Heavy-tail behavior of SGD, Part 2

Junghyun “Nick” Lee

(Undergraduate researcher, OSI Lab)

Overview

- Brief recap of last talk
- Heavy-tailedness **helps** generalization?
- **Origin** of heavy-tailed behavior?
- Can we **control** such heavy-tailed behavior?

Why does SGD work “well”?

- There are many perspectives of analyzing SGD:
  - Generalization capability
  - Algorithmic stability
  - Convergence rate
  - **Dynamics**
- By dynamics, we mean how the sequence of weight vectors $\{\boldsymbol{w}_t\}$ behaves.
  - Does it resemble a continuous process?
  - Does it go through a jump at some point?
  - …etc.

Analyzing the dynamics of SGD: Discretization of a Langevin SDE

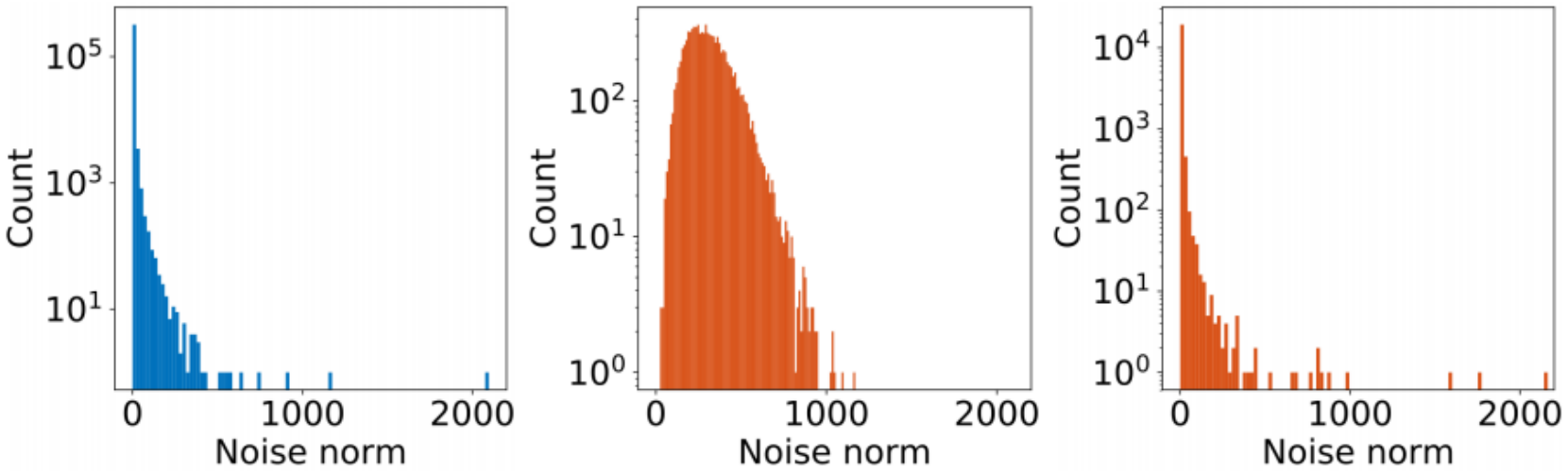
- Let us define the stochastic gradient noise:

$$U_t(\mathbf{w}) \triangleq \nabla \tilde{f}_t(\mathbf{w}) - \nabla f(\mathbf{w})$$

- And rewrite the SGD as follows:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \nabla f(\mathbf{w}_t) - \eta U_t(\mathbf{w}_t)$$

Empirical Evidence (Şimşekli et al., 2019)



(a) Real (b) Gaussian (c) α-stable

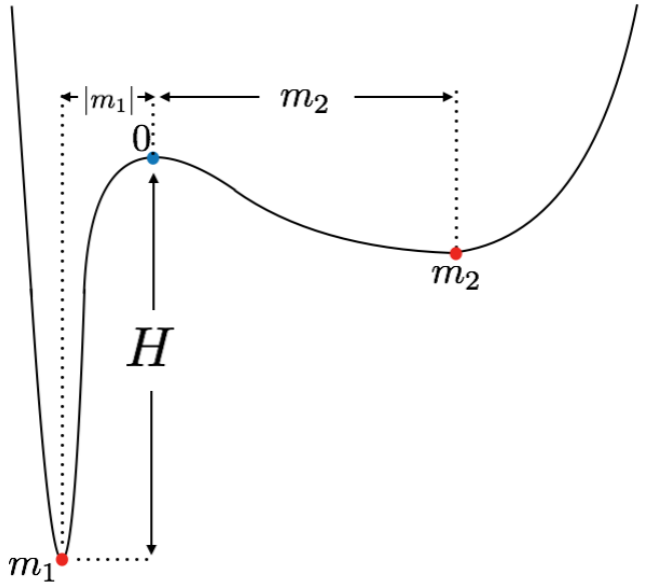
Figure 1. (a)The histogram of the norm of the gradient noises computed with AlexNet on Cifar10. (b) and (c) the histograms of the norms of (scaled) Guassian and α-stable random variables.

