# The Heavy-Tail Phenomenon in Decentralized Stochastic Gradient Descent

Mohammad Rafiqul Islam

Florida State University

November 20, 2023

# Table of Contents

# Optimization Problem

The learning or training of the neural network involves a very well-known optimization problem:

$$\min_{x \in \mathbb{R}^d} F(x) :\triangleq \mathbb{E}_{z \sim \mathcal{D}} [f(x, z)] \tag{1}$$

where

- $\mathbf{z} \in \mathbb{R}^p$ denotes a random data point,

# Optimization Problem

The learning or training of the neural network involves a very well-known optimization problem:

$$\min_{x \in \mathbb{R}^d} F(x) :\triangleq \mathbb{E}_{z \sim \mathcal{D}} \left[ f(x, z) \right] \tag{1}$$

where

- $\mathbf{z} \in \mathbb{R}^p$ denotes a random data point,
- $\mathcal{D}$ is a probability distribution on $\mathbb{R}^p$ that denotes the law of the data points,

# Optimization Problem

The learning or training of the neural network involves a very well-known optimization problem:

$$\min_{x \in \mathbb{R}^d} F(x) :\triangleq \mathbb{E}_{z \sim \mathcal{D}} [f(x, z)] \tag{1}$$

where

- $\mathbf{z} \in \mathbb{R}^p$ denotes a random data point,
- $\mathcal{D}$ is a probability distribution on $\mathbb{R}^p$ that denotes the law of the data points,
- $x \in \mathbb{R}^d$ denotes the parameters of the neural network to be optimized,

# Optimization Problem

The learning or training of the neural network involves a very well-known optimization problem:

$$\min_{x \in \mathbb{R}^d} F(x) :\triangleq \mathbb{E}_{z \sim \mathcal{D}} [f(x, z)] \tag{1}$$

where

- $\mathbf{z} \in \mathbb{R}^p$ denotes a random data point,
- $\mathcal{D}$ is a probability distribution on $\mathbb{R}^p$ that denotes the law of the data points,
- $x \in \mathbb{R}^d$ denotes the parameters of the neural network to be optimized,
- $f : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}_+$ denotes a measurable cost function, which maybe convex or non-convex in $x$.

# Stochastic Gradient Descent (SGD) Method

If we have a training dataset, $S = \{z_1, z_2, \cdots, z_n\}$ with $n$ i.i.d observations, i.e., $z_i \sim_{i.i.d} \mathcal{D}$ for $i = 1, 2, \cdots, n$

# Stochastic Gradient Descent (SGD) Method

If we have a training dataset, $S = \{z_1, z_2, \cdots, z_n\}$ with $n$ i.i.d observations, i.e., $z_i \sim_{i.i.d} \mathcal{D}$ for $i = 1, 2, \cdots, n$

- Empirical risk minimization problem:

$$\min_{x \in \mathbb{R}^d} f(x, S) :\triangleq \frac{1}{n} \sum_{i=1}^{n} f^{(i)}(x) \tag{2}$$

# Stochastic Gradient Descent (SGD) Method

If we have a training dataset, $S = \{z_1, z_2, \cdots, z_n\}$ with $n$ i.i.d observations, i.e., $z_i \sim_{i.i.d} \mathcal{D}$ for $i = 1, 2, \cdots, n$

- Empirical risk minimization problem:

$$\min_{x \in \mathbb{R}^d} f(x, S) :\triangleq \frac{1}{n} \sum_{i=1}^{n} f^{(i)}(x) \tag{2}$$

- $f^{(i)}$ denotes the cost or (instantaneous) loss function that is contributed by the data point $z_i$ for $i \in \{1, 2, 3, \cdots, n\}$.

# Stochastic Gradient Descent (SGD) Method

If we have a training dataset, $S = \{z_1, z_2, \cdots, z_n\}$ with $n$ i.i.d observations, i.e., $z_i \sim_{i.i.d} \mathcal{D}$ for $i = 1, 2, \cdots, n$

- Empirical risk minimization problem:

$$\min_{x \in \mathbb{R}^d} f(x, S) :\triangleq \frac{1}{n} \sum_{i=1}^{n} f^{(i)}(x) \tag{2}$$

- $f^{(i)}$ denotes the cost or (instantaneous) loss function that is contributed by the data point $z_i$ for $i \in \{1, 2, 3, \cdots, n\}$.
- The SGD iteration: $x_k = x_{k-1} - \eta \nabla \tilde{f}_k(x_{k-1})$

# Stochastic Gradient Descent (SGD) Method

If we have a training dataset, $S = \{z_1, z_2, \cdots, z_n\}$ with $n$ i.i.d observations, i.e., $z_i \sim_{i.i.d} \mathcal{D}$ for $i = 1, 2, \cdots, n$

- Empirical risk minimization problem:

$$\min_{x \in \mathbb{R}^d} f(x, S) :\triangleq \frac{1}{n} \sum_{i=1}^{n} f^{(i)}(x) \tag{2}$$

- $f^{(i)}$ denotes the cost or (instantaneous) loss function that is contributed by the data point $z_i$ for $i \in \{1, 2, 3, \cdots, n\}$.
- The SGD iteration: $x_k = x_{k-1} - \eta \nabla \tilde{f}_k(x_{k-1})$
- $\nabla \tilde{f}_k$ denotes the stochastic gradient at iteration $k$, which is given as

$$\nabla \tilde{f}_k(x) \triangleq \nabla \tilde{f}_{\Omega_k}(x) \triangleq \frac{1}{b} \sum_{i \in \Omega_k} \nabla f^{(i)}(x) \tag{3}$$

# Stochastic Gradient Descent (SGD) Method

If we have a training dataset, $S = \{z_1, z_2, \cdots, z_n\}$ with $n$ i.i.d observations, i.e., $z_i \sim_{i.i.d} \mathcal{D}$ for $i = 1, 2, \cdots, n$

- Empirical risk minimization problem:

$$\min_{x \in \mathbb{R}^d} f(x, S) :\triangleq \frac{1}{n} \sum_{i=1}^{n} f^{(i)}(x) \tag{2}$$

- $f^{(i)}$ denotes the cost or (instantaneous) loss function that is contributed by the data point $z_i$ for $i \in \{1, 2, 3, \cdots, n\}$.
- The SGD iteration: $x_k = x_{k-1} - \eta \nabla \tilde{f}_k(x_{k-1})$
- $\nabla \tilde{f}_k$ denotes the stochastic gradient at iteration $k$, which is given as

$$\nabla \tilde{f}_k(x) \triangleq \nabla \tilde{f}_{\Omega_k}(x) \triangleq \frac{1}{b} \sum_{i \in \Omega_k} \nabla f^{(i)}(x) \tag{3}$$

- Stochasticity: $\Omega_k \subset \{1, 2, 3, \cdots, n\}$ and $b = |\Omega_k|$

# Assumptions

- Stochastic gradient noise (GN): $U_k(\mathbf{x}) \triangleq \nabla \tilde{f}_k(\mathbf{x}) - \nabla f_k(\mathbf{x})$

# Assumptions

- Stochastic gradient noise (GN): $U_k(\mathbf{x}) \triangleq \nabla \tilde{f}_k(\mathbf{x}) - \nabla f_k(\mathbf{x})$
- If $b$ is large enough, then by Central Limit Theorem

$$U_k(\mathbf{x}) \sim \mathcal{N}(0, \sigma^2 \mathbf{I}) \tag{4}$$

# Assumptions

- Stochastic gradient noise (GN): $U_k(\mathbf{x}) \triangleq \nabla \tilde{f}_k(\mathbf{x}) - \nabla f_k(\mathbf{x})$
- If $b$ is large enough, then by Central Limit Theorem

$$U_k(\mathbf{x}) \sim \mathcal{N}(0, \sigma^2 \mathbf{I}) \tag{4}$$

- Under this assumption, the SGD can be written as follows:

$$\mathbf{x}_k = \mathbf{x}_{k-1} - \eta \nabla f(\mathbf{x}_{k-1}) + \sqrt{\eta}\sqrt{\eta\sigma^2} Z_{k-1} \tag{5}$$

where $\mathbf{Z}_k$ denotes a standard normal random variable in $\mathbb{R}^d$.

# Assumptions

- Stochastic gradient noise (GN): $U_k(\mathbf{x}) \triangleq \nabla \tilde{f}_k(\mathbf{x}) - \nabla f_k(\mathbf{x})$
- If $b$ is large enough, then by Central Limit Theorem

$$U_k(\mathbf{x}) \sim \mathcal{N}(0, \sigma^2 \mathbf{I}) \tag{4}$$

- Under this assumption, the SGD can be written as follows:

$$\mathbf{x}_k = \mathbf{x}_{k-1} - \eta \nabla f(\mathbf{x}_{k-1}) + \sqrt{\eta}\sqrt{\eta\sigma^2} Z_{k-1} \tag{5}$$

  where $\mathbf{Z}_k$ denotes a standard normal random variable in $\mathbb{R}^d$.
- If $\eta$ is small enough then the continuous version of (5) is the following stochastic differential equation (SDE)

$$d\mathbf{x}_t = -\nabla f(\mathbf{x}_t) dt + \sqrt{\eta\sigma^2} d\mathbf{B}_t \tag{6}$$

  where $\mathbf{B}_t$ denotes the standard Brownian motion.

