Eliminating Sharp Minima from SGD with Truncated Heavy-tailed Noise

Xingyu Wang*, Sewoong Oh†, Chang-Han Rhee*

Northwestern University*, University of Washington†

ICLR 2022

Intro: Generalization Gap and Flat Minima

- **Generalization Mystery of Stochastic Gradient Descent (SGD)**

Intro: Generalization Gap and Flat Minima

- **Generalization Mystery of Stochastic Gradient Descent (SGD)**

Training Set

Intro: Generalization Gap and Flat Minima

- **Generalization Mystery of Stochastic Gradient Descent (SGD)**

Training Set

Test Set

Image Source: https://www.flickr.com/photos/mrsdkrebs/9728631593

Intro: Generalization Gap and Flat Minima

- **Generalization Mystery of Stochastic Gradient Descent (SGD)**
- **Nonconvex Landscape, Numerous Local Minima**

Intro: Generalization Gap and Flat Minima

- **Generalization Mystery of Stochastic Gradient Descent (SGD)**
- **Nonconvex Landscape, Numerous Local Minima**

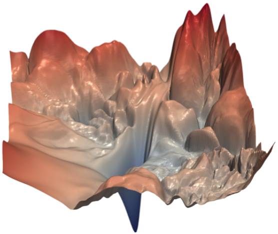


Image Source: Visualizing the Loss Landscape of Neural Nets, Li et al., 2018

Intro: Generalization Gap and Flat Minima

- **Generalization Mystery of Stochastic Gradient Descent (SGD)**
- **Flat minima** (as opposed to sharp minima) generalize better. (Jiang et al., 2020)

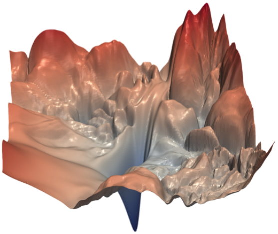
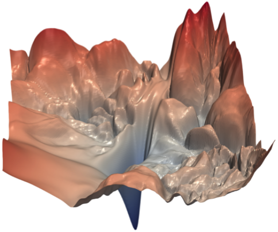


Image Source: Visualizing the Loss Landscape of Neural Nets, Li et al., 2018

Intro: Generalization Gap and Flat Minima

- **Generalization Mystery of Stochastic Gradient Descent (SGD)**
- **Flat minima** (as opposed to sharp minima) generalize better. (Jiang et al., 2020)



- **Q:** SGD prefers flat minima?

Image Source: Visualizing the Loss Landscape of Neural Nets, Li et al., 2018

Intro: Heavy-tailed SGD Prefers Flat Minima

GD $$X_j = X_{j-1} - \eta \, \nabla f(X_{j-1})$$

Intro: Heavy-tailed SGD Prefers Flat Minima

$$\text{SGD} \qquad X_j = X_{j-1} - \eta\big(\nabla f(X_{j-1}) + Z_j\big)$$

Intro: Heavy-tailed SGD Prefers Flat Minima

Traditional Assumption: Light-tailed ↘

SGD $$X_j = X_{j-1} - \eta\big(\nabla f(X_{j-1}) + Z_j\big)$$

Intro: Heavy-tailed SGD Prefers Flat Minima

~~Traditional Assumption: Light-tailed~~ ↘

SGD $$X_j = X_{j-1} - \eta\big(\nabla f(X_{j-1}) + Z_j\big)$$

Intro: Heavy-tailed SGD Prefers Flat Minima

~~Traditional Assumption: Light-tailed~~ ↘

SGD $$X_j = X_{j-1} - \eta\big(\nabla f(X_{j-1}) + Z_j\big)$$

↖ Heavy-tailed

Intro: Heavy-tailed SGD Prefers Flat Minima

~~Traditional Assumption: Light-tailed~~ ↘

$$\text{SGD} \qquad X_j = X_{j-1} - \eta\big(\nabla f(X_{j-1}) + Z_j\big)$$

↖ Heavy-tailed

- **Heavy-tailed Assumption:** $\mathbb{E}Z_j = 0, \ \ \mathbb{P}(\|Z_j\| > x) \approx x^{-\alpha}$