Theoretical Evidence 1: Dependency on the invariant measure

- Previous works depend on analyzing the invariant measure of the SDE
  - But how many iterations does it take for such convergence to take place?
  - $\mathrm{O}(\mathrm{Exp}(d))$ iterations (Raginsky et al., 2017)
  - In our deep learning settings, d is often huge, but we don't need that much iterations…!

Theoretical Evidence 2: Metastability

- Empirically (and maybe partially theoretically), it is known that SGD "prefers" to stay at a wide basin, resulting in a better generalization.
- BUT,
  - First exit time $\sim \mathrm{Exp}(height\ of\ basin)$ and $\sim \mathrm{Poly}(width\ of\ basin)$



The Main Fix

- We assume that SGN follows the d-dimensional stable α-distribution *with independent components*

$$[U_t(\mathbf{w})]_i \sim \mathcal{S}\alpha\mathcal{S}(\sigma), \quad \forall i \in \{1, \ldots, d\}$$

- This then induces a new SDE, which has a fundamentally different behavior than that of the original SDE.

NEW Continuous-time SGD
(Şimşekli, 2017)(Şimşekli et al., 2019)

Lemma (Informal). *The following (fractional) Langevin-type SDE is a weak approximation to the SGD (described previously):*

$$d\mathbf{w}_t = -\nabla f(\mathbf{w}_t)dt + \eta^{(\alpha-1)/\alpha}\sigma dL_t^\alpha$$

where L_t^α is the d-dimensional α-stable Lévy process with independent components.

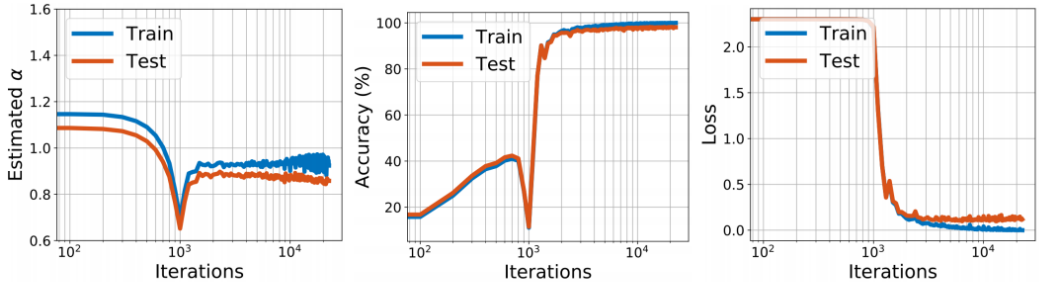
- Let us write above as the following:

$$d\mathbf{w}_t^\varepsilon = -\nabla f(\mathbf{w}_t^\varepsilon)dt + \varepsilon dL_t^\alpha$$

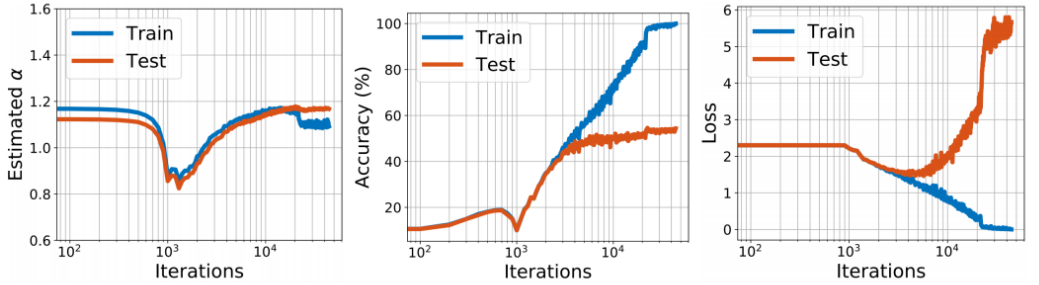
- ε controls the amount of perturbation on the dynamical system, caused by the Lévy process.