# Discussion

- Equation (6) is known as the Langevin diffusion

# Discussion

- Equation (6) is known as the Langevin diffusion
- Under mild regularity assumptions on $f$, one can show that the Markov process $\{\mathbf{x}_t\}_{t\geq 0}$ is ergodic with its unique invariant measure, whose density is proportional to

$$e^{-\frac{f(x)}{\eta\sigma^2}}$$

# Discussion

- Equation (6) is known as the Langevin diffusion
- Under mild regularity assumptions on $f$, one can show that the Markov process $\{\mathbf{x}_t\}_{t\geq 0}$ is ergodic with its unique invariant measure, whose density is proportional to

$$e^{-\frac{f(x)}{\eta\sigma^2}}$$

- Based on this observation, Jastrzębski et al.[1] focused on the relation between this invariant measure and the algorithm parameters, $\eta$ and $b$ as a function of $\sigma^2$.

[1] S. Jastrzębski, Z. Kenton, D. Arpit, N. Ballas, A. Fischer, Y. Bengio, and A. Storkey. Three factors influencing minima in SGD. arXiv preprint arXiv:1711.04623, 2017.

# Discussion

- Equation (6) is known as the Langevin diffusion
- Under mild regularity assumptions on $f$, one can show that the Markov process $\{\mathbf{x}_t\}_{t\geq 0}$ is ergodic with its unique invariant measure, whose density is proportional to

$$e^{-\frac{f(x)}{\eta\sigma^2}}$$

- Based on this observation, Jastrzębski et al.[1] focused on the relation between this invariant measure and the algorithm parameters, $\eta$ and $b$ as a function of $\sigma^2$.
- Their conclusion: ratio $\eta/b$ is the control parameter that determines the width of the minima found by SGD

[1] S. Jastrzębski, Z. Kenton, D. Arpit, N. Ballas, A. Fischer, Y. Bengio, and A. Storkey. Three factors influencing minima in SGD. arXiv preprint arXiv:1711.04623, 2017.

# Wide Minima Folklore

- They revisited the famous wide-minima folklore[2]

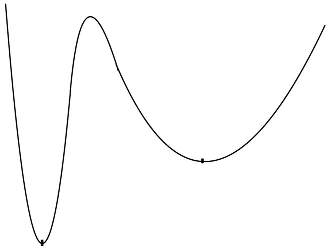
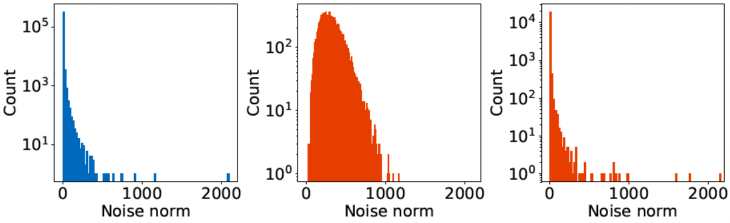


Figure 1: Hypothetical Loss function

- "Among the minima found by SGD, the wider it is, the better it performs on the test set"

[2] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. Neural computation, 9(1):1–42, 1997.

# Empirical issues: Gaussianity assumption




(a) Real (b) Gaussian (c) $\alpha$-stable

Figure 2: (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 $\alpha$-stable random variables.

[2] AlexNet is a convolutional neural network that is 8 layers deep

# Theoretical issues: Complexity and wide minima

- Large number of iterations required to converge to an invariant measure[a]: No. of iterations $\approx \mathcal{O}(e^d)$

# Theoretical issues: Complexity and wide minima

- Large number of iterations required to converge to an invariant measure[a]:
  No. of iterations $\approx \mathcal{O}(e^d)$
- Transition time $\approx e^H \times poly(|m_1|)$

# Theoretical issues: Complexity and wide minima

- Large number of iterations required to converge to an invariant measure[a]: No. of iterations $\approx \mathcal{O}(e^d)$
- Transition time $\approx e^H \times poly(|m_1|)$
- Therefore, SGD prefers wide minima within a considerably small number of iterations cannot be explained using the asymptotic distribution of the SDE given in (6).

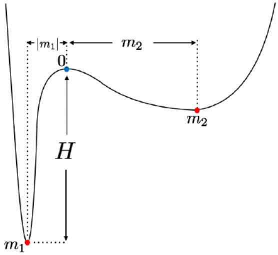



Figure 3: An objective with two local minima $m_1, m_2$ separated by a local maxima at $s_1 = 0$

[a] P. Xu, J. Chen, D. Zou, and Q. Gu. Global convergence of Langevin dynamics based algorithms for nonconvex optimization. Advances in Neural Information Processing Systems, 31, 2018.

# Lévy-Driven SDE Assumptions

- If the Gaussian assumption is not adequate, by generalized CLT, one can model stochastic gradient noise by:

$$[U_k(\mathbf{x})]_i \sim \mathcal{S}\alpha\mathcal{S}(\sigma(\mathbf{x})),\ \forall\ i = 1, 2, \cdots, n \tag{7}$$

where $[v]_i$ denotes the $i$th component of a vector $v$

# Lévy-Driven SDE Assumptions

- If the Gaussian assumption is not adequate, by generalized CLT, one can model stochastic gradient noise by:

$$[U_k(\mathbf{x})]_i \sim \mathcal{S}\alpha\mathcal{S}(\sigma(\mathbf{x})), \ \forall\ i = 1, 2, \cdots, n \tag{7}$$

  where $[v]_i$ denotes the $i$th component of a vector $v$

- Based on the assumption above, (7), we can rewrite the SGD recursion as follows:

$$\mathbf{x}_k = \mathbf{x}_{k-1} - \eta \nabla f(\mathbf{x}_{k-1}) + \eta^{\frac{1}{\sigma}} \left( \eta^{\frac{\alpha-1}{\alpha}} \sigma \right) S_{k-1} \tag{8}$$

  where $\mathbf{S}_k \in \mathbb{R}^d$ is a random vector such that $[S_k]_i \sim \mathcal{S}\alpha\mathcal{S}(1)$.

# Lévy-Driven SDE Assumptions

- If the Gaussian assumption is not adequate, by generalized CLT, one can model stochastic gradient noise by:

$$[U_k(\mathbf{x})]_i \sim \mathcal{S}\alpha\mathcal{S}(\sigma(\mathbf{x})),\ \forall\ i = 1, 2, \cdots, n \tag{7}$$

where $[v]_i$ denotes the $i$th component of a vector $v$

- Based on the assumption above, (7), we can rewrite the SGD recursion as follows:

$$\mathbf{x}_k = \mathbf{x}_{k-1} - \eta\nabla f(\mathbf{x}_{k-1}) + \eta^{\frac{1}{\sigma}} \left(\eta^{\frac{\alpha-1}{\alpha}}\sigma\right) S_{k-1} \tag{8}$$

where $\mathbf{S}_k \in \mathbb{R}^d$ is a random vector such that $[S_k]_i \sim \mathcal{S}\alpha\mathcal{S}(1)$.

- If $\eta$ is small enough then the continuous-time limit of this eq. (8) is the following SDE driven by an $\alpha$-stable Lévy process:

$$d\mathbf{x}_t = -\nabla f(\mathbf{x}_t)dt + \eta^{\frac{\alpha-1}{\alpha}}\sigma d\mathbf{L}_t^{\alpha} \tag{9}$$

where $\mathbf{L}_t^{\alpha}$ denotes the $d$-dimensional $\alpha$-stable Lévy motion

# $\alpha$-stable distribution

- $X \sim \mathcal{S}\alpha\mathcal{S}(\sigma)$ if its characteristic function is

$$\mathbb{E}\left[e^{iuX}\right] = e^{-\sigma^\alpha |u|^\alpha} \text{ for } u \in \mathbb{R}$$

- $d-$dimensional version:

$$\mathbb{E}\left[e^{i\langle u, X\rangle}\right] = e^{-\sigma^\alpha \|u\|_2^\alpha} \text{ for } u \in \mathbb{R}^d$$

- $\sigma > 0$ : scale parameter measures the spread of $X$ around 0 & $\alpha \in (0, 2]$ : determines the heaviness of the distribution tails.

