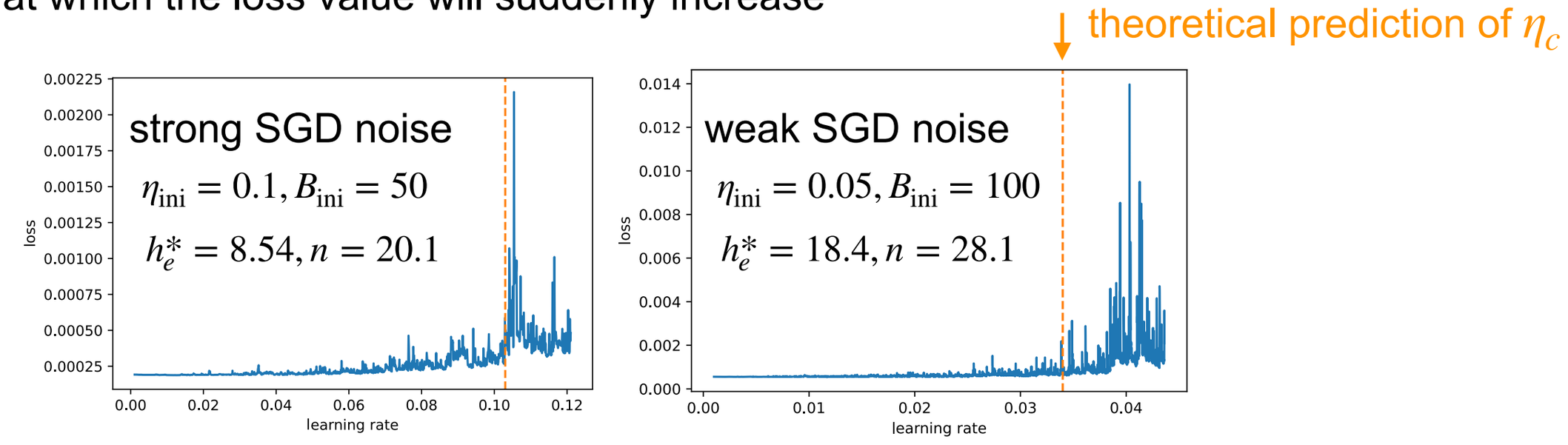
- SGD prefers flat (i.e. small h_e^*) minima *with small effective dimension* n
- stability condition $\quad \frac{B}{\eta h_e^*}+1-\frac{n}{2}>0 \longrightarrow \eta<\eta_c := \frac{2}{n-2}\frac{B}{h_e^*}$

experiment: binary classification of 5,000 samples from MNIST dataset using a FNN with 3 hidden layers, each of which has 100 neurons

(1) a minimum is found by 10^4 iterations of SGD update with fixed values of $\eta = \eta_{\text{ini}}$ and $B = B_{\text{ini}}$

(2) set $B = 8$ and gradually increase η from a small value 0.001

(3) measure η_c at which the loss value will suddenly increase



Conclusion

- SGD noise covariance is proportional to the Hessian *and the loss value*

- SGD is described by a SDE with additive noise on the *logarithmic loss landscape*

- Power-law escape rate from minima crucially depends on the effective dimension

- SGD prefers minima that are flat and have a small effective dimension