A TAIL-INDEX ANALYSIS OF STOCHASTIC GRADIENT NOISE IN DEEP NEURAL NETWORKS

Umut Şimşekli

LTCI, Télécom ParisTech,
Université Paris-Saclay

TELECOM ParisTech

Artificial Intelligence and Physics
March 21, 2019

CREDITS

Joint work with

- Levent Sagun – EPFL
- Mert Gürbüzbalaban – Rutgers University

Paper is on arxiv:

https://arxiv.org/pdf/1901.06053.pdf

DEEP LEARNING & SGD

- Deep learning (in general)

$$\mathbf{w}^\star = \underbrace{\arg\min_{\mathbf{w}\in\mathbb{R}^p}}_{\text{network weights}} \Big\{ \underbrace{f(\mathbf{w})}_{\text{non-convex cost func.}} \triangleq \frac{1}{n}\sum_{\underbrace{i=1}_{\text{data points}}}^{n} f^{(i)}(\mathbf{w}) \Big\}$$

which one is better?

- Optimization Algorithm – **S**tochastic **G**radiend **D**escent → local optimum

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \underbrace{\eta}_{\substack{\text{step-size} \\ \text{(learning rate)}}} \underbrace{\nabla \tilde{f}_k(\mathbf{w}_k)}_{\substack{\text{stochastic} \\ \text{gradient}}} \longrightarrow \nabla \tilde{f}_k(\mathbf{w}) \triangleq \frac{1}{\underbrace{b}_{\substack{\text{minibatch} \\ \text{size}}}} \sum_{i\in \underbrace{\Omega_k}_{\text{minibatch}}} \nabla f^{(i)}(\mathbf{w})$$

Assumption: Stochastic gradients are **unbiased**

‘WIDE MINIMA’ PHENOMENON

- **“The wider the minimum, the better the performance on the test set”**

Hochreiter & Schmidhuber, 1997

Current folklore:

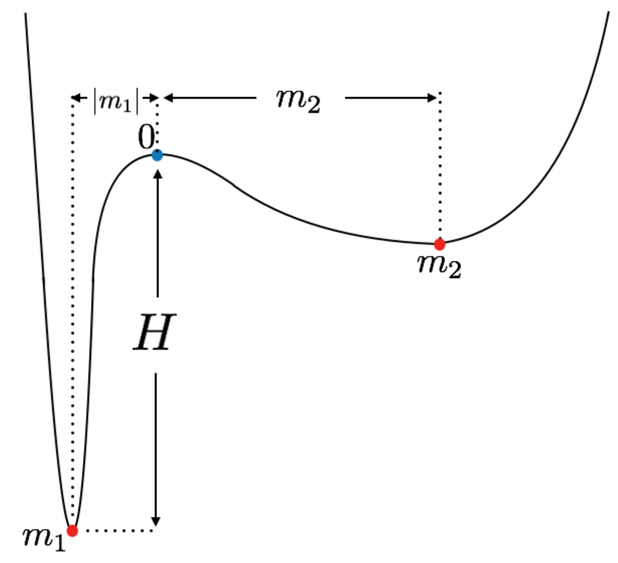
- Gradient Descent (full batch) overfits:

  smaller minibatch → better performance

  (Keskar et al., 2017)

- SGD ‘prefers’ wide minima

  (Jastrzebski et al.,2017)



MAIN QUESTION (THAT WE ASK IN THIS STUDY)

Why would SGD 'prefer' wide minima?

- Something must be going on with the "noise"

$$U_k(\mathbf{w}) \triangleq \left[\nabla \tilde{f}_k(\mathbf{w}) - \nabla f(\mathbf{w})\right] = \frac{1}{b}\sum_{i\in\Omega_k}\left[\nabla f^{(i)}(\mathbf{w}) - \nabla f(\mathbf{w})\right]$$

stochastic gradient noise

Zero-mean and i.i.d. random variables (unbiasedness)

- Additional assumption: $U_k(\mathbf{w})$ has **finite variance**

(Mandt et al.'16, Jastrzebski et al.'17, Zhu et al.'18 …)

- **Central Limit Theorem (CLT) →** $U_k(\mathbf{w}) \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$

DIFFUSION REPRESENTATION

$U_k(\mathbf{w}) \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$

- With the **Gaussian** assumption:

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta \nabla f(\mathbf{w}_k) + \sqrt{\eta}\sqrt{\eta\sigma^2} Z_k$$

standard Gaussian random variable

- Small step-size → the stochastic differential equation **(SDE)**:

$$\mathrm{d}\mathbf{w}_t = -\nabla f(\mathbf{w}_t)\mathrm{d}t + \sqrt{\eta\sigma^2}\mathrm{dB}_t$$

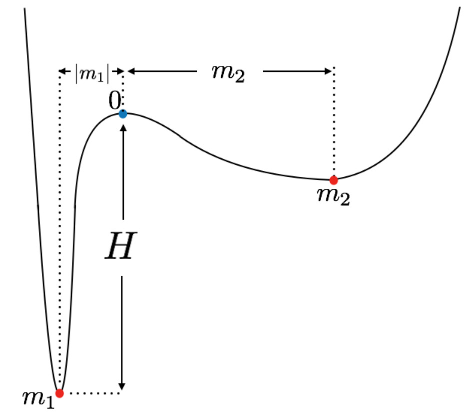
Brownian Motion

- We can now use all the nice + rich theory of SDEs!
- Jastrzebski et al. → the width is determined by: η/b
- They are not the only ones…

SOME ISSUES

- The results are based on the **invariant** distribution:

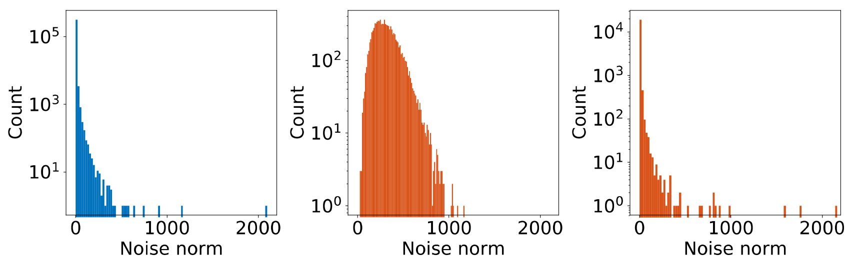
  requires exp(O(p)) many iterations → doesn't reflect the practice



- Confliction with **metastability** results:

"Transition time" ≈ **exp**(H) x **poly**($|m_1|$)

- How accurate is the Gaussianity assumption?



(a) Real (b) Gaussian (c) ???

Can we find a better assumption?

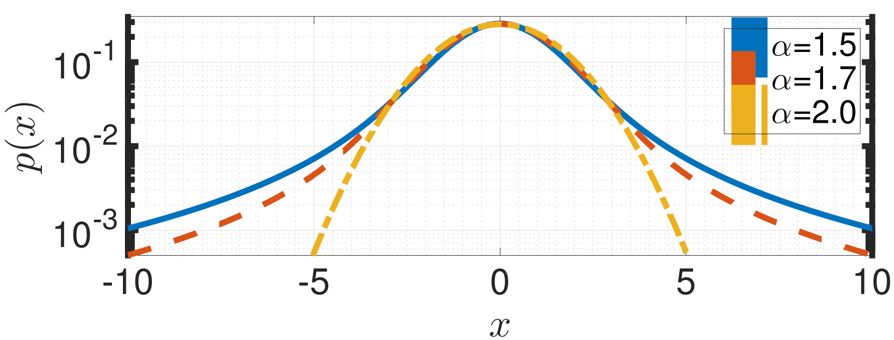
GENERALIZED CLT

- Go back to:

$$\underbrace{U_k(\mathbf{w}) \triangleq \left[\nabla \tilde{f}_k(\mathbf{w}) - \nabla f(\mathbf{w})\right]}_{\text{stochastic gradient noise}} = \frac{1}{b} \sum_{i \in \Omega_k} \Big[\underbrace{\nabla f^{(i)}(\mathbf{w}) - \nabla f(\mathbf{w})}_{\text{Zero-mean and i.i.d. random variables}}\Big]$$