Experiments

- Out of all the results that the authors(Şimşekli et al., 2019) showed, let us focus on the most important one:



(a) MNIST



(b) CIFAR10

Figure 7. The iteration-wise behavior of of α for the FCN.

Main Question #1

Does heavy-tailedness have anything to do with generalization?

Short answer (from Prof. Şimşekli's twitter)

"SGD generates **fractals** and the generalization error is bounded by the **intrinsic dimension** of these fractals, which is determined by the **tail-behavior** at convergence"

Main contributions

- New notion of **complexity** for the trajectories of a stochastic learning algorithm, coined “uniform Hausdorff dimension”
  - sample paths of Feller processes admit a uniform Hausdorff dimension, which is closely related to the tail properties of the process

- Generalization error is controlled by the Hausdorff dimension of the process
  - It can be significantly smaller than the standard Euclidean dimension.
  - It acts as an ‘intrinsic dimension’ of the problem, mimicking the role of VC dimension in classical generalization bounds.

Feller process

- SGD is modeled by the following Feller process SDE:

$$\mathrm{d}\mathrm{W}_t = -\nabla f(\mathrm{W}_t)\mathrm{d}t + \Sigma_1(\mathrm{W}_t)\mathrm{d}\mathrm{B}_t + \Sigma_2(\mathrm{W}_t)\mathrm{d}\mathrm{L}_t^{\boldsymbol{\alpha}(\mathrm{W}_t)}$$

- All the previously considered SDEs (of a lot of stochastic algorithms) can be thought of as special cases of the above SDE
  - SGD
  - SGD with momentum
  - SGD under Gaussian/heavy-tailed noise

Why Hausdorff dimension?

- Many works (modern probability theory) have uncovered characteristics of Markov process
  - **Markov process induces a fractal-like structure.**
- Note that SGD (Feller process SDE in general) is Markov!
- We intend to make a connection between the induced fractal structure and the generalization capability
  - The intrinsic dimension of fractal is measured by Hausdorff dimension

What is Hausdorff dimension? (bit of topology…)(optional)

- Direct fractional generalization of usual definition of “dimension”

$$\mathcal{H}^s_\delta(G) := \inf \sum_{i=1}^{\infty} \operatorname{diam}(A_i)^s,$$

$$\mathcal{H}^s(G) := \lim_{\delta \to 0} \mathcal{H}^s_\delta(G).$$

Definition 1 *The Hausdorff dimension of* $G \subset \mathbb{R}^d$ *is defined as follows.*

$$\dim_{\mathrm{H}} G := \sup\{s > 0 : \mathcal{H}^s(G) > 0\} = \inf\{s > 0 : \mathcal{H}^s(G) < \infty\}. \tag{2.5}$$

Statistical learning framework

- Basic setting:
  - Space of data points: $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$
  - (unknown) data distribution: μ_z
  - (accessible) training dataset: $S = \{z_1, \ldots, z_n\}$
  - Loss function: $\ell : \mathbb{R}^d \times \mathcal{Z} \mapsto \mathbb{R}_+$
  - Population/empirical risk:

$$\mathcal{R}(w) := \mathbb{E}_z[\ell(w, z)] \qquad \hat{\mathcal{R}}(w, S) := \frac{1}{n} \sum_{i=1}^{n} \ell(w, z_i).$$

Statistical learning framework

- Iterative training algorithm: $\mathcal{A} = \mathcal{A}(S, U)$
  - U encapsulates all the randomness, whose law is given as μ_u
  - Returns the entire evolution of the parameters in $[0,1]$ (WLOG)
- Here, we consider the case when $\mathcal{A}$ is chosen to be the Feller process
- **Assume** that S and U are independent

$$\mathcal{W}_S := \{w \in \mathbb{R}^d : \exists t \in [0,1], w = [\mathcal{A}(S)]_t\}$$

$$\mathcal{W} = \bigcup_{S \in \mathcal{Z}^n} \mathcal{W}_S.$$

Main definition + result

Definition 2 *An algorithm $\mathcal{A}$ has uniform Hausdorff dimension d_{H} if for any training set $S \in \mathcal{Z}^n$*

$$\dim_{\mathrm{H}} \mathcal{W}_S = \dim_{\mathrm{H}}\{w \in \mathbb{R}^d : \exists t \in [0,1], w = [\mathcal{A}(S,U)]_t\} \leq d_{\mathrm{H}}, \quad \mu_u\text{-almost surely.} \tag{3.5}$$