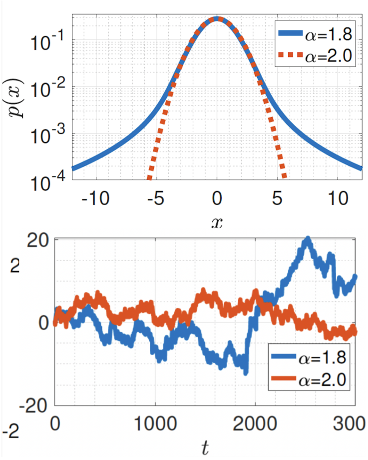



Figure 4: $\alpha$-stable Distribution

# Lévy-driven SDE approximation of SGD (cont.)

- For simplicity of the presentation we rewrite equation (9) for $d = 1$ case (Multidimensional case[3])

$$d\mathbf{x}_t^\epsilon = -\nabla f(\mathbf{x}_t^\epsilon)dt + \epsilon d\mathbf{L}_t^\alpha \tag{10}$$

for $t \geq 0$, started from the initial point $\mathbf{x}_0 \in \mathbb{R}$, $\epsilon \geq 0$ is a parameter and $f$ is a non-convex objective with $r \geq 2$ local minima.

[3] P. Imkeller, I. Pavlyukevich, and M. Stauch. First exit times of non-linear dynamical systems in $\mathbb{R}^d$ perturbed by multifractal Lévy noise. Journal of Statistical Physics, 141:94–119, 2010a

# Lévy-driven SDE approximation of SGD (cont.)

- For simplicity of the presentation we rewrite equation (9) for $d = 1$ case (Multidimensional case[3])

$$d\mathbf{x}_t^\epsilon = -\nabla f(\mathbf{x}_t^\epsilon)dt + \epsilon d\mathbf{L}_t^\alpha \tag{10}$$

  for $t \geq 0$, started from the initial point $\mathbf{x}_0 \in \mathbb{R}$, $\epsilon \geq 0$ is a parameter and $f$ is a non-convex objective with $r \geq 2$ local minima.
- For $\epsilon = 0$ gradient descent: $d\mathbf{x}_t^0 = -\nabla f(\mathbf{x}_t^0)dt$

[3] P. Imkeller, I. Pavlyukevich, and M. Stauch. First exit times of non-linear dynamical systems in $\mathbb{R}^d$ perturbed by multifractal Lévy noise. Journal of Statistical Physics, 141:94–119, 2010a

# Lévy-driven SDE approximation of SGD (cont.)

- For simplicity of the presentation we rewrite equation (9) for $d = 1$ case (Multidimensional case[3])

$$d\mathbf{x}_t^{\epsilon} = -\nabla f(\mathbf{x}_t^{\epsilon})dt + \epsilon d\mathbf{L}_t^{\alpha} \tag{10}$$

  for $t \geq 0$, started from the initial point $\mathbf{x}_0 \in \mathbb{R}$, $\epsilon \geq 0$ is a parameter and $f$ is a non-convex objective with $r \geq 2$ local minima.
- For $\epsilon = 0$ gradient descent: $d\mathbf{x}_t^0 = -\nabla f(\mathbf{x}_t^0)dt$
- When $\epsilon > 0$ these states become 'metastable', there is a positive probability for $x_t^{\epsilon}$ to transition from one basin to another.

[3] P. Imkeller, I. Pavlyukevich, and M. Stauch. First exit times of non-linear dynamical systems in $\mathbb{R}^d$ perturbed by multifractal Lévy noise. Journal of Statistical Physics, 141:94–119, 2010a

# Lévy-driven SDE approximation of SGD (cont.)

- When $\alpha = 2$ (i.e. Gaussianity assumption), the process $x_t^\epsilon$ requires to 'climb' the basin all the way up in order to transfer from one basin to another.

# Lévy-driven SDE approximation of SGD (cont.)

- When $\alpha = 2$ (i.e. Gaussianity assumption), the process $x_t^\epsilon$ requires to 'climb' the basin all the way up in order to transfer from one basin to another.
- But when $\alpha < 2$ the process can incur discontinuous jumps and do not need to cross the boundaries of the basin in order to transition to another one since it can directly jump.

# Lévy-driven SDE approximation of SGD (cont.)

- When $\alpha = 2$ (i.e. Gaussianity assumption), the process $x_t^\epsilon$ requires to 'climb' the basin all the way up in order to transfer from one basin to another.
- But when $\alpha < 2$ the process can incur discontinuous jumps and do not need to cross the boundaries of the basin in order to transition to another one since it can directly jump.
- Under some conditions[4] on $f$, the process (10) admits a stationary density.

[4] G. Samorodnitsky and M. Grigoriu. Tails of solutions of certain nonlinear stochastic differential equations driven by heavy tailed Lévy motions. Stochastic processes and their applications, 105(1):69–97, 2003.

# Validation of the wide minima phenomenon: Setup

- Assume that $f$ is smooth with $r$ local minima $\{m_i\}_{i=1}^r$ separated by $r-1$ local maxima $\{s_i\}_{i=1}^{r-1}$, i.e.,

$$-\infty := s_0 < m_1 < s_1 < \cdots < s_{r-1} < m_r < s_r := \infty$$

**Theorem 1 (Umut Şimşekli, Levent Sagun, and Mert Gürbüzbalaban)**

Under mild conditions, $x^{\epsilon}_{t\epsilon^{-\alpha}} \to Y_m(t)$ as $\epsilon \to 0$, in the sense of finite-dimensional distributions, where $Y = (Y_m(t))_{t\geq 0}$ is a continuous-time Markov chain on a state space $\{m_1, m_2, \cdots, m_r\}$ with the infinitesimal generator $Q = (q_{ij})_{i,j=1}^r$ with

$$q_{ij} = \frac{1}{\alpha}\left|\frac{1}{|s_{j-1} - m_i|^{\alpha}} - \frac{1}{|s_j - m_i|^{\alpha}}\right|; \quad q_{ii} = -\sum_{j\neq i} q_{ij} \tag{11}$$

This process admits a density $\pi$ satisfying $Q^T\pi = 0$.

# Validation of the wide minima phenomenon (cont.)

A consequence of this theorem: Equilibrium probabilities $p_i$ are typically larger for "wide valleys". To see this, consider the case illustrated in Figure (3) with $r = 2$ local minima $m_1 < s_1 = 0 < m_2$

- For this example, $m_2 > |m_1|$, and the second local minimum lies in a wider valley
- A simple computation reveals

$$\pi_1 = \frac{|m_1|^\alpha}{|m_1|^\alpha + m_2^\alpha};\ \pi_2 = \frac{|m_2|^\alpha}{|m_1|^\alpha + |m_2|^\alpha}$$

- We see $\pi_2 > \pi_1$, and $\frac{\pi_2}{\pi_1} = \left(\frac{m_2}{|m_1|}\right)^\alpha$ grows with an exponent $\alpha$ when the ratio $\frac{m_2}{|m_1|}$ of the width of the valleys grows

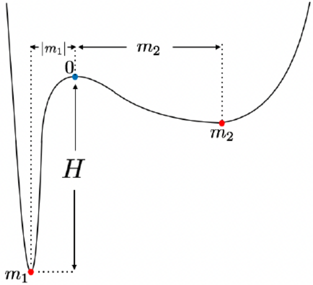



Figure 5: An objective with two local minima $m_1, m_2$ separated by a local maxima at $s_1 = 0$

# Where does the heavy-tail coming from?

- Consider the quadratic loss function $f$ in a linear regression

$$\min_{x \in \mathbb{R}^d} F(x) := \frac{1}{2}\mathbb{E}_{(a,y)\sim\mathcal{D}}\left[(a^T x - y)^2\right] \tag{12}$$

where the data $(a, y)$ comes from an unknown distribution $\mathcal{D}$ with support $\mathbb{R}^d \times \mathbb{R}$.

# Where does the heavy-tail coming from?

- Consider the quadratic loss function $f$ in a linear regression

$$\min_{x \in \mathbb{R}^d} F(x) := \frac{1}{2}\mathbb{E}_{(a,y)\sim\mathcal{D}}\left[(a^T x - y)^2\right] \tag{12}$$

  where the data $(a, y)$ comes from an unknown distribution $\mathcal{D}$ with support $\mathbb{R}^d \times \mathbb{R}$.
- Assume we have access to i.i.d. samples $(a_i, y_i)$ from the distribution $\mathcal{D}$ where $\nabla f^{(i)}(x) = a_i(a_i^T x - y_i)$ is an unbiased estimator of the true gradient $\nabla F(x)$.