Intro: Heavy-tailed SGD Prefers Flat Minima

~~Traditional Assumption: Light-tailed~~ ↘

$$\text{SGD} \qquad X_j = X_{j-1} - \eta\big(\nabla f(X_{j-1}) + Z_j\big)$$

↖ Heavy-tailed

- **Heavy-tailed Assumption:** $\mathbb{E}Z_j = 0, \;\; \mathbb{P}(\|Z_j\| > x) \approx x^{-\alpha}$
- **Heavy tails in deep learning:** Srinivasan et al. (2021); Garg et al. (2021);

Intro: Heavy-tailed SGD Prefers Flat Minima

~~Traditional Assumption: Light-tailed~~ ↘

$$\text{SGD} \qquad X_j = X_{j-1} - \eta\big(\nabla f(X_{j-1}) + Z_j\big)$$

↖ Heavy-tailed

- **Heavy-tailed Assumption:** $\mathbb{E}Z_j = 0,\ \ \mathbb{P}(\|Z_j\| > x) \approx x^{-\alpha}$
- **Heavy tails in deep learning:** Srinivasan et al. (2021); Garg et al. (2021);
- **Why heavy tails arise:** Hodgkinson & Mahoney (2020);

Intro: Heavy-tailed SGD Prefers Flat Minima

~~Traditional Assumption: Light-tailed~~ ↘

$$\text{SGD} \qquad X_j = X_{j-1} - \eta\big(\nabla f(X_{j-1}) + Z_j\big)$$

↖ Heavy-tailed

- **Heavy-tailed Assumption:** $\mathbb{E}Z_j = 0, \ \mathbb{P}(\|Z_j\| > x) \approx x^{-\alpha}$
- **Heavy tails in deep learning:** Srinivasan et al. (2021); Garg et al. (2021);
- **Why heavy tails arise:** Hodgkinson & Mahoney (2020);
- **Heavy-tailed SGD prefers flat minima:** Simsekli et al. (2019)

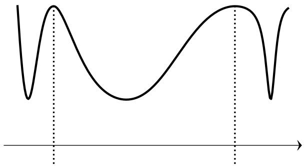
Intro: Heavy-tailed SGD Prefers Flat Minima

~~Traditional Assumption: Light-tailed~~ ↘

$$\textit{SGD} \qquad X_j = X_{j-1} - \eta\big(\nabla f(X_{j-1}) + Z_j\big)$$

↖ Heavy-tailed

- **Heavy-tailed Assumption:** $\mathbb{E}Z_j = 0, \ \mathbb{P}(\|Z_j\| > x) \approx x^{-\alpha}$
- **Heavy tails in deep learning:** Srinivasan et al. (2021); Garg et al. (2021);
- **Why heavy tails arise:** Hodgkinson & Mahoney (2020);
- **Heavy-tailed SGD prefers flat minima:** Simsekli et al. (2019)



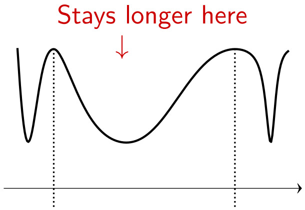
Intro: Heavy-tailed SGD Prefers Flat Minima

~~Traditional Assumption: Light-tailed~~ ↘

$$\text{SGD} \qquad X_j = X_{j-1} - \eta\big(\nabla f(X_{j-1}) + Z_j\big)$$

↖ Heavy-tailed

- **Heavy-tailed Assumption:** $\mathbb{E}Z_j = 0, \ \ \mathbb{P}(\|Z_j\| > x) \approx x^{-\alpha}$
- **Heavy tails in deep learning:** Srinivasan et al. (2021); Garg et al. (2021);
- **Why heavy tails arise:** Hodgkinson & Mahoney (2020);
- **Heavy-tailed SGD prefers flat minima:** Simsekli et al. (2019)



Intro: Heavy-tailed SGD Prefers Flat Minima

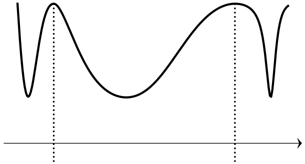
~~Traditional Assumption: Light-tailed~~ ↘

