High Probability Guarantees for Nonconvex Stochastic Gradient Descent with Heavy Tails

ICML 2022

Shaojie Li and Yong Liu

Gaoling School of Artificial Intelligence
Renmin University of China

June 3, 2022

Background

- Stochastic gradient descent (SGD) is the workhorse in modern machine learning and data-driven optimization. As iterative algorithms, SGD works by querying an oracle for an unbiased gradient estimate built on one or several training examples in place of the exact gradients.

- Since its simplicity in implementation, low memory requirement, and low computational complexity per iteration, as well as good practical behavior, SGD is becoming ubiquitous in the big data era.

Motivations

- Existing literature provides a quite comprehensive understanding regarding the expected guarantees of SGD. However, expectation bounds do not capture the behavior of SGD within a single or few runs, which is related to the probabilistic nature of SGD.

- Many recent works suggest that SGD exhibits heavier noise than light sub-Gaussian tails. It is significant to investigate the high probability theoretical guarantees of nonconvex SGD in a heavy-tailed noise setting since it is towards a more realistic analysis.

- Moreover, existing learning guarantees of SGD are mainly derived separately either from the point of optimization performance or generalization performance. Optimization performance concerns how the learning algorithm minimizes the empirical risk, while generalization performance concerns how the predictive models learned from training samples behave on the testing samples.

Notations

- Let P be a probability measure defined on a sample space $\mathcal{Z}$, we focus on the following stochastic optimization problem with a hypothesis space indexed by $\mathcal{W} \subseteq \mathbb{R}^d$

$$\min_{\mathbf{w} \in \mathcal{W}} F(\mathbf{w}) := \mathbb{E}_{z \sim P}[f(\mathbf{w}; z)],$$

  where the objective $f : \mathcal{W} \times \mathcal{Z} \mapsto \mathbb{R}_+$ is possibly non-convex and $\mathbb{E}_{z \sim P}$ denotes the expectation with respect to (w.r.t.) the random variable z drawn form P.

- In practice, we often sample a set of i.i.d. training data $S = \{z_1, ..., z_n\}$ from P and minimize the following empirical risk

$$F_S(\mathbf{w}) := \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{w}; z_i).$$

- Let $B(\mathbf{w}_0, R) := \{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w} - \mathbf{w}_0\| \le R\}$ denote a ball with center $\mathbf{w}_0 \in \mathbb{R}^d$ and radius R. In this paper, we mainly assume that the set $\mathcal{W}$ satisfies $\mathcal{W} := B(\mathbf{0}, R)$, denoted by B_R. Let $\mathbf{w}(S) \in \arg\min_{\mathbf{w} \in \mathcal{W}} F_S(\mathbf{w})$ and $\mathbf{w}^* \in \arg\min_{\mathcal{W}} F(\mathbf{w})$.

Sub-Weibull Distribution

- We now introduce the definition of sub-Weibull random variables, which is characterized by the moment generating function (MGF).
- Definition: A random variable X, satisfying

$$\mathbb{E}\left[\exp\left((|X|/K)^{\frac{1}{\theta}}\right)\right] \leq 2, \tag{1}$$

  for some positive K and θ, is called a sub-Weibull random variable with tail parameter θ, which is denoted by $X \sim subW(\theta, K)$.
- Sub-Weibull distributions are parameterized by a positive tail index θ and reduced to sub-Gaussian distributions for $\theta = 1/2$ and to sub-Exponential distributions for $\theta = 1$.
- The higher tail parameter θ corresponds to the heavier tails.

Assumptions

We introduce some standard assumptions.