# Where does the heavy-tail coming from?

- Consider the quadratic loss function $f$ in a linear regression

$$\min_{x\in\mathbb{R}^d} F(x) := \frac{1}{2}\mathbb{E}_{(a,y)\sim\mathcal{D}}\left[(a^T x - y)^2\right] \tag{12}$$

  where the data $(a, y)$ comes from an unknown distribution $\mathcal{D}$ with support $\mathbb{R}^d \times \mathbb{R}$.
- Assume we have access to i.i.d. samples $(a_i, y_i)$ from the distribution $\mathcal{D}$ where $\nabla f^{(i)}(x) = a_i(a_i^T x - y_i)$ is an unbiased estimator of the true gradient $\nabla F(x)$.
- In this settings, SGD with batch-size $b$ leads to the iteration

$$x_k = M_k x_{k-1} + q_k, \tag{13}$$

  with

$$M_k := I - \frac{\eta}{b}H_k, \qquad H_k := \sum_{i\in\Omega_k} a_i a_i^T, \qquad q_k := \frac{\eta}{b}\sum_{i\in\Omega_k}(a_i y_i),$$

  where $\Omega_k := \{b(k-1)+1, b(k-1)+2, \cdots, bk\}$

# Heavy-Tail Phenomenon in Linear Regression

- Assumption (A1): $a_i$'s are i.i.d with a continuous distribution supported on $\mathbb{R}^d$ with all the moments finite. All the moments of $a_i$ are finite

# Heavy-Tail Phenomenon in Linear Regression

- Assumption (A1): $a_i$'s are i.i.d with a continuous distribution supported on $\mathbb{R}^d$ with all the moments finite. All the moments of $a_i$ are finite
- Assumption (A2): $y_i$ are i.i.d with a continuous density whose support is $\mathbb{R}$ with all the moments finite.

# Heavy-Tail Phenomenon in Linear Regression

- Assumption (A1): $a_i$'s are i.i.d with a continuous distribution supported on $\mathbb{R}^d$ with all the moments finite. All the moments of $a_i$ are finite
- Assumption (A2): $y_i$ are i.i.d with a continuous density whose support is $\mathbb{R}$ with all the moments finite.
- Let us define

$$h(s) := \lim_{k\to\infty} \left(\mathbb{E}\|M_k M_{k-1} \cdots M_1\|^s\right)^{\frac{1}{k}} \tag{14}$$

which arises in stochastic matrix recursions[5]

[5] D. Buraczewski, E. Damek, Y. Guivarc'h, and S. Mentemeier. On multidimensional mandelbrot cascades. Journal of Difference Equations and Applications, 20(11):1523–1567, 2014.

# Heavy-Tail Phenomenon in Linear Regression (cont.)

Since $\mathbb{E}\|M_k\|^s < \infty$ for all $k$ and $s > 0$, we have $h(s) < \infty$. Let us also define $\Pi_k := M_k M_{k-1} \cdots M_1$ and

$$\rho := \lim_{k\to\infty} \frac{1}{2k} \log\left(\text{largest eigenvalue of } \Pi_k^T \Pi_k\right) \tag{15}$$

The latter quantity is called the top Lyapunov exponent of the stochastic recursion.

# Heavy-Tail Phenomenon in Linear Regression (cont.)

Since $\mathbb{E}\|M_k\|^s < \infty$ for all $k$ and $s > 0$, we have $h(s) < \infty$. Let us also define $\Pi_k := M_k M_{k-1} \cdots M_1$ and

$$\rho := \lim_{k\to\infty} \frac{1}{2k} \log\left(\text{largest eigenvalue of } \Pi_k^T \Pi_k\right) \tag{15}$$

The latter quantity is called the top Lyapunov exponent of the stochastic recursion.

**Theorem 2 (Gürbüzbalaban, Şimşekli, and Zhu(2021))**

Consider the SGD iterations (13). If $\rho < 0$ and there exists a unique positive $\alpha$ such that $h(\alpha) = 1$, then (13) admits a unique stationary solution $x_\infty$ and the SGD iterations converge to $x_\infty$ in distribution, where the distribution of $x_\infty$ satisfies

$$\lim_{t\to\infty} t^\alpha \mathbb{P}(u^T x_\infty > t) = e_\alpha(u),\ \ u \in \mathbb{S}^{d-1} \tag{16}$$

for some positive and continuous function $e_\alpha$ on $\mathbb{S}^{d-1}$

# Heavy-Tail Phenomenon in Linear Regression (cont.)

- In order to have a more explicit characterization of the tail-index, we will make the following additional assumption
  Assumption (A3): $a_i \sim \mathcal{N}(0, \sigma^2 I_d)$ are Gaussian for every $i$.

# Heavy-Tail Phenomenon in Linear Regression (cont.)

- In order to have a more explicit characterization of the tail-index, we will make the following additional assumption
  Assumption (A3): $a_i \sim \mathcal{N}(0, \sigma^2 I_d)$ are Gaussian for every $i$.
- Under (A3), next result shows that the formulas for $\rho$ and $h(s)$ can be simplified.

# Heavy-Tail Phenomenon in Linear Regression (cont.)

- In order to have a more explicit characterization of the tail-index, we will make the following additional assumption
  Assumption (A3): $a_i \sim \mathcal{N}(0, \sigma^2 I_d)$ are Gaussian for every $i$.
- Under (A3), next result shows that the formulas for $\rho$ and $h(s)$ can be simplified.
- Let $H$ be a matrix with the same distribution as $H_k$, and $e_1$ be the first basis vector. Define

$$\begin{aligned} \tilde{\rho} &:= \mathbb{E} \log \left\| \left( I - \frac{\eta}{b} H \right) e_1 \right\| \\ \tilde{h}(s) &:= \mathbb{E} \left[ \left\| \left( I - \frac{\eta}{b} H \right) e_1 \right\|^s \right] \text{ for } \rho < 0 \end{aligned} \tag{17}$$

# Heavy-Tail Phenomenon in Linear Regression (cont.)

**Theorem 3 (Gürbüzbalaban, Şimşekli, and Zhu(2021)**

Assume (A3) holds. Consider the SGD iterations (13). If $\rho < 0$, then (i) there exists a unique positive $\alpha$ such that $h(\alpha) = 1$ and (16) holds; (ii) we have $\rho = \tilde{\rho}$ and $h(s) = \tilde{h}(s)$, where $\tilde{\rho}$ and $\tilde{h}(s)$ are defined in (17).

# Heavy-Tail Phenomenon in Linear Regression (cont.)

**Theorem 3 (Gürbüzbalaban, Şimşekli, and Zhu(2021)**

Assume (A3) holds. Consider the SGD iterations (13). If $\rho < 0$, then (i) there exists a unique positive $\alpha$ such that $h(\alpha) = 1$ and (16) holds; (ii) we have $\rho = \tilde{\rho}$ and $h(s) = \tilde{h}(s)$, where $\tilde{\rho}$ and $\tilde{h}(s)$ are defined in (17).

**Theorem 4 (Gürbüzbalaban, Şimşekli, and Zhu(2021)**

Assume (A3) holds. The tail-index $\alpha$ is strictly increasing in batch-size $b$ and strictly decreasing in stepsize $\eta$ and variance $\sigma^2$ provided that $\alpha \geq 1$. Moreover, the tail-index $\alpha$ is strictly decreasing in dimension $d$.

# Characterization of the tail-index $\alpha$

- Under assumption (A3) next we notice the characterization of the tail-index $\alpha$ depending on the choice of the batch-size $b$, the variance $\sigma^2$, which determines the curvature around the minimum and the stepsize.

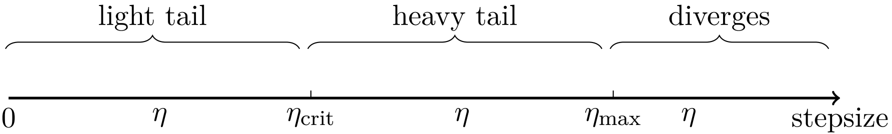
# Characterization of the tail-index $\alpha$

- Under assumption (A3) next we notice the characterization of the tail-index $\alpha$ depending on the choice of the batch-size $b$, the variance $\sigma^2$, which determines the curvature around the minimum and the stepsize.
- In particular we show that if the stepsize exceeds an explicit threshold, the stationary distribution will become heavy tailed with an infinite variance.