- Note that any algorithm $\mathcal{A}$ posseses above property with $d_H = d$
- However, $\boldsymbol{d_H}$ **can be (and usually) much smaller than** $\boldsymbol{d}$**:**

Proposition 1 *Let $\{\mathrm{W}^{(S)}\}_{S\in\mathcal{Z}^n}$ be a family of Feller processes. Assume that for each S, $\mathrm{W}^{(S)}$ is decomposable at a point w_S with sub-symbol ψ_S. Consider the algorithm $\mathcal{A}$ that returns $[\mathcal{A}(S)]_t = \mathrm{W}_t^{(S)}$ for a given $S \in \mathcal{Z}^n$ and for every $t \in [0,1]$. Then, we have*

$$\dim_{\mathrm{H}} \mathcal{W}_S \leq \beta_S, \quad \textit{where} \quad \beta_S := \inf\left\{\lambda \geq 0 : \lim_{\|\xi\|\to\infty} \frac{|\psi_S(\xi)|}{\|\xi\|^\lambda} = 0\right\}, \tag{3.6}$$

μ_u-almost surely. Furthermore, $\mathcal{A}$ has uniform Hausdorff dimension with $d_{\mathrm{H}} = \sup_{S\in\mathcal{Z}^n} \beta_S$.

Main definition + result

- β_S is often termed as the **upper Blumenthal-Getoor (BG) index** of the Lévy process with an exponent ψ_S

- In general, β_S decreases as the process gets heavier-tailed i.e. **heavier-tailed processes have smaller Hausdorff dimension**
  - Smaller complexity

Algorithm-dependent definition of generalization

$$\sup_{t\in[0,1]} |\hat{\mathcal{R}}([\mathcal{A}(S)]_t, S) - \mathcal{R}([\mathcal{A}(S)]_t)| = \sup_{w\in\mathcal{W}_S} |\hat{\mathcal{R}}(w, S) - \mathcal{R}(w)|,$$

Main results

Theorem 1 *Assume that* **H***1 to 4 hold, and* $\mathcal{Z}$ *is countable. Then, for a sufficiently large* n*, we have*

$$\sup_{w \in \mathcal{W}_S} |\hat{\mathcal{R}}(w, S) - \mathcal{R}(w)| \leq B \sqrt{\frac{2 d_{\mathrm{H}} \log(nL^2)}{n} + \frac{\log(1/\gamma)}{n}}, \tag{3.8}$$

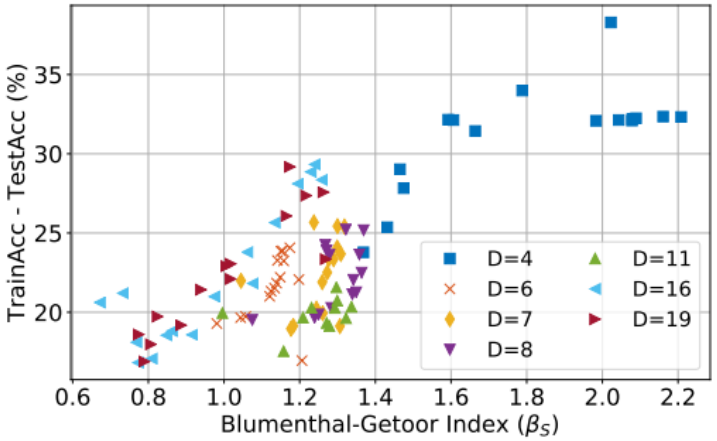
with probability at least $1 - \gamma$ *over* $S \sim \mu_z^{\otimes n}$ *and* $U \sim \mu_u$.

Theorem 2 *Assume that* **H***1, 2 and 5 hold, and* **H***4 holds with* $\mathcal{W}_S$ *in place of* $\mathcal{W}$ *for all* $S \in \mathcal{Z}^n$ *(with* a*,* b*,* r_0*,* s *potentially depending on* S*). Then, for* n *sufficiently large, we have*