- **[Assumption 1: Sub-Weibull Noise]** Conditioned on the previous iterates, we assume the gradient noise $\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)$ is centered and $\|\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)\| \sim subW(\theta, K)$ such that $\theta \geq \frac{1}{2}$, i.e.,

  $$\mathbb{E}_{j_t}[\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)] = 0,$$

  and

  $$\mathbb{E}_{j_t}\left[\exp\left((\|\nabla f(\mathbf{w}_t; z_{j_t}) - \nabla F_S(\mathbf{w}_t)\|/K)^{\frac{1}{\theta}}\right)\right] \leq 2.$$

- **[Assumption 2: Smoothness]** Let $\beta > 0$. For any sample $z \in \mathcal{Z}$ and $\mathbf{w}, \mathbf{w}' \in \mathcal{W}$, a differentiable function $\mathbf{w} \mapsto f(\mathbf{w}; z)$ is β-smooth if

  $$\|\nabla f(\mathbf{w}; z) - \nabla f(\mathbf{w}'; z)\| \leq \beta\|\mathbf{w} - \mathbf{w}'\|.$$

Assumptions

We introduce some standard assumptions.

- **[Assumption 3]** There exists $G > 0$ such that for all $S \in \mathcal{Z}^n$,
$$\eta_t \|\nabla F_S(\mathbf{w}_t)\| \leq G, \forall t \in \mathbb{N}.$$

- **[Assumption 4]** There exists $G_* > 0$ such that for all $2 \leq k \leq n$,
$$\mathbb{E}_z \left[\|\nabla f(\mathbf{w}^*, z)\|^k \right] \leq 2^{-1} k! \mathbb{E}_z \left[\|\nabla f(\mathbf{w}^*, z)\|^2 \right] G_*^{k-2}.$$

- **[Assumption 5: PL condition]** Assume that for any $S \in \mathcal{Z}^n$, there exists an $\mu_S > 0$ such that
$$F_S(\mathbf{w}) - F_S(\mathbf{w}(S)) \leq (4\mu_S)^{-1} \|\nabla F_S(\mathbf{w})\|^2.$$
Additionally, we assume F satisfies the PL condition for some positive constant μ:
$$F(\mathbf{w}) - F(\mathbf{w}^*) \leq \frac{1}{2\mu} \|\nabla F(\mathbf{w})\|, \forall \mathbf{w} \in \mathcal{W}.$$

Main Results

We study heavy-tailed SGD with joint consideration of optimization and generalization performance.

- 1: General Nonconvex Learning.
- 2: Nonconvex Learning with PL Condition.
- 3: Towards Sharper Learning Guarantees.

Main Results: General Nonconvex Learning

Optimization Bounds: Suppose Assumptions 1 and 2 hold. Let $\mathbf{w}_t$ be the iterate produced by Algorithm SGD. Assume $\eta_t = \eta_1 t^{-\frac{1}{2}}$ with $\eta_1 \leq 1/(2\beta)$. For any $\delta \in (0,1)$, with probability $1-\delta$, (a.) if $\theta = \frac{1}{2}$, then we have the following inequality

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \eta_t \|\nabla F_S(\mathbf{w}_t)\|^2 = \mathcal{O}\Big(\frac{\log(1/\delta)\log T}{\sqrt{T}}\Big);$$

(b.) if $\theta \in (\frac{1}{2}, 1]$ and Assumption 3 holds, then we have

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \eta_t \|\nabla F_S(\mathbf{w}_t)\|^2 = \mathcal{O}\Big(\frac{\log^{2\theta}(1/\delta)\log T}{\sqrt{T}}\Big);$$

(c.) if $\theta > 1$ and Assumption 3 holds, then we have

$$\frac{1}{\sqrt{T}} \sum_{t=1}^{T} \eta_t \|\nabla F_S(\mathbf{w}_t)\|^2 = \mathcal{O}\Big(\frac{\log^{\theta-1}(T/\delta)\log(1/\delta) + \log^{2\theta}(1/\delta)\log T}{\sqrt{T}}\Big).$$

Main Results: General Nonconvex Learning

Generalization Bounds: Suppose Assumptions 1 and 2 hold. Let $\mathbf{w}_t$ be the iterate produced by Algorithm SGD. Assume $\eta_t = \eta_1 t^{-\frac{1}{2}}$ with $\eta_1 \leq 1/(2\beta)$. Selecting $T \asymp n/d$. For any $\delta \in (0,1)$, with probability $1-\delta$,