**Proposition 1 (Gürbüzbalaban, Şimşekli, and Zhu(2021))**

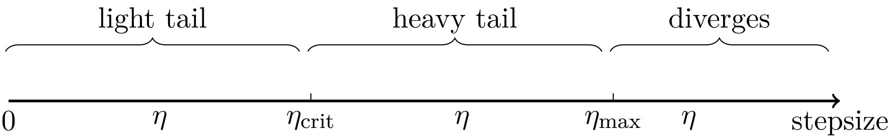
Assume (A3) holds. Let $\eta_{crit} := \frac{2b}{\sigma^2(d+b+1)}$
The following holds:
i. There exists $\eta_{max} > \eta_{crit}$ such that for any $\eta_{crit} < \eta < \eta_{max}$, Theorem 2 holds with tail-index $0 < \alpha < 2$
ii. If $\eta = \eta_{crit}$, Theorem 2 holds with tail-index $\alpha = 2$
iii. If $\eta \in (0, \eta_{crit})$, then Theorem 2 holds with tail-index $\alpha > 2$

# Three regimes for $\eta$



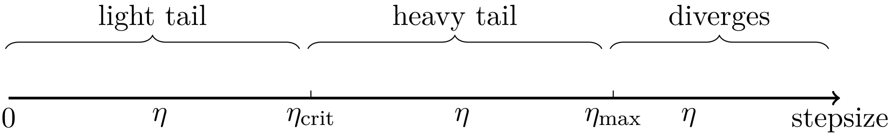

- (I) convergence to a limit with a finite variance if $\rho < 0$ and $\alpha > 2$

# Three regimes for $\eta$



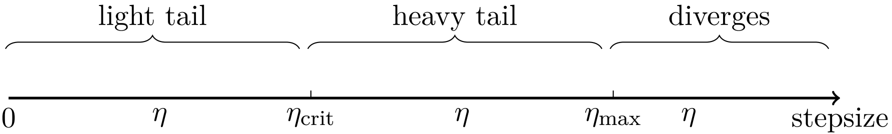

- (I) convergence to a limit with a finite variance if $\rho < 0$ and $\alpha > 2$
- (II) convergence to a heavy-tailed limit with infinite variance if $\rho < 0$ and $\alpha < 2$

# Three regimes for $\eta$




- (I) convergence to a limit with a finite variance if $\rho < 0$ and $\alpha > 2$
- (II) convergence to a heavy-tailed limit with infinite variance if $\rho < 0$ and $\alpha < 2$
- (III) $\rho > 0$ when convergence cannot be guaranteed.

# Three regimes for $\eta$




- (I) convergence to a limit with a finite variance if $\rho < 0$ and $\alpha > 2$
- (II) convergence to a heavy-tailed limit with infinite variance if $\rho < 0$ and $\alpha < 2$
- (III) $\rho > 0$ when convergence cannot be guaranteed.
- For Gaussian input
  - if $\eta < \eta_{crit}$, by Proposition 1, $\rho < 0$ and $\alpha > 2$, therefore regime (I) applies
  - if $\eta_{crit} < \eta < \eta_{max}$, then $\alpha < 2$ thus regime II applies

# Experiment on synthetic data

Model setup:

- $x_0 \sim \mathcal{N}(0, \sigma_x^2 I)$,
- $a_i \sim \mathcal{N}(0, \sigma^2 I)$
- $y_i | a_i, x_0 \sim \mathcal{N}\left(a_i^T x_0, \sigma_y^2\right)$
- where $x_0, a_0 \in \mathbb{R}^d$, $y_i \in \mathbb{R}$ for all $i = 1, 2, \cdots, n$ and $\sigma, \sigma_x, \sigma_y > 0$

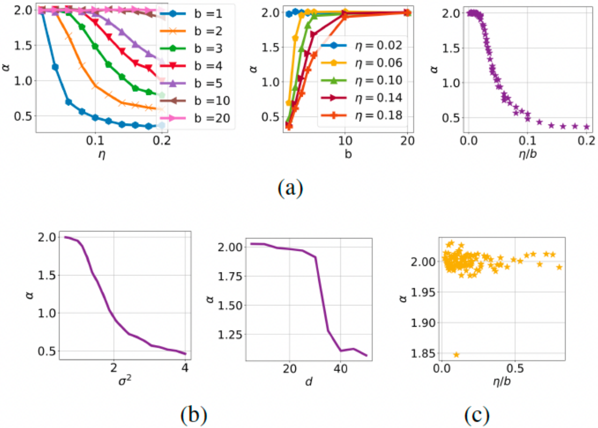



Figure 6: Behavior of $\alpha$ with (a) variying the step size $\eta$ and batch-size $b$, (b) $d$ and $\sigma$, (c) under RMSProp

# Results from a fully connected neural network

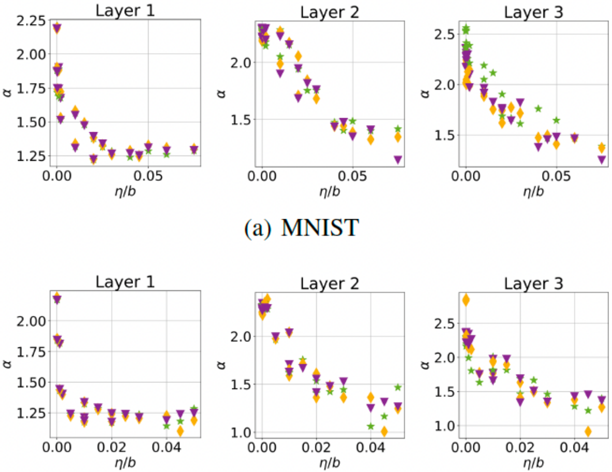



Figure 7: Results on FCNs. Different markers represent different initialization with the same $\eta, b$.

# What is decentralized SGD?

- This present era is the era of big data, artificial intelligence, and machine learning.

# What is decentralized SGD?

- This present era is the era of big data, artificial intelligence, and machine learning.
- There has been an exponential growth in the amount of data collection through various source such as smart phones, tablets, sensors or video cameras are major sources of data generation.

# What is decentralized SGD?

- This present era is the era of big data, artificial intelligence, and machine learning.
- There has been an exponential growth in the amount of data collection through various source such as smart phones, tablets, sensors or video cameras are major sources of data generation.
- Often these devices are connected over a communication network (such as a wireless network or a sensor network) that has a high latency or a limited bandwidth.

# What is decentralized SGD?

- This present era is the era of big data, artificial intelligence, and machine learning.
- There has been an exponential growth in the amount of data collection through various source such as smart phones, tablets, sensors or video cameras are major sources of data generation.
- Often these devices are connected over a communication network (such as a wireless network or a sensor network) that has a high latency or a limited bandwidth.
- Because of communication constraints and privacy constraints, gathering all these data for centralized processing is often impractical or infeasible.

# What is decentralized SGD?

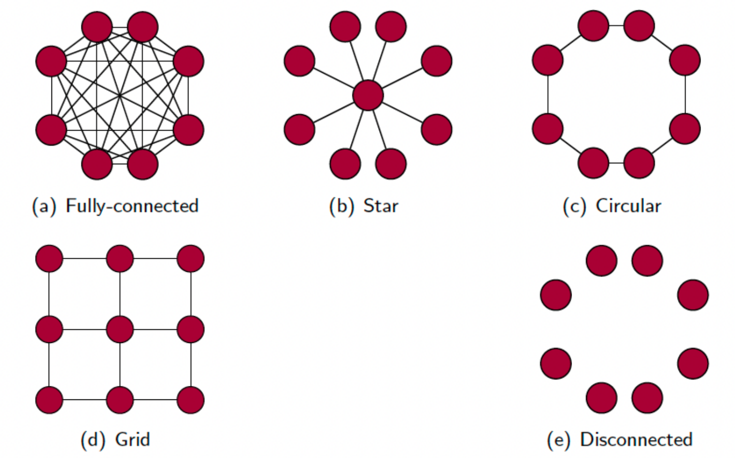



Figure 8: Illustration of the network architectures.

# Preliminaries before jump into DE-SGD

Decentralized stochastic optimization problems of the form

$$\min_{x\in\mathbb{R}^d} f(x) = \sum_{i=1}^{N} f_i(x);\;\; f_i(x) = \mathbb{E}_{z\sim\mathcal{D}_i}[\ell(x, z_i)] \tag{18}$$

where $\ell(x, z_i)$ represents the instantaneous loss at node $i$ based on the predictor $x$ and the data point $z_i$

# Preliminaries before jump into DE-SGD

Decentralized stochastic optimization problems of the form

$$\min_{x \in \mathbb{R}^d} f(x) = \sum_{i=1}^{N} f_i(x); \;\; f_i(x) = \mathbb{E}_{z \sim \mathcal{D}_i}[\ell(x, z_i)] \tag{18}$$

where $\ell(x, z_i)$ represents the instantaneous loss at node $i$ based on the predictor $x$ and the data point $z_i$

- We have $N$ computation nodes lying on a connected undirected graph $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, 2, \cdots, N\}$ is the set of (vertices) nodes and $\mathcal{E} \subseteq (\mathcal{V} \times \mathcal{V})$ is the set of edges that define the connectivity patterns between the nodes.

# Preliminaries before jump into DE-SGD

Decentralized stochastic optimization problems of the form

$$\min_{x \in \mathbb{R}^d} f(x) = \sum_{i=1}^{N} f_i(x); \;\; f_i(x) = \mathbb{E}_{z \sim \mathcal{D}_i}[\ell(x, z_i)] \tag{18}$$

where $\ell(x, z_i)$ represents the instantaneous loss at node $i$ based on the predictor $x$ and the data point $z_i$

- We have $N$ computation nodes lying on a connected undirected graph $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, 2, \cdots, N\}$ is the set of (vertices) nodes and $\mathcal{E} \subseteq (\mathcal{V} \times \mathcal{V})$ is the set of edges that define the connectivity patterns between the nodes.
- The objective $f_i$ is only available at the node $i$, and the aim is to train models locally at each agent where only local parameters vectors are shared among the neighbors.