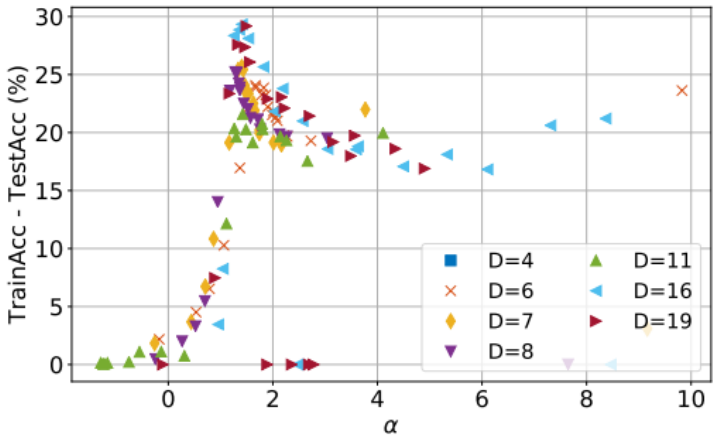
$$\sup_{w \in \mathcal{W}_S} |\hat{\mathcal{R}}(w, S) - \mathcal{R}(w)| \leq B \sqrt{\frac{2 \dim_{\mathrm{H}} \mathcal{W}_S \log(nL^2)}{n} + \frac{\log(4M/\gamma)}{n}}, \tag{3.11}$$

with probability at least $1 - \gamma$ *over* $S \sim \mu_z^{\otimes n}$ *and* $U \sim \mu_u$.

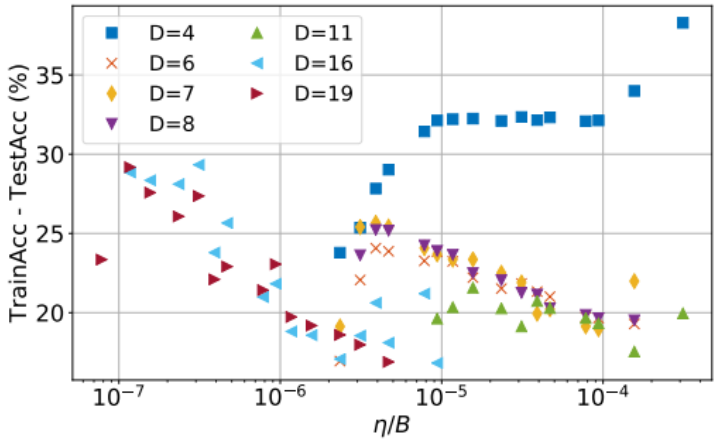
Experiments



(a) Hausdorff Dimension [Ours]



(c) α [MM19]



(b) η/B [JKA+17]

Figure 2: Empirical comparison to other capacity metrics.

Experiments

- **No relationship between the number of parameters and $\boldsymbol{dim_H W_S}$**
  - Increasing depth is **not** always beneficial from the generalization perspective
  - Choices of learning rate and bath size effects β_S
- β_S is the best choice for comparing the generalization capability of architecture+algorithm pair.

Main Question #2

What is the origin of such heavy-tailed phenomenon?

Main contributions

- Even in the simplest case of linear regression over Gaussian data, SGD iterates **can lead to a heavy-tailed stationary distribution** with an infinite variance:
  - The tails become *monotonically heavier* for increasing curvature, increasing η, or decreasing b
  - The law of the iterates converges exponentially fast towards the stationary distribution in the Wasserstein metric
  - There exists a higher-order moment (e.g., variance) of the iterates that diverges *at most* polynomially-fast, depending on the heaviness of the tails at stationarity.

- Empirically, this is shown to hold for synthetic settings, and deep neural networks.

(Informal) Motivation

- Assume:
  - x_0 is in the domain of attraction of a local minimum x_* of f
  - f is smooth and **well-approximated by a quadratic function in this basin**
- Then using Taylor expansion, SGD can be approximated as follows:

$$x_k \approx x_{k-1} - (\eta/b) \sum_{i \in \Omega_k} \nabla^2 f^{(i)}(x_\star) x_{k-1} + (\eta/b) \sum_{i \in \Omega_k} \left(\nabla^2 f^{(i)}(x_\star) x_\star - \nabla f^{(i)}(x_\star) \right)$$
$$=: \left(I - (\eta/b) H_k\right) x_{k-1} + q_k, \tag{3.1}$$

- This is a **linear stochastic recursion (LSR)!**

(Informal) Motivation

- Consider the case of f being quadratic i.e.

$$\min_{x \in \mathbb{R}^d} F(x) := (1/2)\mathbb{E}_{(a,y)\sim\mathcal{D}}\left[\left(a^T x - y\right)^2\right]$$

- We have access to i.i.d samples (a_i, y_i) from some distribution $\mathcal{D}$ whose support is in $\mathbb{R}^d \times \mathbb{R}$
- Assume the following:

(A1) $a_i \sim \mathcal{N}(0, \sigma^2 I_d)$ are i.i.d.

(A2) y_i are i.i.d. with a continuous density whose support is $\mathbb{R}$ with all the moments finite.

