Towards Theoretically Understanding Why SGD Generalizes Better Than ADAM in Deep Learning

Pan Zhou, Jiashi Feng, Chao Ma, Caiming Xiong, Steven Hoi, and Weinan E

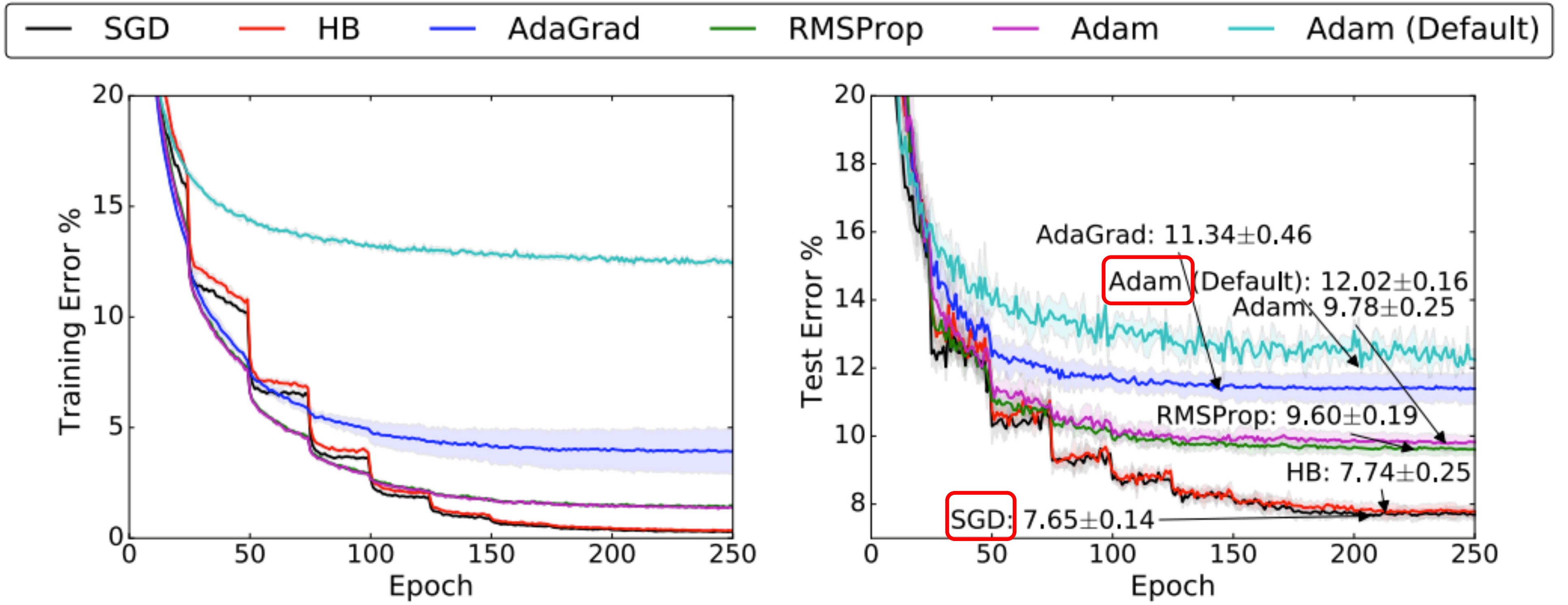
Salesforce Research

pzhou@salesforce.com

Dec 06, 2020

Observation: SGD Generalizes Better Than ADAM in Deep Learning

Important observations: **SGD often generalizes better than adaptive gradient algorithms, e.g. ADAM**