- In many domains the "finite variance" might not hold

- **Extended** CLT: $U_k(\mathbf{w})$ converges → **heavy-tailed α-stable r.v.**



Gaussian	**when α=2**
Infinite variance	**when α≠2**

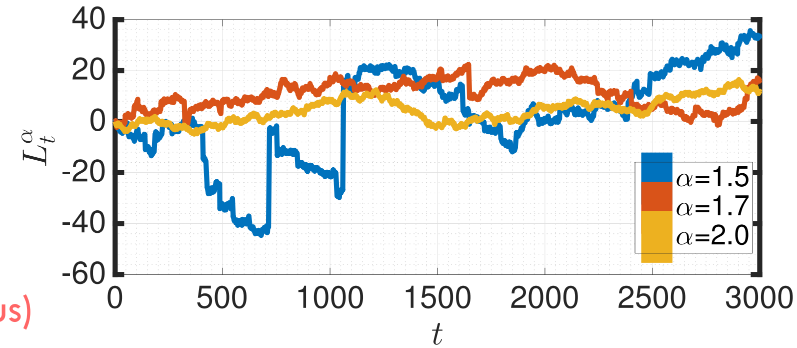
NEW FRAMEWORK + IMPLICATIONS

- Proposed assumption: $U_k(\mathbf{w}) \sim \mathcal{S}\alpha\mathcal{S}(\sigma(\mathbf{w}))$

- The resulting SDE:

$$\mathrm{d}\mathbf{w}_t = -\nabla f(\mathbf{w}_t)\mathrm{d}t + \eta^{(\alpha-1)/\alpha}\sigma\,\mathrm{dL}_t^{\alpha}$$

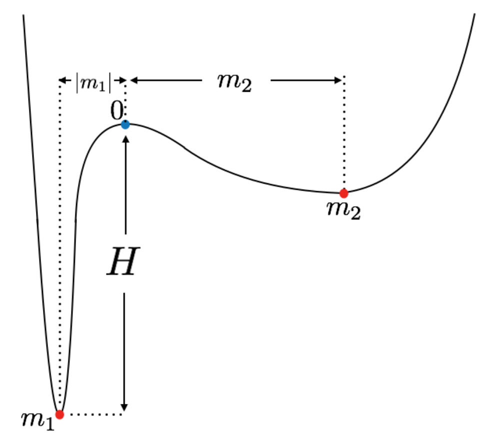
Lévy Motion (discontinuous)



- Metastability: *(Pavlyukevich'07)*

Transition time - does **not** depend on **H**,

- **poly**(|m₁|, α)



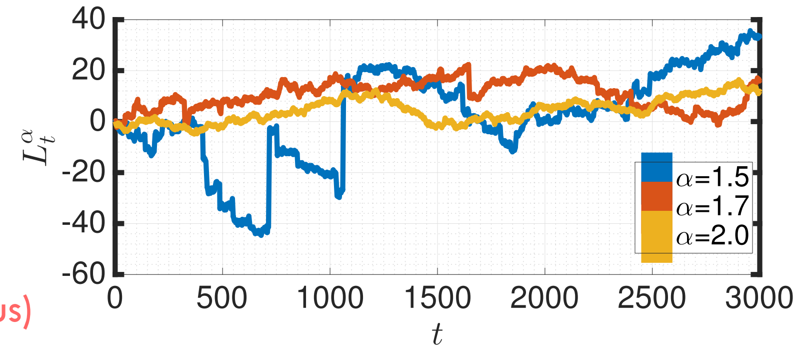
NEW FRAMEWORK + IMPLICATIONS

- Proposed assumption: $U_k(\mathbf{w}) \sim \mathcal{S}\alpha\mathcal{S}(\sigma(\mathbf{w}))$

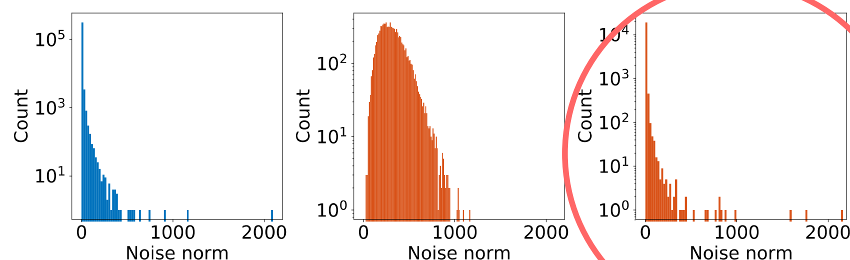
- The resulting SDE:

$$\mathrm{d}\mathbf{w}_t = -\nabla f(\mathbf{w}_t)\mathrm{d}t + \eta^{(\alpha-1)/\alpha}\sigma\,\mathrm{dL}_t^\alpha$$

Lévy Motion
(discontinuous)



- **BIG QUESTION:** is SGD noise really α-stable?