# Decentralized SGD

- The DE-SGD algorithm consists of a weighted averaging with the local variables $x_i^{(k)}$ of node $i$'s immediate neighbors $j \in \Omega_i := \{j : (i,j) \in \mathcal{G}\}$ as well as a stochastic gradient step over the node's component function $f_i(x)$, i.e.,

$$x_i^{(k)} = \sum_{\ell \in \Omega_i} W_{i\ell} x_\ell^{(k-1)} - \eta \tilde{\nabla} f_i \left( x_i^{(k-1)} \right) \tag{19}$$

# Decentralized SGD

- The DE-SGD algorithm consists of a weighted averaging with the local variables $x_i^{(k)}$ of node $i$'s immediate neighbors $j \in \Omega_i := \{j : (i,j) \in \mathcal{G}\}$ as well as a stochastic gradient step over the node's component function $f_i(x)$, i.e.,

$$x_i^{(k)} = \sum_{\ell \in \Omega_i} W_{i\ell} x_\ell^{(k-1)} - \eta \tilde{\nabla} f_i \left( x_i^{(k-1)} \right) \tag{19}$$

- $\tilde{\nabla} f_i(x)$ is an estimate of the gradient of the loss $f_i(x)$ at node $i$ based on a batch size $b_i$, satisfying,

$$\tilde{\nabla} f_i(x_i^{(k)}) := \frac{1}{b_i} \sum_{j=1}^{b_i} \nabla \ell \left( x_i^{(k)}, z_{i,j}^{(k)} \right) \tag{20}$$

# Decentralized SGD

- The DE-SGD algorithm consists of a weighted averaging with the local variables $x_i^{(k)}$ of node $i$'s immediate neighbors $j \in \Omega_i := \{j : (i,j) \in \mathcal{G}\}$ as well as a stochastic gradient step over the node's component function $f_i(x)$, i.e.,

$$x_i^{(k)} = \sum_{\ell \in \Omega_i} W_{i\ell} x_\ell^{(k-1)} - \eta \tilde{\nabla} f_i \left( x_i^{(k-1)} \right) \tag{19}$$

- $\tilde{\nabla} f_i(x)$ is an estimate of the gradient of the loss $f_i(x)$ at node $i$ based on a batch size $b_i$, satisfying,

$$\tilde{\nabla} f_i(x_i^{(k)}) := \frac{1}{b_i} \sum_{j=1}^{b_i} \nabla \ell \left( x_i^{(k)}, z_{i,j}^{(k)} \right) \tag{20}$$

- $W \in \mathbb{R}^{N \times N}$ is a symmetric double stochastic weight matrix, with $W_{ij} = W_{ji} > 0$ if $j \in \Omega_i$, $W_{ij} = W_{ji} = 0$ if $j \notin \Omega_i$ and $i \neq j$, and $W_{ii} = 1 - \sum_{j \neq i} W_{ij} > 0$ for every $1 \leq i \leq N$

# DE-SGD as centralized SGD

Following Fallah et al.[6] we can express the DE-SGD iterations as

$$x^{(k)} = \mathcal{W}x^{(k-1)} - \eta\tilde{\nabla}F\left(x^{(k-1)}\right)$$

where, $\mathcal{W} := W \otimes I_d$, $F : \mathbb{R}^{Nd} \to \mathbb{R}$ defined as
$F(x) := F(x_1, x_2, \cdots, x_N) = \sum_{i=1}^{N} f_i(x_i)$, with

$$x^{(k)} := \left[\left(x_1^{(k)}\right)^T, \left(x_2^{(k)}\right)^T, \cdots, \left(x_N^{(k)}\right)^T\right]^T \in \mathbb{R}^{Nd}$$

[6] A. Fallah, M. Gürbüzbalaban, A. Ozdaglar, U. Şimşekli, and L. Zhu. Robust distributed accelerated stochastic gradient methods for multi-agent networks. The Journal of Machine Learning Research, 23(1):9893–9988, 2022.

# DE-SGD as centralized SGD

Following Fallah et al.[6] we can express the DE-SGD iterations as

$$x^{(k)} = \mathcal{W}x^{(k-1)} - \eta\tilde{\nabla}F\left(x^{(k-1)}\right)$$

where, $\mathcal{W} := W \otimes I_d$, $F : \mathbb{R}^{Nd} \to \mathbb{R}$ defined as
$F(x) := F(x_1, x_2, \cdots, x_N) = \sum_{i=1}^{N} f_i(x_i)$, with

$$x^{(k)} := \left[\left(x_1^{(k)}\right)^T, \left(x_2^{(k)}\right)^T, \cdots, \left(x_N^{(k)}\right)^T\right]^T \in \mathbb{R}^{Nd}$$

and

$$\tilde{\nabla}F\left(x^k\right) := \left[\left(\tilde{\nabla}f_1\left(x_1^{(k)}\right)\right)^T, \left(\tilde{\nabla}f_2\left(x_2^{(k)}\right)\right)^T, \cdots, \left(\tilde{\nabla}f_N\left(x_N^{(k)}\right)\right)^T\right] \tag{21}$$

[6] A. Fallah, M. Gürbüzbalaban, A. Ozdaglar, U. Şimşekli, and L. Zhu. Robust distributed accelerated stochastic gradient methods for multi-agent networks. The Journal of Machine Learning Research, 23(1):9893–9988, 2022.

# DE-SGD as centralized SGD (cont.)

- We can alternatively view DE-SGD as C-SGD iterations

$$x^{(k)} = x^{(k-1)} - \eta \tilde{\nabla} F_{\mathcal{W}} \left( x^{(k-1)} \right) \tag{22}$$

# DE-SGD as centralized SGD (cont.)

- We can alternatively view DE-SGD as C-SGD iterations

$$x^{(k)} = x^{(k-1)} - \eta \tilde{\nabla} F_{\mathcal{W}} \left( x^{(k-1)} \right) \tag{22}$$

- On a modified objective $F_{\mathcal{W}}$ defined as

$$F_{\mathcal{W}}(x) := F(x) + \frac{1}{2\eta} x^T (I_{Nd} - \mathcal{W}) x \tag{23}$$

with the convention that $\tilde{\nabla} F_{\mathcal{W}}(x) = \tilde{\nabla} F(x) + \frac{1}{\eta}(I_{Nd} - \mathcal{W})x$

# DE-SGD as centralized SGD (cont.)

- Similar to (20), we can define the stochastic Hessian as

$$\tilde{\nabla}^2 f_i\left(x_i^{(k)}\right) := \frac{1}{b_i}\sum_{j=1}^{b_i} \nabla^2 \ell\left(x_i^{(k)}, z_{i,j}^{(k)}\right)$$

with

$$\tilde{\nabla}^2 F\left(x^{(k)}\right) := \left[\left(\tilde{\nabla}^2 f_1\left(x_1^{(k)}\right)\right)^T, \left(\tilde{\nabla}^2 f_2\left(x_2^{(k)}\right)\right)^T, \cdots, \left(\tilde{\nabla}^2 f_N\left(x_N^{(k)}\right)\right)^T\right] \tag{24}$$

when $f_i$'s are twice differentiable for every $i = 1, 2, \cdots, N$

# DE-SGD and heavy-tail for quadratic loss

- The loss function in (18) is a quadratic of the form $\ell(x, z_i) = \frac{1}{2}\left(a_i^T x - y_i\right)^2$ for every node $i$, where $z_i = (a_i, y_i)$ is the local data at agent $i$ with $a_i$ representing the input feature vector and $y_i$ being the output.