SGD $$X_j = X_{j-1} - \eta\big(\nabla f(X_{j-1}) + Z_j\big)$$

↖ Heavy-tailed

- **Heavy-tailed Assumption:** $\mathbb{E}Z_j = 0$, $\mathbb{P}(\|Z_j\| > x) \approx x^{-\alpha}$
- **Heavy tails in deep learning:** Srinivasan et al. (2021); Garg et al. (2021);
- **Why heavy tails arise:** Hodgkinson & Mahoney (2020);
- **Heavy-tailed SGD prefers flat minima:** Simsekli et al. (2019)

Our Work: Complete Elimination of Sharp Minima



Theoretical Results

$$X_j = X_{j-1} - \varphi_b\big(\eta \nabla f(X_{j-1}) + \eta Z_j\big); \quad \varphi_b(x) = \min\{b, \|x\|\} \cdot \frac{x}{\|x\|}$$

Theoretical Results

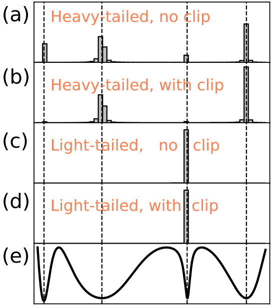
Gradient Clipping

$$X_j = X_{j-1} - \underset{\downarrow}{\varphi_b}\big(\eta \nabla f(X_{j-1}) + \eta Z_j\big); \quad \varphi_b(x) = \min\{b, \|x\|\} \cdot \frac{x}{\|x\|}$$

Theoretical Results

Gradient Clipping
↓

$$X_j = X_{j-1} - \varphi_b\big(\eta \nabla f(X_{j-1}) + \eta Z_j\big); \quad \varphi_b(x) = \min\{b, \|x\|\} \cdot \frac{x}{\|x\|}$$

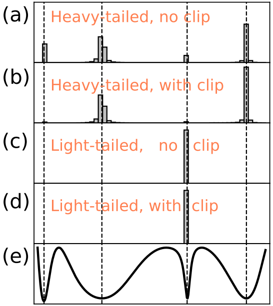


Theoretical Results

Theorem (Wang, Oh, Rhee, 2022)

Under suitable conditions, for any β large enough and any $t > 0$,

$$\frac{1}{\lfloor t/\eta^{\beta} \rfloor} \int_0^{\lfloor t/\eta^{\beta} \rfloor} \mathbb{1}\Big\{ X^{\eta}_{\lfloor u \rfloor} \textit{ is around ``narrow'' minima} \Big\} du \xrightarrow{\mathrm{P}} 0 \textit{ as } \eta \downarrow 0.$$



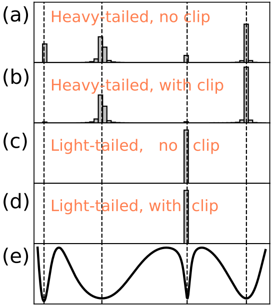
Theoretical Results

Theorem (Wang, Oh, Rhee, 2022)

Under suitable conditions, for any β large enough and any $t > 0$,

$$\frac{1}{\lfloor t/\eta^{\beta} \rfloor} \int_0^{\lfloor t/\eta^{\beta} \rfloor} \mathbb{1}\Big\{ X^{\eta}_{\lfloor u \rfloor} \text{ is around “narrow” minima} \Big\} du \xrightarrow{\mathrm{P}} 0 \text{ as } \eta \downarrow 0.$$

↖Proportion of time at narrow minima



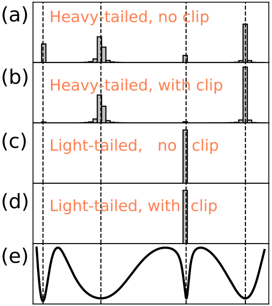
Theoretical Results

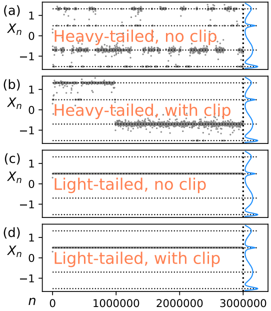
Theorem (Wang, Oh, Rhee, 2022)