(a) Real (b) Gaussian (c) α-stable

EMPIRICAL STUDY

- Aim: estimate $\alpha \rightarrow$ if $\alpha = 2$ the noise is Gaussian

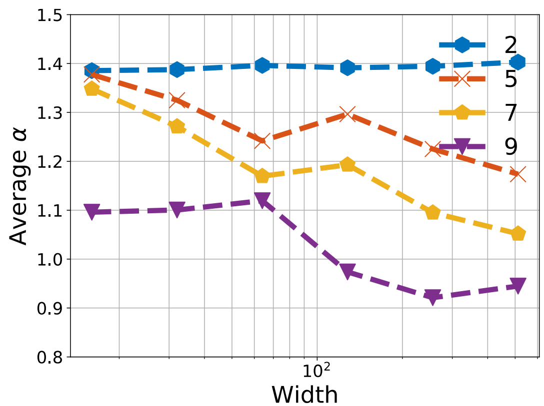
We have tested different (in the paper)

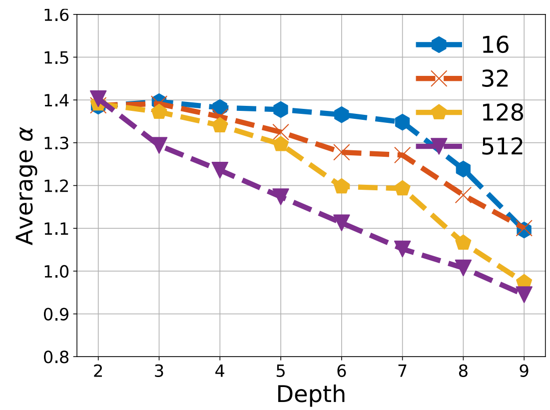
1) **datasets** (MNIST, CIFAR10, CIFAR100)
2) **architectures** (fully connected, convolutional)
3) **loss functions** (cross entropy, linear hinge)
4) **network sizes** (width, depth)
5) **minibatch sizes**