(a.) if $\theta = \frac{1}{2}$, then we have the following inequality

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla F(\mathbf{w}_t)\|^2 = \mathcal{O}\Big((\frac{d}{n})^{\frac{1}{2}} \log(\frac{n}{d}) \log^3(\frac{1}{\delta}) \Big);$$

(b.) if $\theta \in (\frac{1}{2}, 1]$ and Assumption 3 holds, then we have

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla F(\mathbf{w}_t)\|^2 = \mathcal{O}\Big((\frac{d}{n})^{\frac{1}{2}} \log(\frac{n}{d}) \log^{(2\theta+2)}(\frac{1}{\delta}) \Big);$$

(c.) if $\theta > 1$ and Assumption 3 holds, then we have

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla F(\mathbf{w}_t)\|^2 = \mathcal{O}\Big((\frac{d}{n})^{\frac{1}{2}} \big(\log(\frac{n}{d}) \log^{(2\theta+2)}(\frac{1}{\delta}) + \log^{\theta-1}(\frac{n}{d\delta}) \log^2(\frac{1}{\delta}) \big) \Big).$$

Main Results: Nonconvex Learning with PL Condition

Otimization Bounds: Suppose Assumptions 1, 2, and 5 hold. Let $\mathbf{w}_t$ be the iterate produced by Algorithm SGD. Assume $\eta_t = \frac{2}{\mu_S(t+t_0)}$ with $t_0 \geq \max\{\frac{4\beta}{\mu_S}, 1\}$. Then for any $\delta \in (0,1)$, with probability $1-\delta$,

(a.) if $\theta = \frac{1}{2}$, then we have the following inequality

$$F_S(\mathbf{w}_{T+1}) - F_S(\mathbf{w}(S)) = \mathcal{O}\Big(\frac{\log(1/\delta)}{T}\Big);$$

(b.) if $\theta \in (\frac{1}{2}, 1]$ and Assumption 3 holds, then we have

$$F_S(\mathbf{w}_{T+1}) - F_S(\mathbf{w}(S)) = \mathcal{O}\Big(\frac{\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{1}{2}} T}{T}\Big);$$

(c.) if $\theta > 1$ and Assumption 3 holds, then we have

$$F_S(\mathbf{w}_{T+1}) - F_S(\mathbf{w}(S)) = \mathcal{O}\Big(\frac{\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{3(\theta-1)}{2}}(T/\delta)\log^{\frac{1}{2}} T}{T}\Big).$$

Main Results: Nonconvex Learning with PL Condition

Generalization Bounds: Suppose Assumptions 1, 2, and 5 hold. Let $\mathbf{w}_t$ be the iterate produced by Algorithm SGD. Assume $\eta_t = \frac{2}{\mu_S(t+t_0)}$ with $t_0 \geq \max\{\frac{4\beta}{\mu_S}, 1\}$. Selecting $T \asymp n$. Then for any $\delta \in (0,1)$, with probability $1-\delta$,

(a.) if $\theta = \frac{1}{2}$, then we have the following inequality

$$F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) = \mathcal{O}\Big(\frac{d + \log(\frac{1}{\delta})}{n} \log^2(\frac{1}{\delta}) \log n\Big);$$

(b.) if $\theta \in (\frac{1}{2}, 1]$ and Assumption 3 holds, then we have

$$F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) = \mathcal{O}\Big(\frac{d + \log(\frac{1}{\delta})}{n} \log^{(2\theta+1)}(\frac{1}{\delta}) \log n\Big);$$

(c.) if $\theta > 1$ and Assumption 3 holds, then we have

$$F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) = \mathcal{O}\Big(\frac{d + \log(\frac{1}{\delta})}{n} \log^{(2\theta+1)}(\frac{1}{\delta}) \log^{\frac{3(\theta-1)}{2}}(\frac{n}{\delta}) \log n\Big).$$

Main Results: Towards Sharper Learning Guarantees

Generalization Bounds: Suppose Assumptions 1, 2, 4, and 5 hold. Let $\mathbf{w}_t$ be the iterate produced by Algorithm SGD. Assume $\eta_t = \frac{2}{\mu_S(t+t_0)}$ with $t_0 \geq \max\{\frac{4\beta}{\mu_S}, 1\}$. Selecting $T \asymp n^2$.