Under suitable conditions, for any β large enough and any $t > 0$,

$$\frac{1}{\lfloor t/\eta^{\beta} \rfloor} \int_0^{\lfloor t/\eta^{\beta} \rfloor} \mathbb{1}\Big\{ X^{\eta}_{\lfloor u \rfloor} \text{ is around ``narrow'' minima} \Big\} du \xrightarrow{P} 0 \text{ as } \eta \downarrow 0.$$

↖Proportion of time at narrow minima



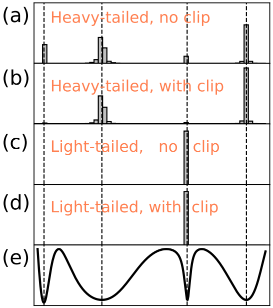


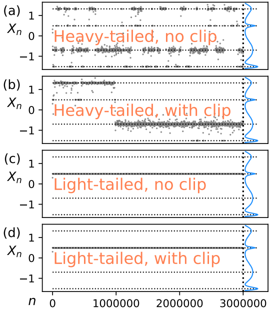
Theoretical Results

Theorem (Wang, Oh, Rhee, 2022)

Under suitable conditions,

$$\{X^{\eta}_{\lfloor t \cdot \lambda(\eta) \rfloor} : \ t \geq 0\} \Rightarrow \{Y_t : \ t \geq 0\} \ \textit{as} \ \eta \downarrow 0$$





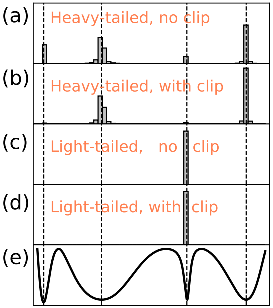
Theoretical Results

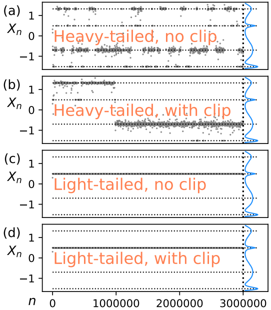
Theorem (Wang, Oh, Rhee, 2022)

Under suitable conditions,

time-scaled SGD

$$\{X^{\eta}_{\lfloor t\cdot\lambda(\eta)\rfloor} : \ t \geq 0\} \Rightarrow \{Y_t : \ t \geq 0\} \text{ as } \eta \downarrow 0$$





Theoretical Results

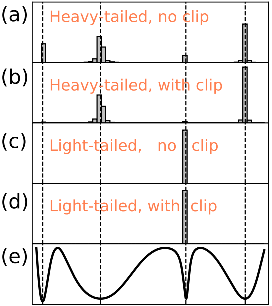
Theorem (Wang, Oh, Rhee, 2022)

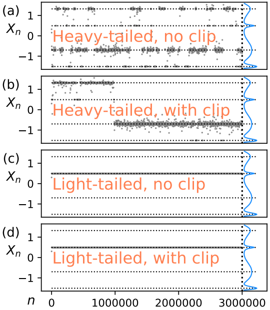
Under suitable conditions,

time-scaled SGD

$$\{X^{\eta}_{\lfloor t\cdot\lambda(\eta)\rfloor} : \ t \geq 0\} \Rightarrow \{Y_t : \ t \geq 0\} \text{ as } \eta \downarrow 0$$

where Y is a continuous-time Markov chain that only visits "wide" minima.