# DE-SGD and heavy-tail for quadratic loss

- The loss function in (18) is a quadratic of the form $\ell(x, z_i) = \frac{1}{2}\left(a_i^T x - y_i\right)^2$ for every node $i$, where $z_i = (a_i, y_i)$ is the local data at agent $i$ with $a_i$ representing the input feature vector and $y_i$ being the output.
- Recall that, each node $i$ has access to $b_i$ samples from data $\left\{z_{i,j}^{(k)} = \left(a_{i,j}^{(k)}, y_{i,j}^{(k)}\right)\right\}_{j=1}^{n_i}$ at every iteration $k$ to form a stochastic gradient estimate, (20) becomes

$$\tilde{\nabla} f_i(x_i^{(k)}) := \frac{1}{b_i} \sum_{j=1}^{b_i} \left[a_{i,j}^{(k)} \left(a_{i,j}^{(k)}\right)^T x_i^{(k)} - y_{i,j}^{(k)} a_{i,j}^{(k)}\right]$$

# DE-SGD and heavy-tail for quadratic loss (cont.)

DE-SGD iteration becomes

$$x^k = M^{(k)}x^{(k-1)} + q^{(k)}, \text{ where } M^{(k)} := \mathcal{W} - \eta H^{(k)} \tag{25}$$

# DE-SGD and heavy-tail for quadratic loss (cont.)

DE-SGD iteration becomes

$$x^k = M^{(k)} x^{(k-1)} + q^{(k)}, \text{ where } M^{(k)} := \mathcal{W} - \eta H^{(k)} \tag{25}$$

with

$$H^{(k)} := \text{blkdiag}\left(\left\{H_i^{(k)}\right\}_{i=1}^{N}\right); \;\; q^{(k)} := \left[\left(q_1^{(k)}\right)^T, \left(q_2^{(k)}\right)^T, \cdots, \left(q_N^{(k)}\right)^T\right]^T$$

# DE-SGD and heavy-tail for quadratic loss (cont.)

DE-SGD iteration becomes

$$x^k = M^{(k)} x^{(k-1)} + q^{(k)}, \text{ where } M^{(k)} := \mathcal{W} - \eta H^{(k)} \tag{25}$$

with

$$H^{(k)} := \text{blkdiag}\left(\left\{H_i^{(k)}\right\}_{i=1}^{N}\right); \; q^{(k)} := \left[\left(q_1^{(k)}\right)^T, \left(q_2^{(k)}\right)^T, \cdots, \left(q_N^{(k)}\right)^T\right]^T$$

where for $i = 1, 2, \cdots, N$

$$H_i^{(k)} := \frac{1}{b_i}\sum_{j=1}^{b_i} a_{i,j}^{(k)} \left(a_{i,j}^{(k)}\right)^T; \; q_i^{(k)} := \frac{\eta}{b_i}\sum_{j=1}^{b_i} a_{i,j}^{(k)} y_{i,j}^{(k)} \tag{26}$$

where $a_{i,j}^{(k)}$ and $y_{i,j}^{(k)}$ are i.i.d over over $k$ random draws from the data with the same distribution as $a_{i,j}, y_{i,j}$ that satisfy some assumptions

# DE-SGD and heavy-tail for quadratic loss (cont.)

**Assumptions**

Assumption (A4): For every $i$, $a_{i,j}$'s are i.i.d over $j$ following a continuous distribution supported on $\mathbb{R}^d$ with all moments finite.
Assumption (A5): For every $i = 1, 2, \cdots, N$, $y_{i,j}$ are i.i.d over $j$ with continuous density whose support is $\mathbb{R}$ with all the moments finite

- We recall the concatenated iterates

$$x^{(k)} = M^{(k)} x^{(k-1)} + q^{(k)}$$

where $M^{(k)}, q^{(k)}$ are defined by (25), (26)

# DE-SGD and heavy-tail for quadratic loss (cont.)

**Assumptions**

Assumption (A4): For every $i$, $a_{i,j}$'s are i.i.d over $j$ following a continuous distribution supported on $\mathbb{R}^d$ with all moments finite.
Assumption (A5): For every $i = 1, 2, \cdots, N$, $y_{i,j}$ are i.i.d over $j$ with continuous density whose support is $\mathbb{R}$ with all the moments finite

- We recall the concatenated iterates

$$x^{(k)} = M^{(k)} x^{(k-1)} + q^{(k)}$$

where $M^{(k)}, q^{(k)}$ are defined by (25), (26)

- Let us introduce

$$h(s) := \lim_{k\to\infty} \left(\mathbb{E} \left\| M^{(k)}, M^{(k-1)}, \cdots, M^{(1)} \right\|^s \right)^{\frac{1}{k}} \tag{27}$$

which arises in stochastic matrix recursions, where $\|\cdot\|$ denotes the matrix 2-norm (i.e. largest singular value of a matrix).

# DE-SGD and heavy-tail for quadratic loss (cont.)

- Since $\mathbb{E}\left\|M^{(k)}\right\|^s < \infty$ for all $k$ and $s > 0$, we have $h(s) < \infty$. We define $\Pi^{(k)} := M^{(k)}M^{(k-1)}\cdots, M^{(1)}$ and

$$\rho := \lim_{k\to\infty} \frac{1}{2k} \log\left(\text{largest eigenvalue of } \left(\Pi^{(k)}\right)^T \Pi^{(k)}\right) \tag{28}$$

The latter quantity is called the top Lyapunov exponent.

**Theorem 5 (Gürbüzbalaban, Hu, Şimşekli, Yuan, and Zhu.)**

Suppose (A4)-(A5) hold. Consider the DE-SGD iterations (25). If $\rho < 0$ and $\exists!$ positive $\alpha \ni h(\alpha) = 1$, then (25) admits a unique stationary solution $x^\infty$ and $x^{(k)} \to x^\infty$ in distribution, where the distribution of $x^\infty$ satisfies

$$\lim_{t\to\infty} t^\alpha \mathbb{P}\left(u^T x^\infty > t\right) = g_\alpha(u)$$

for any $u \in \mathbb{S}^{Nd-1}$, for some positive and continuous function $g_\alpha$ on $\mathbb{S}^{Nd-1}$

# DE-SGD and heavy-tail for quadratic loss (cont.)

- Theorem 5 provides a formula for the tail-index $\alpha$. But there is not closed form formula for $\rho$ and $h(s)$

# DE-SGD and heavy-tail for quadratic loss (cont.)

- Theorem 5 provides a formula for the tail-index $\alpha$. But there is not closed form formula for $\rho$ and $h(s)$
- So we use the following estimates

$$\begin{aligned} \rho \le \tilde{\rho} &:= \mathbb{E} \log \|\mathcal{W} - \eta H\| \\ h(s) \le \tilde{h}(s) &:= \mathbb{E}\left[\|\mathcal{W} - \eta H\|^{s}\right] \end{aligned} \tag{29}$$

where $H$ is a matrix that has the same distribution as $H^{(k)}$ (which does not depend on $k$).

# DE-SGD and heavy-tail for quadratic loss (cont.)

- Theorem 5 provides a formula for the tail-index $\alpha$. But there is not closed form formula for $\rho$ and $h(s)$
- So we use the following estimates

$$\begin{aligned} \rho \leq \tilde{\rho} &:= \mathbb{E} \log \|\mathcal{W} - \eta H\| \\ h(s) \leq \tilde{h}(s) &:= \mathbb{E}\left[\|\mathcal{W} - \eta H\|^s\right] \end{aligned} \tag{29}$$

where $H$ is a matrix that has the same distribution as $H^{(k)}$ (which does not depend on $k$).

**Lower bounds on the tail-index $\alpha$:**

If $\hat{\alpha}$ is such that $\hat{h}(\hat{\alpha}) = 1$ , then by (29), $\hat{\alpha}$ is a lower bound on the tail-index $\alpha$ that satisfies $h(\alpha) = 1$ where $h$ is defined as in (27). In other words, we have $\hat{\alpha} \leq \alpha$ and therefore $\hat{\alpha}$ serves as a lower bound on the tail-index.

# Tail-index comparison between Dis-SGD and C-SGD

We start with defining properly what exactly we mean by Disconnected SGD and Centralized SGD iterations.

## Disconnected SGD (Dis-SGD)

Disconnected SGD (Dis-SGD) corresponds to the case $W = I$ (where nodes do not share information with other nodes), and for every $i = 1, 2, \cdots, N$, the iterates follow the recursion:
$x_i^k = x_i^{(k-1)} - \eta \tilde{\nabla} f_i\left(x_i^{(k-1)}\right)$, where each gradient $\tilde{\nabla} f_i\left(x_i^{(k-1)}\right)$ is based on $b_i$ samples from node $i$'s dataset. The total number of samples (cumulatively over the nodes) equals $\sum_{i=1}^N b_i$.

# Tail-index comparison between Dis-SGD and C-SGD

We start with defining properly what exactly we mean by Disconnected SGD and Centralized SGD iterations.

## Disconnected SGD (Dis-SGD)

Disconnected SGD (Dis-SGD) corresponds to the case $W = I$ (where nodes do not share information with other nodes), and for every $i = 1, 2, \cdots, N$, the iterates follow the recursion: $x_i^k = x_i^{(k-1)} - \eta \tilde{\nabla} f_i \left( x_i^{(k-1)} \right)$, where each gradient $\tilde{\nabla} f_i \left( x_i^{(k-1)} \right)$ is based on $b_i$ samples from node $i$'s dataset. The total number of samples (cumulatively over the nodes) equals $\sum_{i=1}^{N} b_i$.

## Centralized SGD (C-SGD)

Centralized SGD (C-SGD) consists of the iterations $x_k = x_k - \eta \tilde{\nabla} f(x_{k-1})$, where we take (number of data points per iteration) batch-size to be $\sum_{i=1}^{N} b_i$ for centralized SGD.