When $n \geq \frac{c\beta^2(d+\log(\frac{8\log(2nR+2)}{\delta}))}{\mu^2}$ where c is an absolute constant, for any $\delta \in (0,1)$, with probability $1-\delta$,

(a.) if $\theta = \frac{1}{2}$, then we have the following inequality

$$F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) = \mathcal{O}\Big(\frac{\log^2(\frac{1}{\delta})}{n^2} + \frac{\mathbb{E}[\|\nabla f(\mathbf{w}^*, z)\|^2]\log(\frac{1}{\delta})}{n}\Big);$$

(b.) if $\theta \in (\frac{1}{2}, 1]$ and Assumption 3 holds, then we have

$$F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) = \mathcal{O}\Big(\frac{\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{1}{2}} n}{n^2} + \frac{\mathbb{E}[\|\nabla f(\mathbf{w}^*, z)\|^2]\log(1/\delta)}{n}\Big);$$

(c.) if $\theta > 1$ and Assumption 3 holds, then we have

$$F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) = \mathcal{O}\Big(\frac{\mathbb{E}[\|\nabla f(\mathbf{w}^*, z)\|^2]\log(1/\delta)}{n} + \frac{\log^{\frac{3(\theta-1)}{2}}(n/\delta)\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{1}{2}} n}{n^2}\Big).$$

Main Results: Towards Sharper Learning Guarantees

Generalization Bounds: Suppose Assumptions 1, 2, 4, and 5 hold. Let $\mathbf{w}_t$ be the iterate produced by Algorithm SGD. Assume $\eta_t = \frac{2}{\mu_S(t+t_0)}$ with $t_0 \geq \max\{\frac{4\beta}{\mu_S}, 1\}$. Assume that $F(\mathbf{w}^*) = \mathcal{O}(\frac{1}{n})$. Selecting $T \asymp n^2$. When $n \geq \frac{c\beta^2(d+\log(\frac{8\log(2nR+2)}{\delta}))}{\mu^2}$ where c is an absolute constant, for any $\delta \in (0,1)$, with probability $1-\delta$,
(a.) if $\theta = \frac{1}{2}$, we have the following inequality

$$F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) = \mathcal{O}\Big(\frac{\log^2(\frac{1}{\delta})}{n^2}\Big);$$

(b.) if $\theta \in (\frac{1}{2}, 1]$ and Assumption 3 holds, then we have

$$F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) = \mathcal{O}\Big(\frac{\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{1}{2}} n}{n^2}\Big);$$

(c.) if $\theta > 1$ and Assumption 3 holds, then we have

$$F(\mathbf{w}_{T+1}) - F(\mathbf{w}^*) = \mathcal{O}\Big(\frac{\log^{\frac{3(\theta-1)}{2}}(\frac{n}{\delta})\log^{(\theta+\frac{3}{2})}(\frac{1}{\delta})\log^{\frac{1}{2}} n}{n^2}\Big).$$

Conclusion

- This paper establishes high probability learning guarantees for nonconvex SGD.
- In contrast to most theoretical studies, we consider the stochastic gradient noise following a novel class of heavy-tailed sub-Weibull distribution.
- Our analysis involves joint consideration of optimization and generalization performance.
- Under different assumptions, we push the learning guarantees to different orders.
- We also study clipped SGD to remove a very commonly used assumption (see the main paper). Additionally, in this case, the stepsize of SGD is completely oblivious to the knowledge of smoothness.

Thank You