Theoretical Results

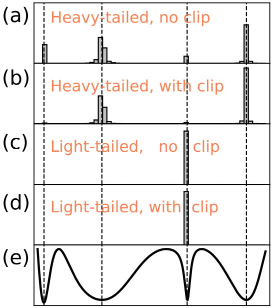
Theorem (Wang, Oh, Rhee, 2022)

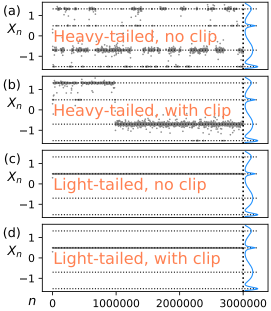
Under suitable conditions,

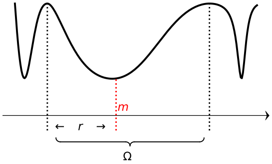
time-scaled SGD

$$\{X^{\eta}_{\lfloor t\cdot\lambda(\eta)\rfloor} : \ t \geq 0\} \Rightarrow \{Y_t : \ t \geq 0\} \text{ as } \eta \downarrow 0$$

where Y is a continuous-time Markov chain that only visits "wide" minima.







Theoretical Results

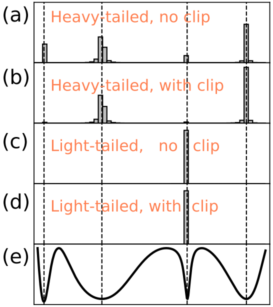
Theorem (Wang, Oh, Rhee, 2022)

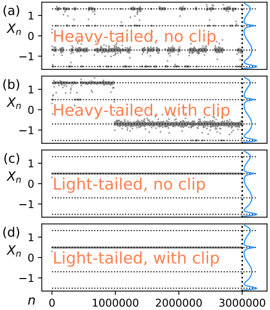
Under suitable conditions,

time-scaled SGD

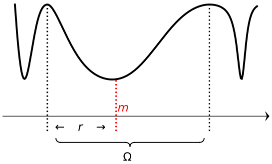
$$\{X^{\eta}_{\lfloor t\cdot\lambda(\eta)\rfloor} :\ t \geq 0\} \Rightarrow \{Y_t :\ t \geq 0\} \text{ as } \eta \downarrow 0$$

where Y is a continuous-time Markov chain that only visits "wide" minima.





Required # of jumps: $l^* = \lceil r/b \rceil$



Theoretical Results

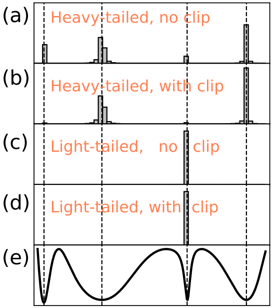
Theorem (Wang, Oh, Rhee, 2022)

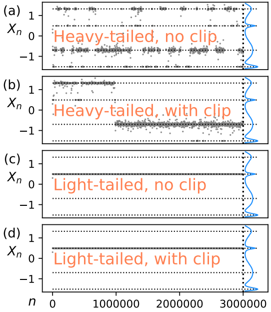
Under suitable conditions,

$$\{X^{\eta}_{\lfloor t\cdot\lambda(\eta)\rfloor} : \ t \geq 0\} \Rightarrow \{Y_t : \ t \geq 0\} \ \textit{as}\ \eta \downarrow 0$$

(↙ time-scaled SGD)

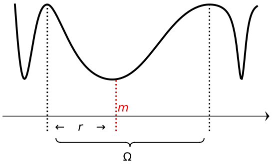
where Y is a continuous-time Markov chain that only visits "wide" minima.





Required # of jumps: $l^* = \lceil r/b \rceil$

Exit Time: $O\big(1/\eta^{\alpha+(l^*-1)(\alpha-1)}\big)$



Tail Inflation and Truncation in Deep Learning

- X: current weights; g_Y: stochastic gradient under method Y.

Tail Inflation and Truncation in Deep Learning

- X: current weights; g_Y: stochastic gradient under method Y.
- **Our Method**: $X \leftarrow X - \varphi_b\big(\eta \cdot g_{\text{heavy}}(X)\big)$ where

Tail Inflation and Truncation in Deep Learning

- X: current weights; g_Y: stochastic gradient under method Y.
- **Our Method**: $X \leftarrow X - \varphi_b\big(\eta \cdot g_{\text{heavy}}(X)\big)$ where

$$g_{\text{heavy}}(X) \triangleq g_{\text{SGD}}(X) + \text{“Heavy-tailed Noise”}$$