# Tail-index comparison between Dis-SGD and C-SGD

- We will first be comparing C-SGD to Dis-SGD and then we will be comparing it to the DE-SGD.

# Tail-index comparison between Dis-SGD and C-SGD

- We will first be comparing C-SGD to Dis-SGD and then we will be comparing it to the DE-SGD.
- To make it easier, we assume that $b_i \equiv b$ and $a_{i,j}$ are i.i.d. over $i$ and $j$

# Tail-index comparison between Dis-SGD and C-SGD

- We will first be comparing C-SGD to Dis-SGD and then we will be comparing it to the DE-SGD.
- To make it easier, we assume that $b_i \equiv b$ and $a_{i,j}$ are i.i.d. over $i$ and $j$
- Under the assumptions (A4)-(A5), $x_i^{(k)}$ are independent and that $\hat{\alpha} = \hat{\alpha}(b)$ is a lower bound of the tail-index of $x_i^{\infty}$ which is the unique positive value satisfying $\hat{h}(\hat{\alpha}(b)) = 1$

# Tail-index comparison between Dis-SGD and C-SGD

- We will first be comparing C-SGD to Dis-SGD and then we will be comparing it to the DE-SGD.
- To make it easier, we assume that $b_i \equiv b$ and $a_{i,j}$ are i.i.d. over $i$ and $j$
- Under the assumptions (A4)-(A5), $x_i^{(k)}$ are independent and that $\hat{\alpha} = \hat{\alpha}(b)$ is a lower bound of the tail-index of $x_i^{\infty}$ which is the unique positive value satisfying $\hat{h}(\hat{\alpha}(b)) = 1$
- Where $\hat{h}(s) = \mathbb{E}\left[\|I - \eta H\|^s\right]$, where $\hat{\alpha}(b)$ emphasize the dependence on the batch-size $b$ such that for each node $i$, $b$ data points are chosen.

# Tail-index comparison between Dis-SGD and C-SGD

- We will first be comparing C-SGD to Dis-SGD and then we will be comparing it to the DE-SGD.
- To make it easier, we assume that $b_i \equiv b$ and $a_{i,j}$ are i.i.d. over $i$ and $j$
- Under the assumptions (A4)-(A5), $x_i^{(k)}$ are independent and that $\hat{\alpha} = \hat{\alpha}(b)$ is a lower bound of the tail-index of $x_i^{\infty}$ which is the unique positive value satisfying $\hat{h}(\hat{\alpha}(b)) = 1$
- Where $\hat{h}(s) = \mathbb{E}\left[\|I - \eta H\|^s\right]$, where $\hat{\alpha}(b)$ emphasize the dependence on the batch-size $b$ such that for each node $i$, $b$ data points are chosen.
- We use $\hat{\alpha}(b)$ as a proxy of the tail-index.

# Tail-index comparison between Dis-SGD and C-SGD

- We will first be comparing C-SGD to Dis-SGD and then we will be comparing it to the DE-SGD.
- To make it easier, we assume that $b_i \equiv b$ and $a_{i,j}$ are i.i.d. over $i$ and $j$
- Under the assumptions (A4)-(A5), $x_i^{(k)}$ are independent and that $\hat{\alpha} = \hat{\alpha}(b)$ is a lower bound of the tail-index of $x_i^{\infty}$ which is the unique positive value satisfying $\hat{h}(\hat{\alpha}(b)) = 1$
- Where $\hat{h}(s) = \mathbb{E}\left[\|I - \eta H\|^s\right]$, where $\hat{\alpha}(b)$ emphasize the dependence on the batch-size $b$ such that for each node $i$, $b$ data points are chosen.
- We use $\hat{\alpha}(b)$ as a proxy of the tail-index.
- In C-SGD, $bN$ data points are chosen at each iteration, and hence the batch-size equals $bN$

# Tail-index comparison between Dis-SGD and C-SGD

- The corresponding tail-index (proxy) is $\hat{\alpha}(bN)$

# Tail-index comparison between Dis-SGD and C-SGD

- The corresponding tail-index (proxy) is $\hat{\alpha}(bN)$
- We have the following observation by adapting the the monotonicity properties of tail-index shown in the paper by Gürbüzbalaban et al.[7], Theorem 4.

**Proposition 2**

The tail-index for disconnected SGD is smaller than that of the centralized SGD. Indeed, their difference gets larger as the network size increases.

[7] Mert Gurbuzbalaban, Umut Simsekli, and Lingjiong Zhu. The heavy-tail phenomenon in sgd. International Conference on Machine Learning, pages 3964–3975. PMLR, 2021.

# Tail-index comparison between DE-SGD and C-SGD

- We assume $d = 1$, $b_i \equiv 1$ and $\sigma_i = \sigma$

# Tail-index comparison between DE-SGD and C-SGD

- We assume $d = 1$, $b_i \equiv 1$ and $\sigma_i = \sigma$
- In Proposition 2, we showed that the tail-index for Dis-SGD is smaller than that of the C-SGD

# Tail-index comparison between DE-SGD and C-SGD

- We assume $d = 1$, $b_i \equiv 1$ and $\sigma_i = \sigma$
- In Proposition 2, we showed that the tail-index for Dis-SGD is smaller than that of the C-SGD
- For DE-SGD, chosen mixing matrix $W = I_N - \delta L$ where $\delta > 0$ is small enough so that the spectral radius of $W$ is not larger than 1, and $L$ is a graph Laplacian.

# Tail-index comparison between DE-SGD and C-SGD

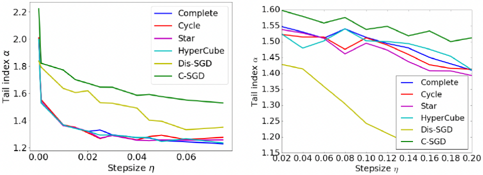
- We assume $d = 1$, $b_i \equiv 1$ and $\sigma_i = \sigma$
- In Proposition 2, we showed that the tail-index for Dis-SGD is smaller than that of the C-SGD
- For DE-SGD, chosen mixing matrix $W = I_N - \delta L$ where $\delta > 0$ is small enough so that the spectral radius of $W$ is not larger than 1, and $L$ is a graph Laplacian.
- Under some mild assumption we can show when $\delta$ is small, the tail-index for the DE-SGD is smaller than that of the Dis-SGD given the stepsize $\eta$ or network size $N$ is large.

**Corollary**

Under some mild assumptions the tail-index $\hat{\alpha}$ of the decentralized SGD is smaller than that of the centralized SGD provided that the stepsize $\eta$ or network size $N$ is large and $\delta$ is small.

# Deep learning experiment

- FCN on MNIST: we assume there are $N = 8$ nodes in the network and batch size is set to $b = 5$
- Trained for $10K$ iterations and step size $\eta \approx 10^{-4}$ to $7.5 \times 10^{-2}$
- ResNet-20 on CIFAR10: with $N = 24$




(a) FCN on MNIST. (b) ResNet-20 on CIFAR10.

Figure 9: Tail-index $\alpha$ for different setting on MNIST and CIFAR10.

# Summary

- Gradient noise in SGD is not guaranteed to be Gaussian and in fact, they admit heavy-tail. These features motivate us to analyze SGD as an SDE driven by $\alpha-$stable Lévy motion.

# Summary

- Gradient noise in SGD is not guaranteed to be Gaussian and in fact, they admit heavy-tail. These features motivate us to analyze SGD as an SDE driven by $\alpha-$stable Lévy motion.
- Next, if the SGD is modeled through heavy-tailed SDE driven by Lévy motion, then the generalization and heaviness of the tail-index have a direct interplay: the heavier tail indicates better generalization.

# Summary

- Gradient noise in SGD is not guaranteed to be Gaussian and in fact, they admit heavy-tail. These features motivate us to analyze SGD as an SDE driven by $\alpha-$stable Lévy motion.
- Next, if the SGD is modeled through heavy-tailed SDE driven by Lévy motion, then the generalization and heaviness of the tail-index have a direct interplay: the heavier tail indicates better generalization.
- Since existing works on SGD about the heavy-tails do not apply in the decentralized settings, therefore, the heaviness of the tail and network structure or architecture have some relationship which are: there are two regimes of parameters (step size and network size), where DE-SGD can have lighter or heavier tails then the disconnected SGD depending on regime.

# Future research directions

- We can investigate the tail-index analysis for the DE-SGD. We want to study the metastability analysis for DE-SGD.

# Future research directions

- We can investigate the tail-index analysis for the DE-SGD. We want to study the metastability analysis for DE-SGD.
- We can study the generalization performance for DE-SGD through the lens of algorithmic stability.

# Future research directions

- We can investigate the tail-index analysis for the DE-SGD. We want to study the metastability analysis for DE-SGD.
- We can study the generalization performance for DE-SGD through the lens of algorithmic stability.
- We can also study the momentum based DE-SGD.

# Lastly

Thank you! Questions?