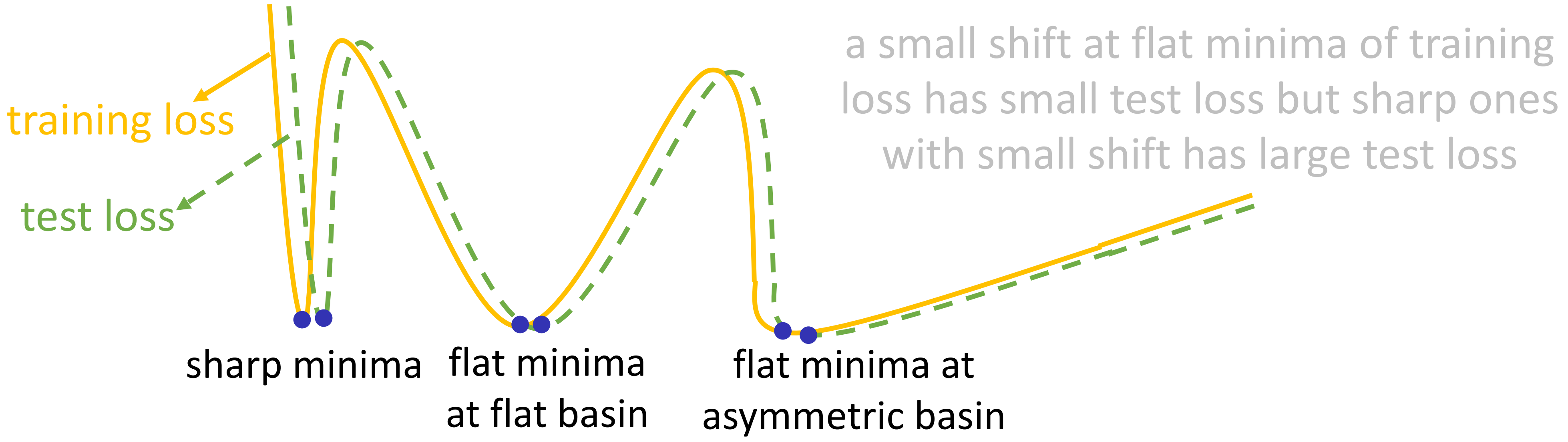


(a) CIFAR-10 (Train) **(b) CIFAR-10 (Test)** (adopted from Wilson et al. NeurIPS'17)

More similar results can be found in Keskar et al. ICLR'17, Wilson et al. NeurIPS'17, Merity et al. ICLR'18 …..

Why SGD Achieves Better Generalization Performance Than ADAM?

Empirical explanation: adaptive gradient algorithms often converge to sharp minima, while SGD prefers to find flat minima at the flat or asymmetric basins/valleys.



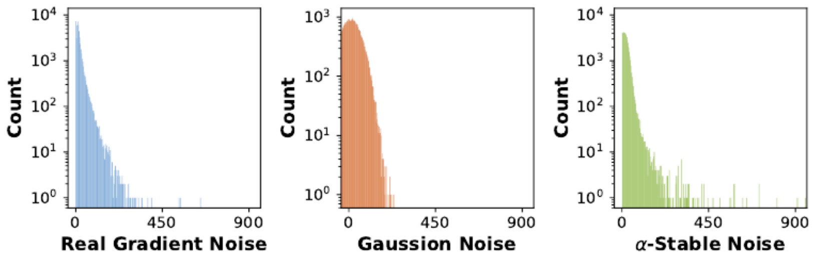
Problem: why SGD often converges to flat minima, while adaptive gradient algorithms do not?

Stochastic Differential Equation (SDE) Based Analysis

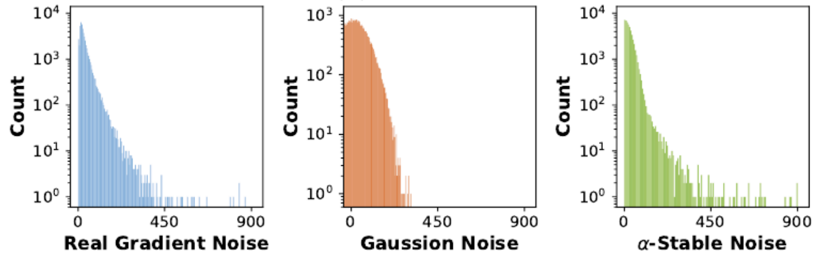
Observation: stochastic gradient noise in SGD/ADAM

approximately obey symmetric α-stable distribution

$$\boldsymbol{u}_t = \underbrace{\nabla \boldsymbol{F}(\boldsymbol{\theta}_t)}_{\text{full gradient}} - \underbrace{\nabla f_{\mathcal{S}_t}(\boldsymbol{\theta}_t)}_{\text{stochastic gradient}}$$



(a) ADAM (AlexNet on CIFAR10)



(b) SGD

Levy-driven SDE of SGD and ADAM:

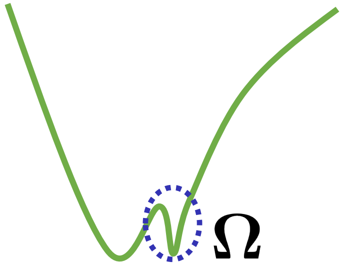
$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \nabla f_{\mathcal{S}_t}(\boldsymbol{\theta}_t) \xRightarrow{\text{SGD}} \mathrm{d}\boldsymbol{\theta}_t = -\nabla \boldsymbol{F}(\boldsymbol{\theta}_t) + \varepsilon \boldsymbol{\Sigma}_t \mathrm{d}L_t.$$

where the levy gradient noise L_t is characterized by tail index α, $\varepsilon = \eta^{(\alpha-1)/\alpha}$, the variance matrix of gradient noise $\boldsymbol{\Sigma}_t = \frac{1}{S}\left[\frac{1}{n}\sum_{i=1}^{n} \nabla f_i(\boldsymbol{\theta}_t) \nabla f_i(\boldsymbol{\theta}_t)^T - \nabla \boldsymbol{F}(\boldsymbol{\theta}_t) \nabla \boldsymbol{F}(\boldsymbol{\theta}_t)^T\right]$.

$$\begin{cases} \boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \boldsymbol{m}_t / (1-\beta_1^t) / \left(\sqrt{\boldsymbol{v}_t/(1-\beta_2^t)} + \epsilon\right), \\ \boldsymbol{m}_t = \beta_1 \boldsymbol{m}_{t-1} + (1-\beta_1) \nabla f_{\mathcal{S}_t}(\boldsymbol{\theta}_t), \\ \boldsymbol{v}_t = \beta_2 \boldsymbol{v}_{t-1} + (1-\beta_2) [\nabla f_{\mathcal{S}_t}(\boldsymbol{\theta}_t)]^2, \end{cases} \xRightarrow{\text{ADAM}} \begin{cases} \mathrm{d}\boldsymbol{\theta}_t = -\mu_t \boldsymbol{Q}_t^{-1} \boldsymbol{m}_t + \varepsilon \boldsymbol{Q}_t^{-1} \boldsymbol{\Sigma}_t \mathrm{d}L_t, \\ \mathrm{d}\boldsymbol{m}_t = \beta_1 (\nabla \boldsymbol{F}(\boldsymbol{\theta}_t) - \boldsymbol{m}_t), \\ \mathrm{d}\boldsymbol{v}_t = \beta_2 ([\nabla f_{\mathcal{S}_t}(\boldsymbol{\theta}_t)]^2 - \boldsymbol{v}_t), \end{cases}$$

where $\boldsymbol{Q}_t = \mathrm{diag}(\sqrt{\omega_t \boldsymbol{v}_t} + \epsilon)$, $\mu_t = 1/(1-e^{-\beta_1 t})$, $\omega_t = 1/(1-e^{-\beta_2 t})$

Escaping Time Analysis



- Assume SGD and ADAM get stuck in a basin $\boldsymbol{\Omega}$, i.e. $\boldsymbol{\theta}_0 \in \boldsymbol{\Omega}$
- Define the escaping time Γ from $\boldsymbol{\Omega}$ as

$$\Gamma = \inf\{t \geq 0 \mid \boldsymbol{\theta}_t \notin \boldsymbol{\Omega}^{-\varepsilon^{\gamma}}\},$$

where $\boldsymbol{\Omega}^{-\varepsilon^{\gamma}} = \{\boldsymbol{y} \in \boldsymbol{\Omega} | \text{dis}(\partial\boldsymbol{\Omega}, \boldsymbol{y}) \geq \varepsilon^{\gamma}\} \approx \boldsymbol{\Omega},$ the constant γ satisfies $\lim_{\varepsilon \to 0} \varepsilon^{\gamma} = 0.$

- Define an escaping set $\boldsymbol{\mathcal{W}}$ of basin $\boldsymbol{\Omega}$

$$\boldsymbol{\mathcal{W}} = \{\boldsymbol{y} \in \mathbb{R}^d \mid \boldsymbol{Q}_{\boldsymbol{\theta}^*}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} \boldsymbol{y} \notin \boldsymbol{\Omega}^{-\varepsilon^{\gamma}}\},$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*} = \lim_{\boldsymbol{\theta}_t \to \boldsymbol{\theta}^*} \boldsymbol{\Sigma}_t$ for both SGD and ADAM, $\boldsymbol{Q}_{\boldsymbol{\theta}^*} = \boldsymbol{I}$ in SGD and $\boldsymbol{Q}_{\boldsymbol{\theta}^*} = \lim_{\boldsymbol{\theta}_t \to \boldsymbol{\theta}^*} \boldsymbol{Q}_t$ in ADAM.

Theorem 1 (Bound of escaping time, informal).