Tail Inflation and Truncation in Deep Learning

Test accuracy	LB	SB	SB + Clip	SB + Noise	Our 1	Our 2
CorrputedFMNIST, LeNet	68.66%	69.20%	68.77%	64.43%	69.47%	**70.06%**
SVHN, VGG11	82.87%	85.92%	85.95%	38.85%	**88.42%**	88.37%
CIFAR10, VGG11	69.39%	74.42%	74.38%	40.50%	75.69%	**75.87%**
Expected Sharpness	LB	SB	SB + Clip	SB + Noise	Our 1	Our 2
CorrputedFMNIST, LeNet	0.032	0.008	0.009	0.047	0.003	**0.002**
SVHN, VGG11	0.694	0.037	0.041	0.012	**0.002**	0.005
CIFAR10, VGG11	2.043	0.050	0.039	2.046	**0.024**	0.037

Tail Inflation and Truncation in Deep Learning

Test accuracy	LB	SB	SB + Clip	SB + Noise	Our 1	Our 2
CorrputedFMNIST, LeNet	68.66%	69.20%	68.77%	64.43%	69.47%	**70.06%**
SVHN, VGG11	82.87%	85.92%	85.95%	38.85%	**88.42%**	88.37%
CIFAR10, VGG11	69.39%	74.42%	74.38%	40.50%	75.69%	**75.87%**
Expected Sharpness	**LB**	**SB**	**SB + Clip**	**SB + Noise**	**Our 1**	**Our 2**
CorrputedFMNIST, LeNet	0.032	0.008	0.009	0.047	0.003	**0.002**
SVHN, VGG11	0.694	0.037	0.041	0.012	**0.002**	0.005
CIFAR10, VGG11	2.043	0.050	0.039	2.046	**0.024**	0.037

- **Flatter geometry & Improved generalization performance**

Tail Inflation and Truncation in Deep Learning

Test accuracy	LB	SB	SB + Clip	SB + Noise	Our 1	Our 2
CorrputedFMNIST, LeNet	68.66%	69.20%	68.77%	64.43%	69.47%	**70.06%**
SVHN, VGG11	82.87%	85.92%	85.95%	38.85%	**88.42%**	88.37%
CIFAR10, VGG11	69.39%	74.42%	74.38%	40.50%	75.69%	**75.87%**
Expected Sharpness	**LB**	**SB**	**SB + Clip**	**SB + Noise**	**Our 1**	**Our 2**
CorrputedFMNIST, LeNet	0.032	0.008	0.009	0.047	0.003	**0.002**
SVHN, VGG11	0.694	0.037	0.041	0.012	**0.002**	0.005
CIFAR10, VGG11	2.043	0.050	0.039	2.046	**0.024**	0.037

- **Flatter geometry & Improved generalization performance**
- Requires both **heavy-tailed** noise and **truncation**

Tail Inflation and Truncation in Deep Learning

CIFAR10-VGG11	SB + Clip	Our 1	Our 2
Test Accuracy	89.54%	**90.76%**	90.45%
Expected Sharpness	0.167	**0.085**	0.096
PAC-Bayes Sharpness	1.31×10^4	$\mathbf{9 \times 10^3}$	10^4
Maximal Sharpness	1.66×10^4	1.29×10^4	$\mathbf{1.22 \times 10^4}$
CIFAR100-VGG16	SB + Clip	Our 1	Our 2
Test Accuracy	56.32%	**65.44%**	62.99%
Expected Sharpness	0.857	**0.441**	0.479
PAC-Bayes Sharpness	2.49×10^4	$\mathbf{1.9 \times 10^4}$	1.98×10^4
Maximal Sharpness	2.75×10^4	$\mathbf{2.12 \times 10^4}$	2.16×10^4

- **More training techniques:** Data augmentation, learning rate scheduler.

Conclusion

- **Theoretical Contribution**
  - Rigorously established that truncated heavy-tailed noises can eliminate sharp minima
  - First exit time analysis and metastability for heavy-tailed SGD
- **Algorithmic Contribution**
  - Proposed a tail-inflation strategy to find flatter solution with better generalization