(Informal) Motivation

- Under such settings, our SGD becomes

$$x_k = M_k x_{k-1} + q_k \text{ with } M_k := I - (\eta/b)H_k,\ H_k := \sum\nolimits_{i\in\Omega_k} a_i a_i^T,\ q_k := (\eta/b)\sum\nolimits_{i\in\Omega_k} a_i y_i\ ;$$

  - Assume one-pass regime, or streaming setting, i.e. datas are only used once.

Preliminaries for LSR

- Quantities useful for analyzing LSR:

$$h(s) := \lim_{k\to\infty} \left(\mathbb{E}\|M_k M_{k-1} \dots M_1\|^s\right)^{1/k}$$

$$\Pi_k := M_k M_{k-1} \dots M_1 \quad \text{and} \quad \rho := \lim_{k\to\infty} (2k)^{-1} \log\left(\text{largest eigenvalue of } \Pi_k^T \Pi_k\right)$$

- In our case, above reduces to the following:

$$\rho = \mathbb{E} \log \|(I - (\eta/b)H)\, e_1\|, \quad h(s) = \mathbb{E}\left[\|(I - (\eta/b)H)\, e_1\|^s\right] \text{ for } \rho < 0,$$

Main Theorem

- Even in the simplest Gaussian setting, SGD iterates can leaad to a heavy-tailed stationary distribution:

Theorem 1 *Consider the SGD iterations (3). If* $\rho < 0$*, then SGD iterations admit a unique stationary distribution* x_∞ *which satisfy*

$$\lim_{t\to\infty} t^{\alpha}\mathbb{P}\left(u^T x_\infty > t\right) = e_\alpha(u)\,, \quad u \in \mathbb{S}^{d-1}\,, \tag{3.6}$$

for some positive and continuous function e_α *on the unit sphere* $\mathbb{S}^{d-1}$*, where* α *is the unique positive value such that* $h(\alpha) = 1$.

Main question

How the tail-index depends on the parameters of the problem including the batch size, dimension and the learning rate?

Main results : tail-index and hyperparameters

- These results theoretically confirm that there is a correlation between the generalization capability, $\frac{\eta}{b}$, and curvature around the minimum:

Theorem 2 *The tail-index α is strictly increasing in batch size b and strictly decreasing in stepsize η and variance σ^2 provided that $\alpha \geq 1$. Moreover, the tail-index α is strictly decreasing in dimension d.*

Proposition 3 *Let $\eta_{crit} = \frac{2b}{\sigma^2(d+b+1)}$. The following holds: (i) There exists $\eta_{max} > \eta_{crit}$ such that for any $\eta_{crit} < \eta < \eta_{max}$, Theorem 1 holds with tail index $0 < \alpha < 2$. (ii) If $\eta = \eta_{crit}$, Theorem 1 holds with tail index $\alpha = 2$. (iii) If $\eta \in (0, \eta_{crit})$, then Theorem 1 holds with tail index $\alpha > 2$.*

Three regimes for learning rate

- (I) $\rho < 0, \alpha > 2$: convergence to a limit with finite variance
- (II) $\rho < 0, \alpha < 2$: convergence to a heavy-tailed limit
- (III) $\rho > 0$: convergence not guaranteed

i.e.

- $\eta < \eta_{crit}$ => (I)
- $\eta_{crit} < \eta < \eta_{max}$ => (II)
- $\eta_{max} < \eta$ => (III)

Linear convergence in p-Wasserstein distance

Theorem 6 *Assume* $\alpha > 1$. *Let* ν_k, ν_∞ *denote the probability laws of* x_k *and* x_∞ *respectively. Then* $\mathcal{W}_p(\nu_k, \nu_\infty) \leq (h(p))^{k/p}\mathcal{W}_p(\nu_0, \nu_\infty)$, *for any* $1 \leq p < \alpha$, *where the convergence rate* $(h(p))^{1/p} \in (0, 1)$.