Under proper assumptions, to escape the basin $\boldsymbol{\Omega}$, the escaping time of SGD and ADAM is

$$\Gamma = \mathcal{O}\Big(\frac{1}{\Theta m(\boldsymbol{\mathcal{W}})}\Big),$$

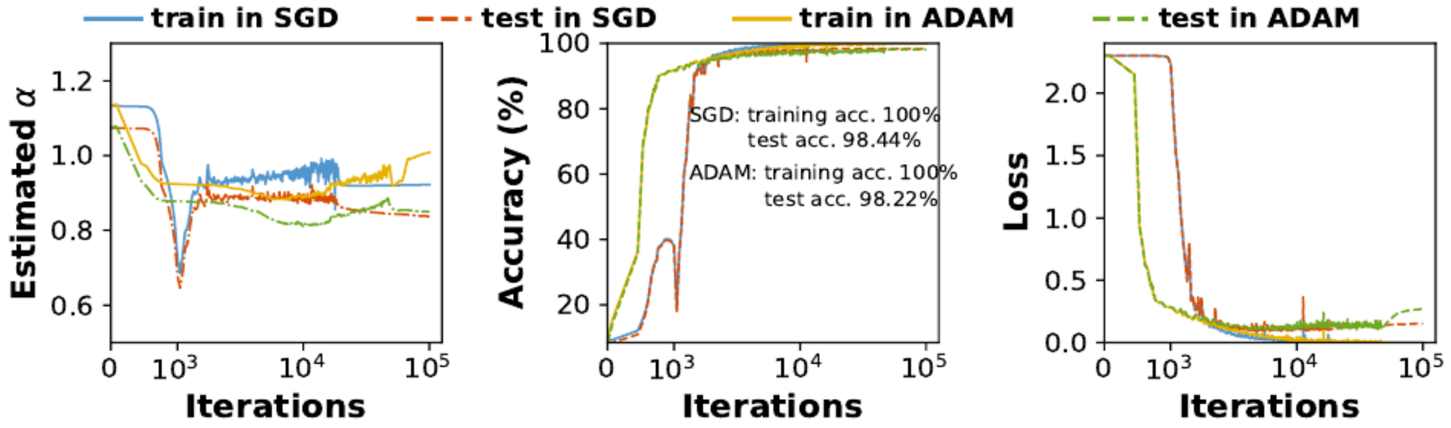
where $m(\boldsymbol{\mathcal{W}})$ is non-zero Radon measure of $\boldsymbol{\mathcal{W}}$, $\Theta = \frac{2}{\alpha}\varepsilon^{\alpha}$ in which α is the tail index of stochastic gradient noise

Escaping Time Comparison of SGD and ADAM

The escaping time Γ to escape from Ω is at the order of

$$\Gamma = \mathcal{O}\Big(\frac{1}{\Theta m(\mathcal{W})}\Big)$$

- **Factor 1.** $\Theta = \frac{2}{\alpha}\varepsilon^{\alpha}$ (the smaller tail index α , the heavier gradient noise)
  - With same learning rate ε in ADAM and SGD, the smaller tail index α , the smaller the escaping time.



(a) MNIST

  - For some iterations, SGD has smaller α , as exponential gradient average in ADAM smooths noise

Escaping Time Comparison of SGD and ADAM

The escaping time Γ to escape from Ω is at the order of

$$\Gamma = \mathcal{O}\Big(\frac{1}{\Theta m(\boldsymbol{\mathcal{W}})}\Big)$$

- **Factor 2.** $m(\boldsymbol{\mathcal{W}})$
  - It positively depends on the volume of escaping set $\boldsymbol{\mathcal{W}} = \{\boldsymbol{y} \in \mathbb{R}^d \mid \boldsymbol{Q}_{\boldsymbol{\theta}^*}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta}^*}\boldsymbol{y} \notin \boldsymbol{\Omega}^{-\varepsilon^{\gamma}}\}$

Theorem 2 (Comparison of escaping set, informal).
Under proper approximation, the escaping set of SGD is much larger than that of ADAM is

$$\boldsymbol{\mathcal{W}}_{\mathrm{ADAM}} < \boldsymbol{\mathcal{W}}_{\mathrm{SGD}}$$

which directly gives $m(\boldsymbol{\mathcal{W}}_{\mathrm{ADAM}}) < m(\boldsymbol{\mathcal{W}}_{\mathrm{SGD}})$ and $\Gamma_{\boldsymbol{\mathcal{W}}_{\mathrm{ADAM}}} > \Gamma_{\boldsymbol{\mathcal{W}}_{\mathrm{SGD}}}$.