- In this talk: **fully connected + cross-entropy**

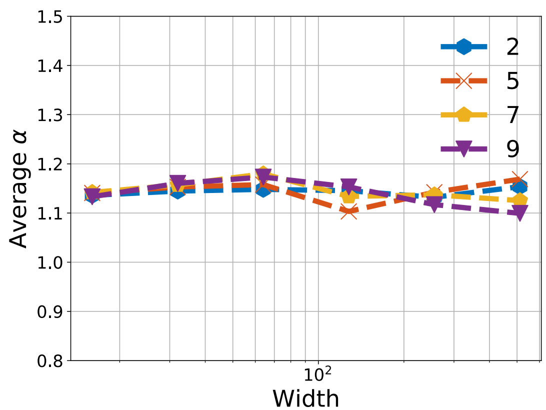
NETWORK SIZE

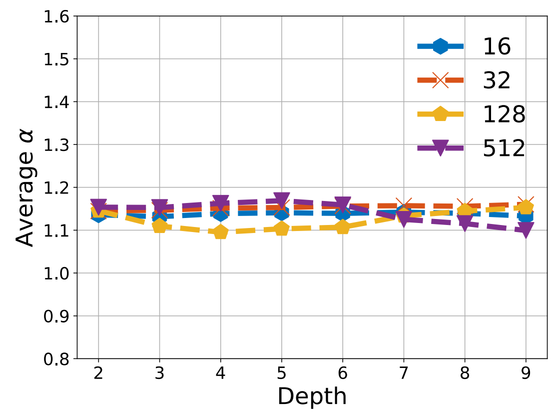
MNIST





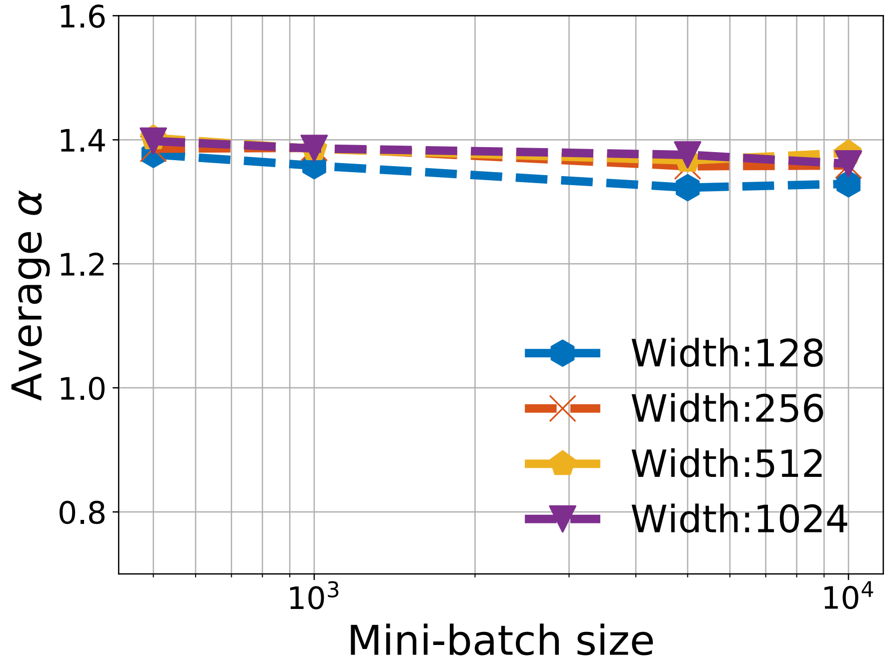
CIFAR10





MINIBATCH SIZE

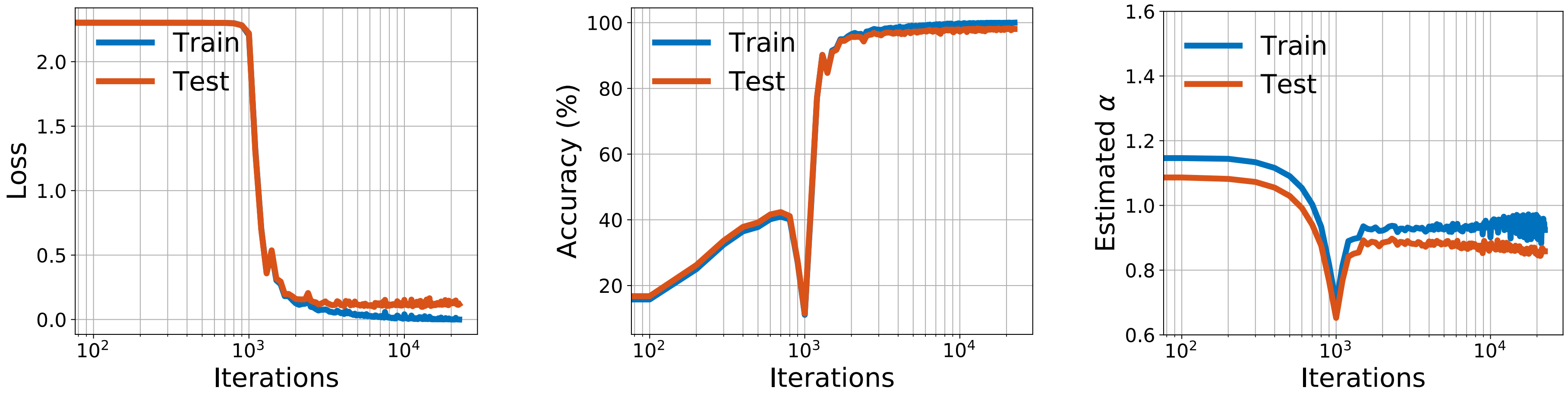
Depth: 2



- Similar results for other depths
- The behavior doesn't become Gaussian

CURIOUS JUMPS

MNIST + Fully Connected



CONCLUSIONS

- SGD noise is highly **non-Gaussian**
- **α-stable** assumption seems **more appropriate**
- Strong **interaction** between **geometry** & **dynamics**
- Existing **theory**: more light on the **wide minima** phenomenon
- Supports: SGD **crosses barriers** in the **initial phase**

THANK YOU FOR YOU ATTENTION!

Any questions?