At most polynomial divergence of α-moment

Proposition 8 *Given the tail-index α, we have $\mathbb{E}\|x_\infty\|^\alpha = \infty$. Moreover, $\mathbb{E}\|x_k\|^\alpha = O(k)$ if $\alpha \leq 1$, and $\mathbb{E}\|x_k\|^\alpha = O(k^\alpha)$ if $\alpha > 1$.*

GCLT for ergodic averages

Corollary 5 *(i) If the tail-index* $\alpha \leq 1$*, then for any* $p \in (0, \alpha)$*,* $\mathbb{E}\|x_\infty\|^p \leq \frac{1}{1-h(p)}\mathbb{E}\|q_1\|^p$*, where* $h(p) < 1$*. (ii) If the tail-index* $\alpha > 1$*, then for any* $p \in (1, \alpha)$*, we have* $h(p) < 1$ *and for any* $\epsilon > 0$ *such that* $(1+\epsilon)h(p) < 1$*, we have* $\mathbb{E}\|x_\infty\|^p \leq \frac{1}{1-(1+\epsilon)h(p)} \frac{(1+\epsilon)^{\frac{p}{p-1}} - (1+\epsilon)}{((1+\epsilon)^{\frac{1}{p-1}} - 1)^p}\mathbb{E}\|q_1\|^p$.

Corollary 9 *Assume* $\rho < 0$ *so that Theorem 1 holds. Then, we have the following:*

(i) If $\alpha \in (0,1) \cup (1,2)$*, then there is a sequence* $d_K = d_K(\alpha)$ *and a function* $C_\alpha : \mathbb{S}^{d-1} \mapsto \mathbb{C}$ *such that as* $K \to \infty$ *the random variables* $K^{-\frac{1}{\alpha}}(S_K - d_K)$ *converge in law to the* α*-stable random variable with characteristic function* $\Upsilon_\alpha(tv) = \exp(t^\alpha C_\alpha(v))$*, for* $t > 0$ *and* $v \in \mathbb{S}^{d-1}$*.*

(ii) If $\alpha = 1$*, then there are functions* $\xi, \tau : (0, \infty) \mapsto \mathbb{R}$ *and* $C_1 : \mathbb{S}^{d-1} \mapsto \mathbb{C}$ *such that as* $K \to \infty$ *the random variables* $K^{-1}S_K - K\xi\left(K^{-1}\right)$ *converge in law to the random variable with characteristic function* $\Upsilon_1(tv) = \exp\left(tC_1(v) + it\langle v, \tau(t)\rangle\right)$*, for* $t > 0$ *and* $v \in \mathbb{S}^{d-1}$*.*

(iii) If $\alpha = 2$*, then there is a sequence* $d_K = d_K(2)$ *and a function* $C_2 : \mathbb{S}^{d-1} \mapsto \mathbb{R}$ *such that as* $K \to \infty$ *the random variables* $(K \log K)^{-\frac{1}{2}}(S_K - d_K)$ *converge in law to the random variable with characteristic function* $\Upsilon_2(tv) = \exp\left(t^2 C_2(v)\right)$*, for* $t > 0$ *and* $v \in \mathbb{S}^{d-1}$*.*

(iv) If $\alpha \in (0,1)$*, then* $d_K = 0$*, and if* $\alpha \in (1,2]$*, then* $d_K = K\bar{x}$*, where* $\bar{x} = \int_{\mathbb{R}^d} x\nu_\infty(dx)$*.*

GCLT for ergodic averages

- There is a **practical implication!**

Instead of trying to estimate the tail-index of some generic heavy-tailed distribution, we can focus on α**-stable distributions**, which has a very efficient and easy estimator for us to use…!

So what causes the heavy tails??

- There are two types of gradient noise:
  - Multiplicative noise (M_k)
  - Additive noise (q_k)

- **Multiplicative noise** is the main source for the heavy tails
  - It was shown that deterministic M_k would not lead to heavy tails.