- From Factors 1 & 2, **SGD is much more unstable, as SGD has smaller escaping time.**

SGD Prefers To Flatter Minima

From theory, **both SGD and ADAM prefers to find minima at flat or asymmetric basins.**

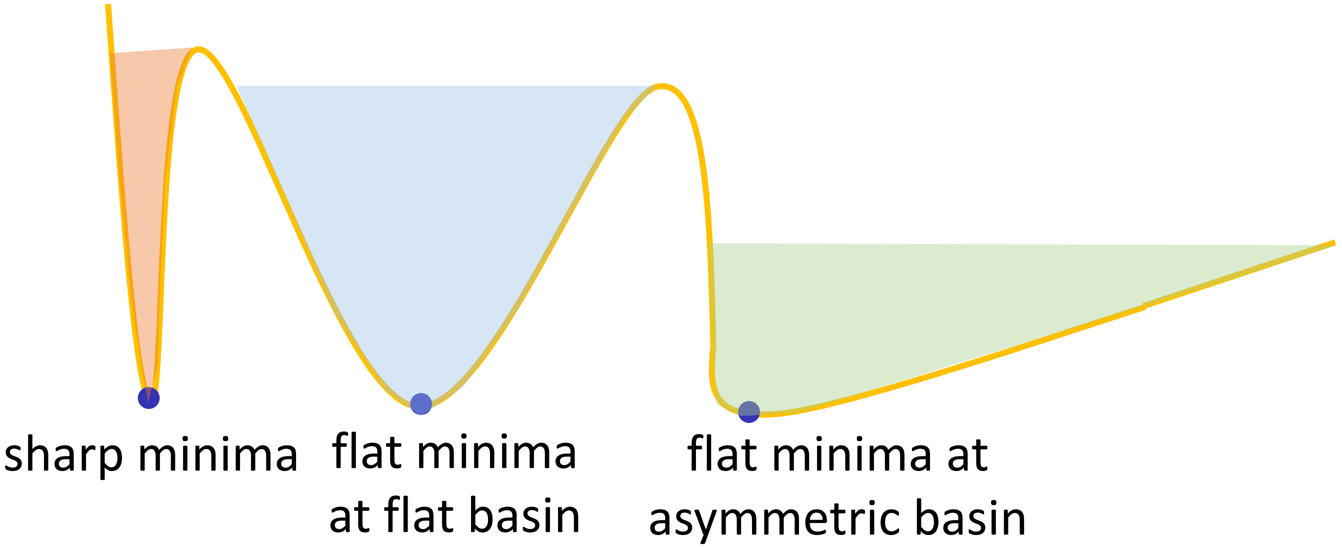
- The escaping time Γ to escape from $\boldsymbol{\Omega}$ is at the order of

$$\Gamma = \mathcal{O}\Big(\frac{1}{\Theta m(\boldsymbol{\mathcal{W}})}\Big)$$

Both SGD and ADAM prefers to escape from the basin with small volume (Radon measure)

$$\text{smaller } \boldsymbol{\Omega} \rightarrow \text{larger } \boldsymbol{\mathcal{W}} \rightarrow \text{larger } m(\boldsymbol{\mathcal{W}}) \rightarrow \text{smaller } \Gamma \rightarrow \text{more unstable at small } \boldsymbol{\Omega}$$

- Flat or asymmetric basins often have large volume than sharp one.



SGD Prefers To Flatter Minima

From theory, **both SGD and ADAM prefers to find minima at flat or asymmetric basins.**

- The escaping time Γ to escape from $\boldsymbol{\Omega}$ is at the order of

$$\Gamma = \mathcal{O}\Big(\frac{1}{\Theta m(\boldsymbol{\mathcal{W}})}\Big)$$

  Both SGD and ADAM prefers to escape from the basin with small volume (Radon measure).

$$\text{smaller } \boldsymbol{\Omega} \rightarrow \text{larger } \boldsymbol{\mathcal{W}} \rightarrow \text{larger } m(\boldsymbol{\mathcal{W}}) \rightarrow \text{smaller } \Gamma \rightarrow \text{more unstable at small } \boldsymbol{\Omega}$$

- Flat or asymmetric basins often have large volume than sharp one.

For the same basin, SGD is more unstable than ADAM. (ADAM could stuck in one basin, but SGD may not)

SGD could better escape from sharp minima and converge to flatter minima.

Thanks !