On the Noisy Gradient Descent that Generalizes as SGD

Jingfeng Wu, Wenqing Hu, Haoyi Xiong, Jun Huan,

Vladimir Braverman, Zhanxing Zhu

Johns Hopkins University, Missouri University of Science and Technology,

Baidu Research, Styling AI, Peking University

Stochastic gradient descent (SGD)

Loss function

$$L(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ell(x_i;\theta)$$

SGD

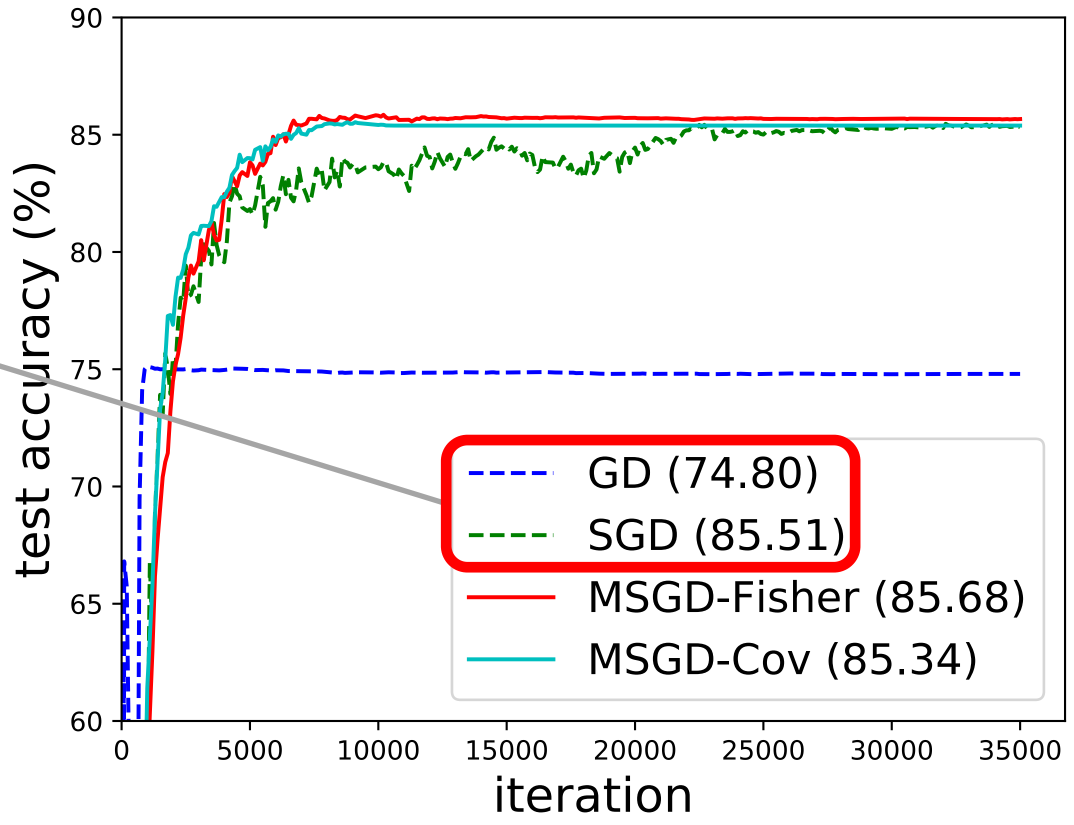
$$\begin{aligned}
\theta_{t+1} &= \theta_t - \eta\, \overbrace{\tilde{g}(\theta_t)}^{\frac{1}{b}\sum_{i\in B_t} \nabla_\theta\, \ell(x_i;\theta_t)} \\
&= \underbrace{\theta_t - \eta\, \nabla_\theta\, L(\theta_t)}_{\text{GD}} - \eta\, \underbrace{\left(\tilde{g}(\theta_t) - \nabla_\theta\, L(\theta_t)\right)}_{v_{\mathrm{sgd}}(\theta_t)}
\end{aligned}$$

(unbiased) gradient noise

Noise matters!