Experiments

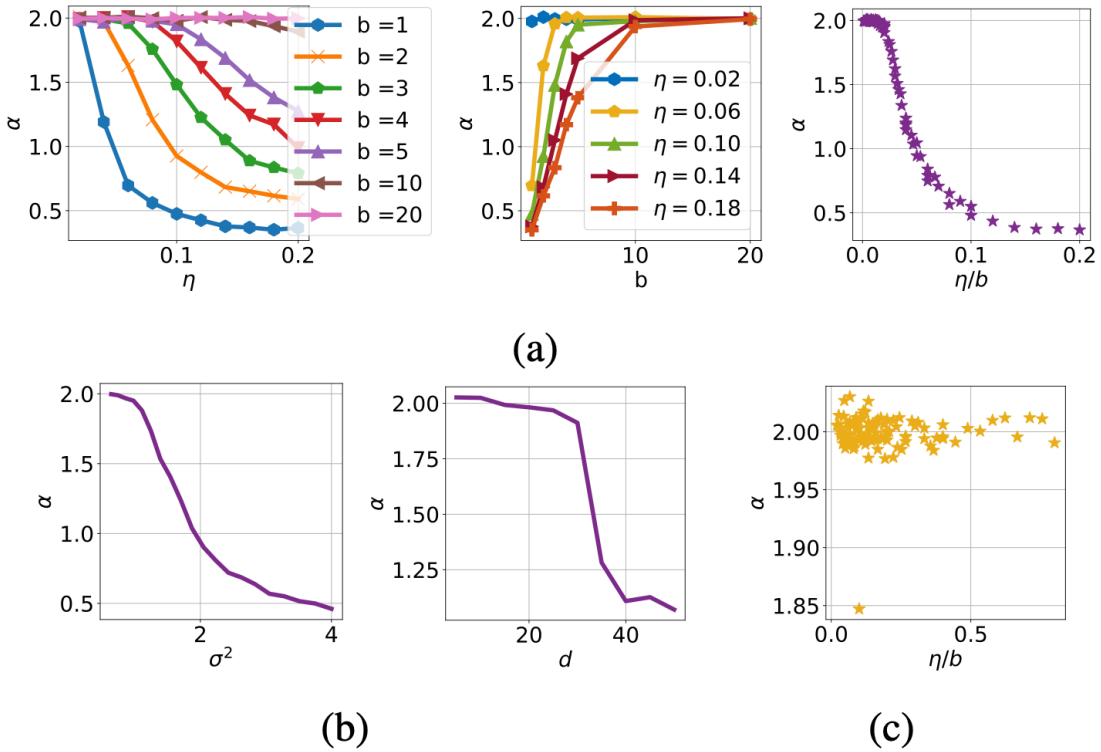
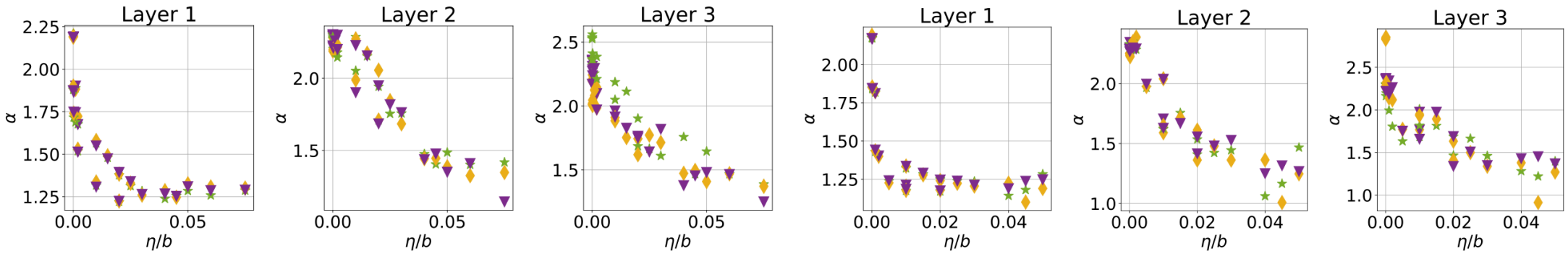


Figure 1: Behavior of α with (a) varying stepsize η and batch-size b, (b) d and σ, (c) under RMSProp.

Experiments



(a) MNIST (b) CIFAR10

Figure 2: Results on FCNs. Different markers represent different initializations with the same η, b.

Summary of everything so far…!

- **Multiplicative noise** is shown to be the main cause for the heavy tails of SGD iterates.
- Changing η or b doesn't only change the scale of the noise; it also changes the tails of the algorithm
- **Heavy-tailed SGD iterates imply less generalization error**
  - Heavy-tailed process have smaller Hausdorff dimension.
  - Hausdorff dimension acts as VC dimension for the generalization error i.e. it is included in the upper bound.
- Theoretically motivated β_S(upper BG index) is shown to have strong correlation with the generalization capability.

Main References

- **Main"** Gürbüzbalaban, M., Şimşekli, U., Zhu, L. The Heavy-Tail Phenomenon in SGD. In *arXiv*, 2020.
- **Sub-main:** Şimşekli, U., Sener, O., Deligiannidis, G., Erdogdu, M. A. Hausdorff Dimension, Stochastic Differential Equations, and Generalization in Neural Networks. In *NIPS*, 2020. (*Spotlight paper!*)
- Şimşekli, U., Gürbüzbalaban, M., Nguyen, T. H., Richard, G., Sagun, L. On the Heavy-Tailed Theory of Stochastic Gradient Descent for Deep Neural Networks. In *arXiv*, 2019.

Thank you!

Any questions?