SGD >> GD

- How? <= Still open...
- Which? <= This work!



CIFAR-10, ResNet-18, w/o weight decay, w/o data augmentation

Which noise matters?

$$v_{\mathrm{sgd}}(\theta) = \tilde{g}(\theta) - \nabla_\theta L(\theta)$$

1. Magnitude <= YES! (e.g., Jastrzkebski et al. 2017)
2. Covariance structure <= YES! (e.g., Zhu et al. 2018)
3. Distribution class <= ? No!!! (this work)

Bernoulli? Gaussian? Levy?...

Intuition

For quadratic loss, the generalization error

$$\mathbb{E}_{x,\theta_T}\left[\ell(x;\theta_T) - \ell(x;\theta_*)\right]$$

only depends on the first two moments of θ_T, which only depend on the first two moments of $v(\theta)$.

$$\theta_{t+1} = \theta_t - \eta \underbrace{\nabla_\theta L(\theta_t)}_{\text{Linear}} - \eta v(\theta_t)$$

Noise matters! But noise class does not!!!

A closer look at the noise of SGD

$$\underbrace{v_{\mathrm{sgd}}(\theta)}_{\text{Gradient noise}} = \underbrace{\tilde{g}(\theta)}_{\nabla_\theta \mathcal{L}(\theta) \cdot \mathcal{W}_{\mathrm{sgd}}} - \underbrace{\nabla_\theta L(\theta)}_{\nabla_\theta \mathcal{L}(\theta) \cdot \frac{1}{n}\mathbb{1}} = \nabla_\theta \mathcal{L}(\theta) \cdot \underbrace{\mathcal{V}_{\mathrm{sgd}}}_{\text{Sampling noise}}$$

- Gradient matrix $\nabla_\theta \mathcal{L}(\theta) = (\nabla_\theta \ell(x_1;\theta), \ldots, \nabla_\theta \ell(x_n, \theta))$
- Sampling vector $\mathcal{W}_{\mathrm{sgd}} : \#\frac{1}{b} = b,\ \#0 = n - b$
- Sampling noise $\mathcal{V}_{\mathrm{sgd}} = \mathcal{W}_{\mathrm{sgd}} - \frac{1}{n}$

Gradient noise vs. sampling noise

$$\underbrace{v(\theta)}_{\textit{Gradient noise}} = \underbrace{\nabla_\theta \mathcal{L}(\theta)}_{\textit{Gradient matrix}} \cdot \underbrace{\mathcal{V}}_{\textit{Sampling noise}}$$

Gradient noise
- State-dependent
- Noise of gradient

Gradient matrix
- State-dependent
- Deterministic

Sampling noise
- State-independent
- Noise of mini-batch sampling

Noisy gradient descent

$$\theta_{t+1} = \underbrace{\theta_t - \eta \nabla_\theta L(\theta_t)}_{\text{GD}} \underset{\text{with}}{-\eta} \underbrace{v(\theta)}_{\text{noise}}$$

- in the same magnitude/covariance
- from different classes

Option 1: use gradient noise $v(\theta)$ ☹

Option 2: use sampling noise $v(\theta) = \nabla_\theta \mathcal{L}(\theta) \cdot \mathcal{V}$ ☺

Multiplicative SGD (MSGD)

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathcal{L}(\theta_t) \cdot \mathcal{W}, \quad \mathcal{W} = \frac{1}{n} + \mathcal{V}$$

Algorithm:

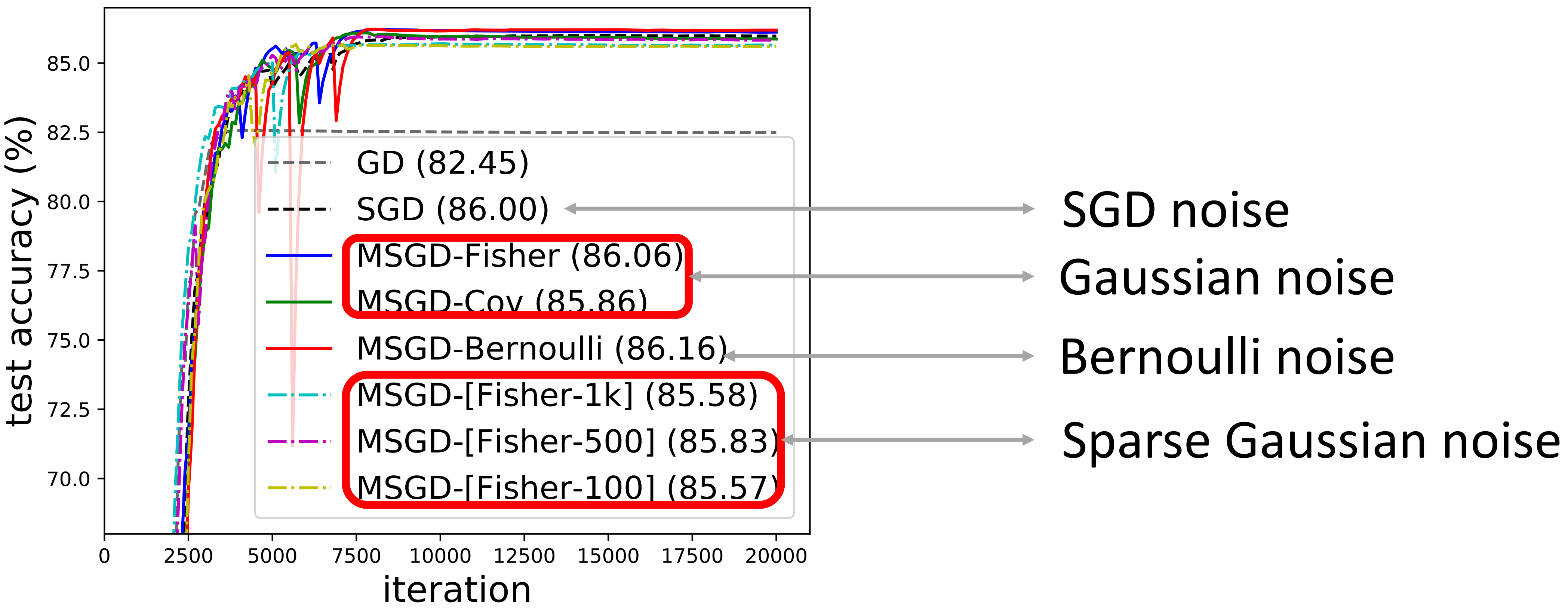
1. Generate sampling vector $\mathcal{W} = 1/n + \mathcal{V}$
2. Compute randomized loss $\tilde{L}(\theta) = \mathcal{L}(\theta) \cdot \mathcal{W}$
3. Compute stochastic gradient $\nabla_\theta \tilde{L}(\theta)$
4. Update parameters $\theta \leftarrow \theta - \eta \nabla_\theta \tilde{L}(\theta)$

Injecting noise by MSGD

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathcal{L}(\theta_t) \cdot \mathcal{W}, \quad \mathcal{W} = \frac{1}{n} + \mathcal{V}$$

1. SGD class $\mathcal{W}_{\text{sgd}} : \#\frac{1}{b} = b,\ \#0 = n - b$
2. Gaussian class $\mathcal{W}_{\text{G}} \sim \mathcal{N}\left(1/n, \text{Var}\left[\mathcal{W}_{sgd}\right]\right)$
3. "Bernoulli" class $\mathbb{P}\left(\mathcal{W}_{\text{B}}^{(i)} = \frac{1}{b}\right) = \frac{b}{n},\ \mathbb{P}\left(\mathcal{W}_{\text{B}}^{(i)} = 0\right) = 1 - \frac{b}{n}$
4. "Sparse Gaussian" class Mini-batch + Gaussian noise

Experiments



Small SVHN. More results are available in the paper!

Take Home

1. Noise class is not crucial

2. Multiplicative SGD algorithm

3. Sampling noise perspective

Get the paper!

Join our poster session for more details!